Feasibility and Utility of Multimodal Micro Ecological Momentary Assessment on a Smartwatch

Ha Le
Northeastern University
Boston, Massachusetts, USA
le.ha1@northeastern.edu

Veronika Potter
Khoury College of Computer Science
Northeastern University
Boston, Massachusetts, USA
potter.v@northeastern.edu

Rithika Lakshminarayanan
Khoury College of Computer Science
Northeastern University
Boston, Massachusetts, USA
lakshminarayanan.r@northeastern.edu

Varun Mishra
Northeastern University
Boston, Massachusetts, USA
v.mishra@northeastern.edu

Stephen Intille
Khoury College of Computer Sciences
and Bouve College of Health Sciences
Northeastern University
Boston, Massachusetts, USA
s.intille@northeastern.edu

Abstract

μEMAs allow participants to answer a short survey quickly with a tap on a smartwatch screen or a brief speech input. The short interaction time and low cognitive burden enable researchers to collect self-reports at high frequency (once every 5-15 minutes) while maintaining participant engagement. Systems with single input modality, however, may carry different contextual biases that could affect compliance. We combined two input modalities to create a multimodal-μEMA system, allowing participants to choose between speech or touch input to self-report. To investigate system usability, we conducted a seven-day field study where we asked 20 participants to label their posture and/or physical activity once every five minutes throughout their waking day. Despite the intense prompting interval, participants responded to 72.4% of the prompts. We found participants gravitated towards different modalities based on personal preferences and contextual states, highlighting the need to consider these factors when designing context-aware multimodal μEMA systems.

CCS Concepts

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**.

Keywords

Ecological momentary assessment, Experience sampling, Ubiquitous computing; Wearable computing; Speech input; Touch input; Multimodal input

ACM Reference Format:
Ha Le, Veronika Potter, Rithika Lakshminarayanan, Varun Mishra, and Stephen Intille. 2025. Feasibility and Utility of Multimodal Micro Ecological Momentary Assessment on a Smartwatch. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3706598.3714086

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714086

1 Introduction

Accurately detecting human behaviors is an important research area in ubiquitous computing, human-computer interaction, and personal health informatics. Researchers can use behavior recognition models to drive context-aware interactive systems, just-in-time interventions, or health monitoring tools. Building behavior inference systems requires datasets that have *high-fidelity labels of behavior* that can be used to effectively train models. Systems that can be used to collect *temporally-dense*, in-situ behavioral data may result in datasets that could be used to help build and validate behavioral recognition models. Such labeled data might also be used to create personalized recognition models that may accurately detect behavioral patterns.

A particularly important human activity recognition (HAR) task in health and other domains is the detection of human posture, physical activity (PA), and sedentary behaviors from wearable sensors. Effectively training, validating, and benchmarking such HAR models, however, requires large, continuously labeled datasets. Most wearable HAR datasets are collected in heavily controlled [36] or semi-controlled [18] settings; the resulting datasets may not reflect the diversity and complexity of activities that people engage in during daily living. Models trained on such datasets, therefore, often perform poorly when tested on data from real-world, less-controlled scenarios [112] — where people have a wider range of activities or move their bodies more naturally. For example, when someone is told to *sit* in the lab, they may move in a controlled way and sit upright, but when they *sit* in real-life, they may plop down on a couch and lounge. Moreover, some activities, such as *driving*, are common in everyday life but difficult to capture realistically in lab protocols. The gold-standard method used to collect labels in-the-wild is to use an egocentric on-body camera to record images or video; the visuals are then used to label posture and behaviors post-hoc [14, 32]. For researchers, labeling the images or videos [45] is time-consuming, tedious and resource-intensive. Further, because human labeling is involved, the method cannot be used for developing systems that gather new labels in real-time and update models in-situ. For the participants in studies, the burden of wearing the front-facing, on-body camera can raise significant privacy concerns [46] about recording the activities of the participant *and* those nearby; thus participant recruiting can be challenging.

Ecological Momentary Assessment (EMA) is a data collection method whereby participants in research studies are prompted to complete short surveys periodically, often on their smartphone [98]. μEMA is a specific form of EMA where where researchers prompt participants to answer a single multiple-choice question using an “at-a-glance,” single-tap interaction, often on a smartwatch screen [43]. Researchers have extended μEMA to enable participants to provide their behavioral labels using speech input, triggered by a vibration on a smartwatch or a *beep* heard through an earable [59, 91]. By design, each μEMA prompt delivers only a single question that can be answered quickly, which enables μEMAs to be prompted at a high interval (e.g., once every 5-15 min) while maintaining a high response rate and thus ensuring temporally dense labels. One challenge with deploying μEMAs is that there may be contextual biases in response rates that depend, in part, on the modality of data entry (e.g., participants may be less inclined to speak to the watch in public settings or tap on the watch during vigorous exercise) [59, 83].

In this work, we combine speech and touch input to allow **multimodal μEMA**, giving participants maximum flexibility when self-reporting behaviors (in this case their posture and activity). We explore the impact of multimodal interaction on reporting burden, where our aim is to test a methodology that may allow participants to maintain a high compliance rate despite data collection using a temporally intensive prompting interval—once every five minutes. Our research questions are:

- **RQ1:** What is the usability and feasibility of using multimodal μEMA to collect temporally dense posture and physical activity labels in-the-wild?
- **RQ2:** What factors affect participants' non-response and modality choice when responding to multimodal μEMA prompts?
- **RQ3:** What are the characteristics and potential utility of the posture and physical activity labels collected with multimodal μEMA?

To address these questions, we conducted a seven-day field study with 20 participants. The key contributions of our work are:

- We introduce a new multimodal μEMA data collection method that allows speech and/or touch input. We explore the usability of our system in the context of collection of posture and activity labels by conducting a mixed-method field study with 20 participants for seven days.
- We quantitatively show that on our acquired dataset, passively sensed contextual parameters (heart rate, wrist movements, location, ambient noises, phone usage, time of day and day of week) are associated with response rate and choice of interaction modality. Furthermore, we qualitatively examine how interruption and interaction burden are associated with multimodal μEMA.
- We explore the characteristics and utility of labels collected using our system, and we demonstrate that automatic, real-time label extraction is possible using an adapted commercial speech recognition model and an open-source large-language model.

2 Related Works

This work builds on prior research on ecological momentary assessment (EMA), multimodal input, and in-situ data collection — the methods that have been used to collect human behavioral data.

2.1 Collecting in-situ behavioral labels using ecological momentary assessment (EMA)

Ecological momentary assessment [95], sometimes called the experience sampling method, is a data collection method widely used in behavioral monitoring research to collect longitudinal data [19, 98]. Using EMA, researchers can collect ecologically valid measurements by using notifications on a mobile phone or wearable device to prompt participants to report data in-situ. The primary disadvantage of EMA is that the participants may perceive notifications as burdensome *because* they are prompted in-situ; pulling out a device and stopping an ongoing activity to answer surveys can disrupt the behaviors being measured. Thus, most research studies seek to balance compensation and burden to ensure a high response rate to EMA [111]. μEMA [43, 52] is a modified version of the standard EMA method where each prompt is guaranteed to include only a single-question survey, often presented on the smartwatch, that can be answered with a quick tap; μEMAs (microinteraction EMAs) are designed to be answerable with “at a glance,” single-tap interactions that take only 2-3 s. Prior research has demonstrated that even when μEMA surveys are delivered at rates of up to four times an hour, participants in research studies can maintain a response rate significantly higher than for standard EMA [43, 82]. μEMAs can be implemented well on a smartwatch, because the watch is easily accessible on the wrist. One limitation of μEMA delivered on a smartwatch, however, is the limited amount of space available on the watch screen, which reasonably can only support multiple-choice questions with less than five options. While this limitation helps ensure μEMA questions do not become too complex, it also makes μEMA more suitable as a prediction confirmation mechanism (e.g., “Are you walking?” “Yes/No”) than for input that involves selecting from a list of possibilities (e.g., "What are you doing now?" with a long list of activities). Another consideration when using smartwatch-based μEMA is that answering a question requires a two-handed interaction, which can be inconvenient during certain activities (e.g., driving or carrying groceries). Audio-μEMA [59] allows participants to use speech input to provide open-ended responses. In audio-μEMA, the system prompts participants using either a short acoustic cue presented through an earable, or if a watch is used, through a vibration on the wrist; the prompt indicates that the participant should speak the answer to a known question (e.g., what is their in-the-moment behavior). Due to the hands-free nature of speech interaction [89], audio-μEMA could allow capture of a wide range of behavioral labels (e.g., postures, activities, or contextual information) while maintaining a high response rate.

A limitation of all EMA-based implementations is that in-the-moment contextual parameters may affect response rate [83]; this contextual reactivity could impact the validity of some types of behavior data collected using this method (e.g., if participants are unwilling to respond while exercising, we would not gather any labels on this activity to train recognition models). Thus, combining different modalities of EMA and a retrospective recall may be required to capture a comprehensive picture of a participant's entire waking day. Multimodal EMA systems have been developed for home environments [63, 110]; in that work, even though participants preferred touch interactions to voice commands, modality preferences varied based on participants' contextual states. However, we are aware of no research to date on using a multimodal μEMA system in free-living settings.

2.2 Multimodal data logging and tracking systems

People interact with the world around them by speaking, touching, gesturing, drawing, and pointing; they use different modes, alternatively or simultaneously, in different contexts [94]. Multimodal input interfaces can improve user experience and system robustness [78] on multiple tasks such as data analytics/exploration [49, 51, 92, 101], tracking exercise, [66] tracking food intake [53, 65, 67, 96], and in-car communication [44, 54, 58, 104].

Touch is the most common input modality used in health-tracking applications, mainly due to the ubiquitous nature of commercial smartphones and smartwatches. Recently, there has been an increase in research on voice-based interfaces for behavioral logging [4, 51, 66]. Researchers have implemented voice-based interfaces on many device form factors because voice does not require a physical interface. Although researchers reported users' positive reactions to speech input, the method presents challenges related to addressing cognitive load, social context, and speech recognition errors [96]. Other modalities like touch or keyboard often complement speech inputs to allow error reconstruction [78]. Researchers have also explored the uses and combination of different input modalities (e.g., voice log [38, 65, 67, 96], photos [27, 28, 65, 69], touch [65, 96], and type [69]) on different device form factors (e.g., mobile [65, 67, 69, 96], desktop [69], smartwatch [4, 38, 50], and smart speaker [63]) for behavioral journaling/tracking.

Even if a system allows multimodal interaction, users are not guaranteed to interact multimodally [77]. A user's modality usage patterns are heavily influenced by external and task contexts. Researchers have shown modality usage can be affected by the cognitive and communication load of the task [79], contextual variables (e.g., surrounding environments, movements/activities, hand usage, visual load) [61, 87], and physical/mental interaction effort [13]. In this paper, we investigate the usability of our multimodal μEMA system and examine different contextual variables that affect user modality choice, given the task of recording posture and physical activity labels.

2.3 Collecting human activity labels in-the-wild

Capturing information about a person's free-living physical activities (PA) labels in-the-wild can support better training of HAR models and more realistic evaluation of such models. One approach is to use participants' self-reported data, often acquired either via end-of-day survey [8], in-situ measurements (EMA) [85], or a mixture of both [41, 106]. Labels collected using end-of-day recall surveys tend to suffer from recall bias, in which events happening before or after can affect the recalled event. Fast-changing, overlapping sequences of activities [6] can often lead to mistakes in labeling activity boundaries, which can significantly reduce model performance [55]. An alternative approach to self-report is to use a body-worn camera, where participants wear a front-facing camera around their neck [24, 34, 47] or on their head [32] to capture an egocentric narrative of their daily life. Although this approach does not create a response burden for the participant, it introduces privacy concerns that may hinder participant recruitment and is not sustainable for longitudinal studies. Furthermore, labeling a large volume of video data after-the-fact is time-consuming. The quality of annotating videos retrospectively relies on the annotators' ability to extract information about a person's activity from the first-person narrative video without any additional self-report context.

In-situ measurement methods, such as EMA and μEMA, enable participants to annotate their data in real-time, reducing the cognitive biases associated with event recall. Voice-based, open-ended μEMA is particularly well-suited for capturing detailed human activity labels, given the complexity and variability of human behavior. We hypothesized that combining voice- and touch-based inputs could mitigate the contextual bias linked to non-responses in voice-based μEMA, while maintaining a high participant response rate and allowing the collection of a diverse range of postures and activities.

3 Multimodal μEMA: System design and implementation

The goal of the multimodal μEMA system described in this work is to collect temporally-dense activity and posture labels from participants while maintaining high compliance. To achieve this goal, we allow participants to answer the prompt using either touch or speech input (Figure 1). In this section, we outline the components of our system and discuss their implementations.

3.1 Prompt design

The smartwatch uses haptic cues to prompt participants to report their in-the-moment posture and activity. The haptic cue is short (~1 s) to avoid distracting participants should they choose to, or need to, ignore the prompt. The cue is ideally intense enough so that participants are unlikely to miss the prompt during bouts of intensive activity or in loud environments. We chose not to use an auditory cue, because sounds from the smartwatch can be overheard by others nearby, potentially causing social disruptions. After the haptic cue, participants have 10 s to start responding to the prompt. To capture temporally dense activity we prompt for μEMA input once every five minutes.

3.2 Touch interaction

After the haptic cue, the watch screen displays four quadrants featuring the system's predictions of the most likely PA or posture

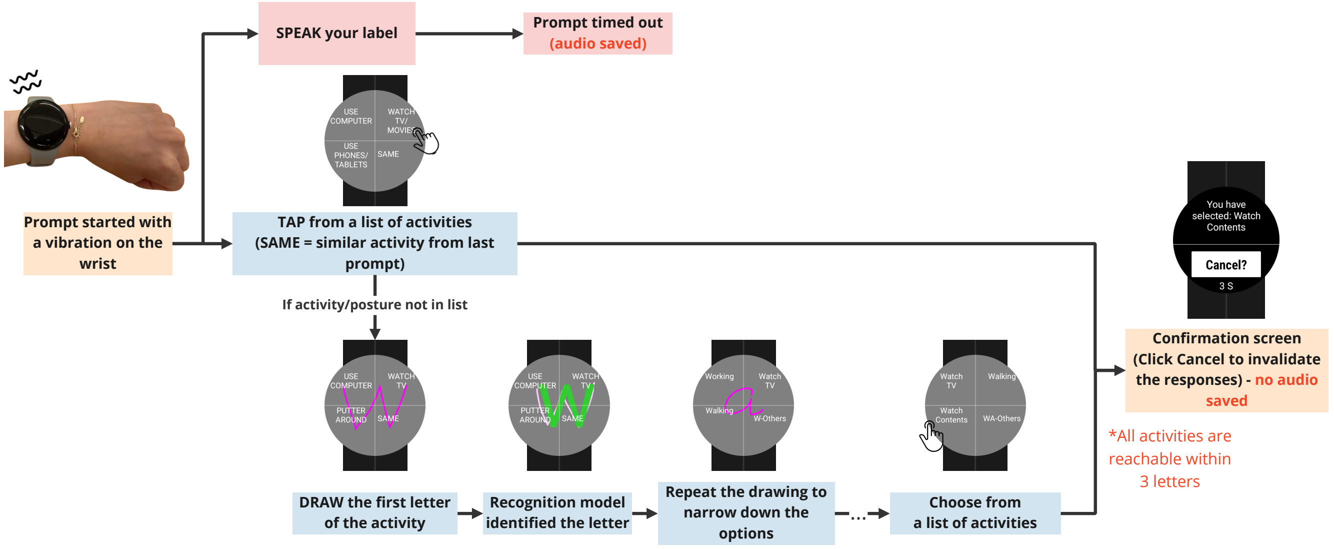

SPEAK your label
Prompt timed out
(audio saved)
USE COMPUTER
WATCH TV/ MOVIES
USE PHONES/ TABLETS
SAME
Prompt started with a vibration on the wrist
TAP from a list of activities (SAME = similar activity from last prompt)
If activity/posture not in list
USE COMPUTER
WATCH TV
PUTTER AROUND
SAME
USE COMPUTER
WATCH TV
PUTTER AROUND
SAME
Working
Watch TV
Walking
W-Others
Watch TV
Walking
Watch Contents
WA-Others
You have selected: Watch Contents
Cancel?
3 S
Confirmation screen (Click Cancel to invalidate the responses) - no audio saved
*All activities are reachable within 3 letters
DRAW the first letter of the activity
Recognition model identified the letter
Repeat the drawing to narrow down the options
...
Choose from a list of activities

Figure 1: Multimodal μEMA allows users to self-report their posture and physical activity labels using either speech or touch input on a smartwatch. When prompted using watch vibration, participants can complete the survey by either (1) speaking normally (being recorded by the watch) or (2) tapping buttons or drawing letters then tapping a button on the watch to complete the survey.

labels. The list of four activities shown in the initial screen is determined by the participant's self-reported most-common activities, a ranked list of which is acquired during a study onboarding session. Subsequently, the system uses the participant's previous μEMA responses, current heart rate (HR), and the last 10 s of wrist motion (determined using accelerometer data from the watch) to populate the screen. We discuss the details of how we narrow down the options in the implementation section and the appendix of this paper.

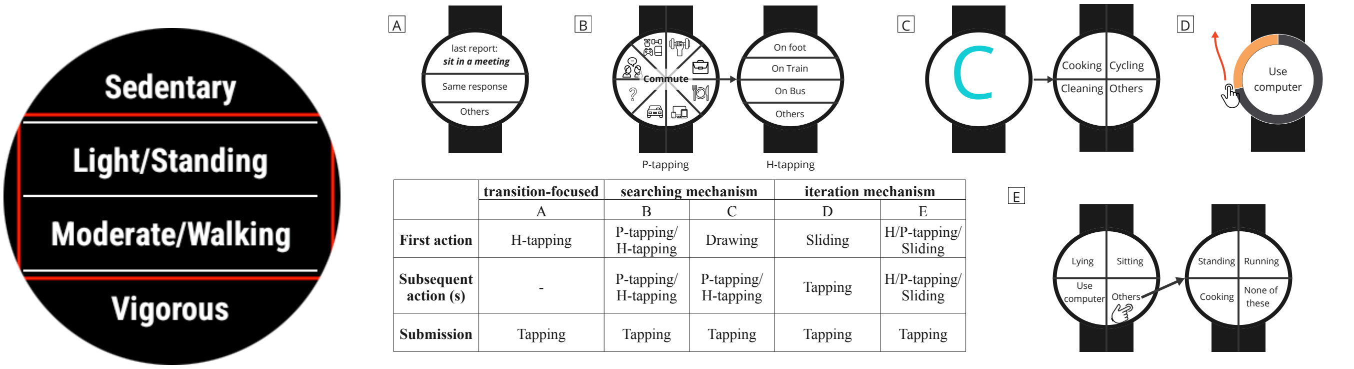
If the participant chooses to interact with the system using a touch interaction, they select a physical activity or posture by tapping on the appropriate button quadrant on the watch screen. If the participant chooses to use speech for their previous self-report, then the watch screen will show an option labeled "SAME" in one of the quadrants — this can be pressed to select the last-reported activity/posture again. If the participant uses touch for the prior self-report, then the name of the label that was selected will replace the "SAME" option. If the participant cannot find their in-the-moment activity on the screen, they can draw the first letter (or more) of their activity on the screen over the buttons; this will trigger a search for a different posture or activity that starts with the letter(s). The system then displays the three most probable activities based on the letters drawn. For example, if the participant draws letter "B"/"b," the screen might show "Biking," "Bus (Riding)," and "Baking." If participants draws "CA," the screen might show "Car (Riding)," "Car (Driving)," and "Carrying Stuff." Participants can click on the fourth quadrant ("Others") if they still cannot find the label.

We considered combinations of three different mechanisms and five different interactions on smartwatch self-reporting interfaces (Figure 2b) [113, 114] based on prior work. Our internal pilot testing suggested options A and C were more intuitive and less error-prone than the other input options while still allowing us to include an unlimited number of posture/activity labels. Option A enables users to indicate the same activity/posture to the last prompt, which helps reduce cognitive burden and interaction time. Option C enables users to search for the label by drawing letters on the watch screen. We displayed the list of labels in a radial/pie-list layout (P-list) rather than the traditional horizontal-list layout (H-list) commonly used in previous μEMA studies [85]. This decision was supported by our internal testing, which suggested the H-list layout made the middle options difficult to click due to limited spacing (Figure 2a). In contrast, the P-list layout provides equal space for all four options and maximizes the interaction area.

Participants can cancel their responses if they make a mistake (such as by tapping the wrong option, drawing incorrect letters, or if the system misinterprets the letters). Participants have 10 s to complete the self-report after every letter. Participants can cancel their self-report by selecting the "Cancel?" button at the end of the prompt or at any time by drawing the letter "X" on the watch face. Participants can also simply not answer the prompt.

3.3 Speech interaction

In addition to touch interaction, system users can report their postures and activities using speech. We use the same interaction design as prior work on an audio-μEMA system [59]. When the watch prompts the participant, the watch records audio either until the participant reports using touch interaction, or 10 s has elapsed. Because they do not need to even look at the watch or move their hand, participants can maintain their movements while reporting with speech. Although it is not required, even when participants



	transition-focused	searching mechanism		iteration mechanism	
	A	B	C	D	E
First action	H-tapping	P-tapping/ H-tapping	Drawing	Sliding	H/P-tapping/ Sliding
Subsequent action (s)	-	P-tapping/ H-tapping	P-tapping/ H-tapping	Tapping	H/P-tapping/ Sliding
Submission	Tapping	Tapping	Tapping	Tapping	Tapping

(a)

(b)

Figure 2: Designing touch interaction. Figure (a) is an example of a μEMA prompt using an h-list; when the buttons are small, the options outlined in red are more difficult to tap than the options at the top and bottom and thus prone to mistakes. Figure (b) shows the different designs proposed in our pilot testing. We considered three interaction mechanism: A) participants tap on "SAME" to indicate if they are doing similar activity with the previous prompt, or "Others" if they cannot find the activity. B and C) participants narrow down the list of labels by tapping on the high-level behavior (e.g., commute, work-related, relaxation) or drawing the first letter of the label. D and E) participants cycled through a list of options by either swiping or tapping on the screen until they found their desired label.

bring their hand closer to their mouth to use the speech input, it is still a one-hand interaction (versus two-hand required for touch).

The system retains audio recordings until participants complete labeling using touch input so that participants can switch to speech input if their desired labels do not appear in the list using touch. When participants opt to report via the touch interaction and complete the interaction, the audio recording for that prompt is deleted to minimize privacy concerns.

3.4 Implementation

We implemented the multimodal μEMA system to work on Android Wear devices. For the evaluation study, we loaned participants a Pixel Watch 2 (Alphabet, Inc) and paired the watch to the participant's personal Android phone, which was running Android 9 or above. The connection to the phone was used to transfer data to our research server during the study. We implemented the system to transfer data from the watch to our server once every hour using Bluetooth and network connections to avoid overflowing the watch storage. From our initial testing, however, the software can function without network connection for up to one month.

We determine the list of activities shown on the watch screen during a touch interaction using previous responses, accelerometer and heart rate (HR) data collected from the smartwatch, and common self-reported activities. The smartwatch estimates physical activity intensity using a real-time algorithm that measures the overall motion of the wrist based on accelerometer data. The smartwatch samples raw tri-axial accelerometer data at 50 Hz and smooths the raw signal using a moving average filter with a window size of 0.5 s (filtered signal). For each axis, it computes the area under the curve (AUC) $AUC_t = |raw_t - filtered_t|$ to compensate for the effect of gravity (DC offset for the axis) and calculates a 10 s summary of AUC by summing AUC values from the three axes to derive a physical activity summary unit [57, 59]. We implemented the system to use a combination of HR and accelerometer data to account for variability in participants' heart rate level and wrist motion during sedentary activities (e.g., hand gesturing during conversation). More details about how the system determined the suggested activities are explained in Appendix A.

If participants decide to draw on the watch screen to narrow down the labels, the system uses a predefined activity and posture list to narrow the search. The research team predetermined the mapping between letter sequences and activities/postures. Four research team members independently coded common activity abbreviations; and we used the abbreviations to set the mapping. We derived the list of common activities from the 2024 Adult Compendium of Physical Activities [37] and added in additional activities identified during our internal pilot testing. We included the list of labels in our system in the supplementary materials. We used Google Firebase's digital ink recognition model [1] (Alphabet, Inc.) to identify drawn letters.

4 STUDY DESIGN

We conducted a mixed-method, seven-day free-living study with 20 participants to evaluate the multimodal μEMA system. The study took place in three parts: a ≈60-minute in-person introduction/training session, a seven-day free-living period where participants used the multimodal μEMA system to record their behavior, and a ≈60-minute exit interview.

To ensure that the study was adequately powered to detect meaningful differences, we conducted an a priori power analysis using the MRT-SS Calculator[2] for a micro-randomized control trial for 7 days, assuming 120 decision time points per day (i.e., 10 hours of prompting μEMA, a constant randomization probability of 0.5 at each decision point, and an expected availability of 70%). This

[1] https://developers.google.com/ml-kit/vision/digital-ink-recognition
[2] https://d3center.shinyapps.io/mrt-ss-calculator/

response rate assumption is based on the response rate from prior works on μEMA and audio-μEMA [38, 43, 59, 85]). We set the desired power at 80% with a significance level of 0.05. Based on these parameters, the analysis indicated that a minimum of 13 participants would be required to detect the hypothesized proximal effect size. Our targeted recruitment of 20 participants exceeded the required sample size.

We recruited participants using posters placed around an academic campus, social media posts, and campus mailing lists. To be eligible to enroll in the study, participants 1) were at least 18 years old, 2) had no cognitive or hearing impairments, 3) were able to read their phone without reading glasses, 4) used an Android phone, 5) were willing to install an Android application developed by the research team on their phone, and 6) were willing to wear a smartwatch provided by the research team for seven days and answer the prompts on the watch. The study protocol was approved by the IRB at Northeastern University. We compensated participants $75 in Amazon gift cards for the study ($20 for each in-person session, and $5 for each day they wore the watch). We targeted recruitment outside the computer science department and non-STEM students. Besides the Amazon Gift Cards, participants in the study did not receive any other incentives, e.g., class credits.

In later sections of the paper, we use the prefix P with a number to denote participants from the field study. We show the demographic summary of the 20 participants in Table 1.

After obtaining informed consent, we began the training session by collecting demographic information about the participant — age, occupation, and self-reported data about daily habits and physical activity level. A research assistant paired the study smartwatch with the participant's personal Android phone and installed the study application to ensure the system would work during the free-living period. We showed participants a video demonstrating how to use the multimodal μEMA system (included in the supplemental materials). We asked participants to practice answering the μEMA prompts with a research assistant so they could receive real-time feedback on their responses and ask clarifying questions.

The seven-day, free-living portion of the study began the day after the introductory training session. We instructed participants to wear the provided smartwatch during their waking hours (or until the watch ran out of battery), and report their in-the-moment physical activity and/or posture each time they received a prompt, which was every five minutes. Participants could respond to each prompt using either speech or touch input. The system prompted participants two hours before their sleep time to answer a daily burden survey on their phone. The survey asked them to report any instances in which they removed the watch, their experiences with the system that day, and their expected sleep/wake times for the following day. Additionally, they answered four questions related to perceived burden. The burden questions were: "*I feel comfortable wearing the smartwatch*," "*I easily responded to the smartwatch prompts*," "*I responded to the smartwatch prompts quickly*," and "*The smartwatch is easy to learn how to use*." These questions were adapted from prior work [59] and used a five-point Likert-scale answer, ranging from "Strongly Disagree" to "Strongly Agree." Through these questions, we attempted to measure participants' interaction burden, including their comfort with the watch and the ease of responding to prompts, specifically examining perceived response speed and difficulties in formulating answers. We also asked if the watch or the phone needed to be recharged at any time during the day. To gain additional insights about the effects of the prompting frequency, we collected qualitative feedback from participants during the exit interviews. Once a day, a research assistant sent a text message to the participants to remind them to wear the watch and answer any questions or concerns.

At the end of the seven-day period, participants returned the watch and attended an in-person semi-structured exit interview. During the interview, we asked participants about their experiences using the system including difficulties they may have had while using the system, scenarios when they chose to use speech verses touch input to answer the prompts, and factors influencing their willingness to respond to a prompt. Participants also provided general feedback they had on the system and how to improve multimodal μEMA in future deployments. Overall, 20 participants consented to enroll in the study, and all 20 participants finished the 7-day study (no dropouts).

5 Analysis Plan

Our analysis tested these following hypotheses and used thematic analysis to analyze the transcript from the exit interviews.

5.1 Hypotheses

We tested four hypotheses to quantitatively evaluate research questions **R1** and **R2**. The hypotheses are motivated by prior works on μEMA and EMA. *H1* and *H2* are related to how ***temporal*** factor (day into study) affects participants' response behavior. *H3* and *H4* are related to how ***passively-measured*** factors affect participant's response behavior.

- *H1:* Participants' response rate to μEMA would decrease over time.
- *H2:* Participants' perceived burden of multimodal μEMA would decrease over time.
- *H3:* There are associations between passively measured contextual factors and participants' modality choice.
- *H4:* There are associations between passively measured contextual factors and participants' response rate.

Prior works have shown that day-into-study has a significant effect on μEMA and EMA response rate. Non-response for EMA tends to be lowest in the beginning of a study and then increase as the study goes on [11, 20, 62, 83, 97]. We also hypothesized that perceived burden would decrease over time as participants acclimated to the smartwatch and responding to μEMA.

To examine the effects of passively-measured factors on participants' modality choices, we choose seven contextual variables: heart rate, wrist movement, location, phone usage, ambient noises, time of day, and day of week. The passively-measured variables were selected based on prior works, as well as the sensing capability on the smartwatch and the phone. Results from prior works on EMA, μEMA, and multimodal interactions and impact on response rates are summarized in Table 2.

5.2 Thematic analysis

We performed inductive coding to assess the usability of multimodal μEMA and identify participants' perceive source of burden.

Table 1: Demographics of participants in the field study (N=20). The categories for ethnicity and race are those recommended by the U.S. National Institute of Health [75]. We do not include categories with no participants in the table.

		N=20
Age (mean (STD))		25.8 (3.1)
Sex (n (%))	Female	7 (35%)
	Male	13 (65%)
Ethnicity (n (%))	Non-Hispanic	19 (95%)
	Hispanic	1 (5%)
Race (n (%))	Asian	16 (80%)
	White	2 (10%)
	Bi-racial	2 (10%)
Occupation (n (%))	Student	13 (65%)
	Full-time employed	4 (20%)
	Part-time employed	3 (15%)
Daily routine (n (%))	Highly structured	3 (15%)
	Fairly structured	9 (45%)
	Moderately structured	4 (20%)
	Not structured	4 (20%)
Activity level (n (%))	Sedentary	11 (55%)
	Moderate	3 (15%)
	Vigorous	6 (30%)
Familiar with tracking technologies (n (%))	Very Familiar	11 (55%)
	Somewhat familiar	6 (30%)
	Not familiar	3 (15%)
Use a smartwatch (n (%))		7 (35%)

Table 2: Summary of prior research examining various passive sensing variables' effects on response rate and modality choice for EMA and μEMA.

Contexts	*Prior Works*	*Passively-Measured Variables*	*Study Findings*
Activity levels	[25, 39, 48, 64, 72, 83]	Heart rate, wrist movement	Increased response rates and voice interaction for physical activity group than sedentary group
Phone usage	[83, 108]	Phone interactive	Decreased non-response with recent phone use
Social context	[87, 103, 116]	Detecting speech or conversation in the background	Increased non-response and decreased voice interaction during social interactions
Environmental noise	[39, 72]	Detecting noises in the background	Increased non-response and decreased voice interaction with environmental noise
Location	[88, 99]	Home vs. not home	Decreased non-response at home than other locations
Time of day	[11, 20, 83, 97]	Morning, afternoon, evening, and night	Decreased non-response in the morning than in the afternoon or late evening
Day of the week	[68, 99]	Weekday vs. weekend	Not a statistically significant predictor for response rate

Two authors carefully read each transcript from our field study and performed open-ended coding using inductive coding [26]. The authors coded the transcript independently and met frequently to reconcile disagreement. The codes were generated and improved iteratively. We merged similar codes/themes and removed codes outside the scope of our research. The interrater agreement (Cohen's kappa) was $\kappa = 0.74$.

6 RESULTS

We report quantitative and qualitative results from the field study to answer each research question (RQ).

6.1 RQ1: Assessing the usability of multimodal μEMA

In the field study, we collected 135 days of data from 20 participants. We lost three days of data for P16 and two days of data for P10 because the participants deleted the app before the exit interviews and the devices had not transmitted data to the research server. Participants responded to 11,320 of the 15,635 delivered prompts (72.4%) — averaging 84 prompts per day. We investigated the usability of the multimodal μEMA system on a smartwatch for seven days by calculating compliance and usability metrics (Table 3). We collected responses on the System Usability Survey (SUS) [12] during the exit interview.

We calculated three prompt response metrics (Table 3): 1) compliance rate refers to the number of prompts that participants interacted with (*promptsAnswered*) over the number of prompts scheduled based on participants' wake/sleep time (*promptsScheduled*); 2) response rate refers to *promptsAnswered* over the number of prompts successfully delivered (*promptsDelivered*); and 3) success rate refers to *promptsCompleted* over the number of prompts answered by the participants. *promptsCompleted* are touch responses that participants did not cancel, or speech responses that are intelligible to human annotators. We computed the compliance rate as in prior works on μEMA [43, 59, 83]. The completion rate shows the level of engagement the participant had with the system (the percentage of the questions answered among the questions successfully prompted and delivered) while the compliance rate shows how much data the system captured within a waking day relative to what was anticipated. Compliance is an important metric to consider for future studies to deploy multimodal μEMA to annotate activities and postures for the *entire waking day.* Overall, participants were highly engaged with the system, with the response rate of 72.4% (45.7% or 5,087 of the responses were speech input). The battery limitation of the watch explains the gap in compliance and response rate. The majority of the responses were captured successfully by the system (success rate of 99.8%). Among the *promptsAnswered*, 6% of the responses were "SAME" (n = 670). The majority (94.8%, 128/135) of the canceled touch inputs were recovered by speech input. Even though there was no upper bound on response time for touch input, the average interaction time was 0.3 s.

We statistically analyzed whether day-into-study affects response rate over time (*H1*) (Figure 3). We used linear mixed-effect models with random intercept for each participant, with the following formula:

$$response\_rate \sim day\_into\_study + (1|subject\_id).$$

Results from the linear mixed-effect model show that response rate decreased over time ($\beta = -.01$, $SE = 0.01$, $p < .05$). This indicates that the response rate declined by 1% for each day of the study.

We collected 106 responses for the daily burden survey (response rate: 84.8%). Among the responses, 71 (66.9%) mentioned the watch or the phone needed to be recharged during the day: "*It was quite smooth to use just that the battery [of the watch] would get drained out fast*" [P21]. For each daily burden survey, we converted the Likert responses into a numeric score, with "Strongly Disagree" as 1 and "Strongly Agree" as 5. We statistically tested whether the participants' responses to individual questions on the daily burden survey changed over time (*H2*). We fitted linear mixed-effect models for each of the four survey questions with random intercept for each participant, following the formula (*converted_score* is the final score converted from the Likert scale for each question):

$$converted\_score \sim day\_into\_study + (1|subject\_id).$$

Our results show a decreasing trend over time for question 4: "*I feel comfortable wearing the smartwatch*" ($\beta = -.03$, $SE = .02$, $p = .04$). This suggests participants feel **less comfortable** with the smartwatch over time, which somewhat contradicts our initial assumption in *H2*. We found no statistically significant result for the other three questions. We show the distribution of responses for each questions in the burden survey in Figure 4.

Our system also received a high usability score (SUS) of 80.1, suggesting high perceived usability [7]. We present the results from the individual questions of SUS in Appendix C.

6.2 RQ2: Factors affecting modality choice and response rate in multimodal μEMA prompts

In this section, we quantitatively examine how in-the-moment contextual factors affected participants' modality choice and response rate *(H3, H4)*. We discuss our qualitative findings on how varying interruption and interaction burden influence participant's decision making process.

6.2.1 Associations of in-the-moment contextual factors with response rate and modality choice. The distribution of modality usage differs widely across all participants (Fig 5). This variance can be attributed to multiple factors: personal preferences, in-the-moment contextual variables, and prompt interaction/interruption burden.

We used a mixed-effect logistic regression with a random intercept for each participant to predict whether the participant responded to the μEMA prompt (response = 1 vs. non-response = 0). Additionally, we use another mixed-effect logistic regression with a random intercept for each participant to evaluate associations of contexts with modality they opted for (speech = 1 vs. touch = 0). Using the passively collected data, we identified seven in-the-moment contextual variables to use as predictors of non-response and modality choice. These variables include heart rate, wrist movement, location, phone usage, ambiance noises, time of day, and type of day.

Heart rate. We used the heart rate measured by the Pixel Watch 2 at the time of the prompt. If the participants were not wearing the watch at the time of the prompt, we imputed the value with the average heart rate of each participant.

Wrist movement. For wrist movement, we used the AUC unit (as computed in Section 3.4) calculated at the closest time before a prompt. The AUC unit is calculated once every 10 seconds.

Location. The study app recorded GPS data (longitude and latitude) from a participant's personal phone once every minute. We used the DBSCAN clustering algorithm to identify prominent location clusters where participants spent time during the seven-day study. We labeled the cluster the participants were at most frequently during their self-reported sleep time as "Home," and other locations were labeled as "Not Home." We used the location label closest to the time of a prompt as the predictor in our models. "Not home" was the reference variable for the mixed-effect models.

Phone usage. We gathered phone usage data from participants' personal phones at one-minute intervals. "Phone in use" was set to 1 when a prompt appeared with the phone screen on, 0 otherwise.

Ambient noises. The watch passively listened to 10 seconds of audio before each μEMA prompt and used Google's YAMNet audio classification model [30] to determine the ambient noises present

Table 3: Usability metrics for multimodal μEMA. $\#promptsCompleted$ are speech responses that were intelligible to human annotators and touch responses that the participants did not cancel. $\#promptsAnswered$ are prompts participants interacted with (e.g., an audio input that is not intelligible would be answered but not completed). $\#promptsDelivered$ are successfully delivered prompts. $\#promptsScheduled$ are the number of hours between participant's self-reported wake and sleep times, multiplied by 12. Interaction time measures the duration from the first touch interaction to the last touch interaction. Error recovery rate is the rate of touch responses being canceled and replaced by speech over all canceled touch input.

Metric	*Formula*	*All-prompt Mean*	*Between-subject Mean (SD)*
Compliance rate	$\frac{\#promptsAnswered}{\#promptScheduled}$ (%)	63.1	65.6 (21.4)
Response rate	$\frac{\#promptsAnswered}{\#promptDelivered}$ (%)	72.4	74.2 (11.5)
Success rate	$\frac{\#promptsCompleted}{\#promptsAnswered}$ (%)	99.8	99.8 (2.34)
SAME rate	$\frac{\#SAME}{\#promptAnswered}$ (%)	6.01	8.85 (12.9)
SUS scores	—	—	80.1 (11.9)
Error recovery rate	$\frac{\#touchToSpeech}{\#touchCancelled}$ (%)	—	94.8 (1.55)
Interaction time (touch input)	$lastTouch - firstTouch$ (s)	0.3 (0.16)	—

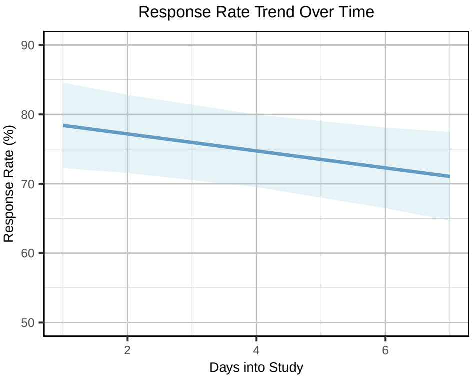


Figure 3: Response rate with regard to day-into-study.

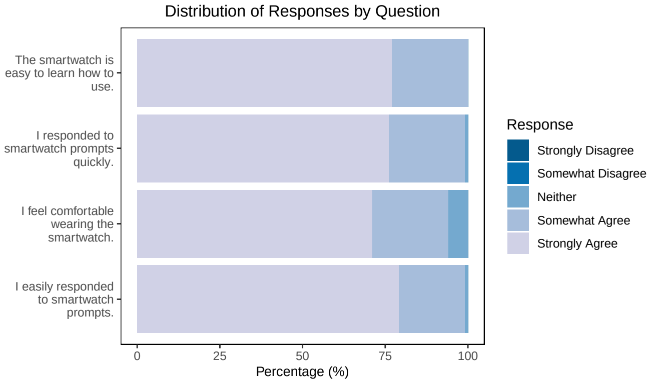


Figure 4: Distribution of responses for individual items in the daily burden survey.

right before a prompt. The model ran locally on the smartwatch, and no raw audio recordings were saved. We categorized the noises into three mutually-exclusive labels: "Speech," "Silence," or "Other noises." If "Speech" was detected during the 10-second period, the ambient noise of the prompt was set to "Speech." If "Silence" was the only noise detected by the YAMNet model, the ambient noise label was set to "Silence." Otherwise, the ambient noise label was set to "Other noises." "Silence" was the reference variable for the mixed-effect models.

Time of day. We converted the 24-hour time of day into four categories: Morning (6 am to 12 pm), Afternoon (12 pm to 6 pm), Evening (6 pm to 12 am), and Night (12 am to 6 am). "Morning" was the reference variable for the mixed-effect models.

Day of week. We converted each day into "weekday" (Mon-Fri) and "weekend" (Sat/Sun). "Weekend" was the reference variable for the mixed-effect models.

Day into study. Day into study ranged from day 1 to day 7.

We show the mixed-effect models below, where *response* is whether the participant responded to a prompt (response vs. non-response), and *modality* is either speech or touch. Results from the mixed-effect models are shown in Table 4.

$$response \sim heart\_rate + location\_ + phone\_usage + ambient\_noise$$
$$+time\_of\_day + day\_of\_week + day\_into\_study + (1|subject\_id). \quad (1)$$

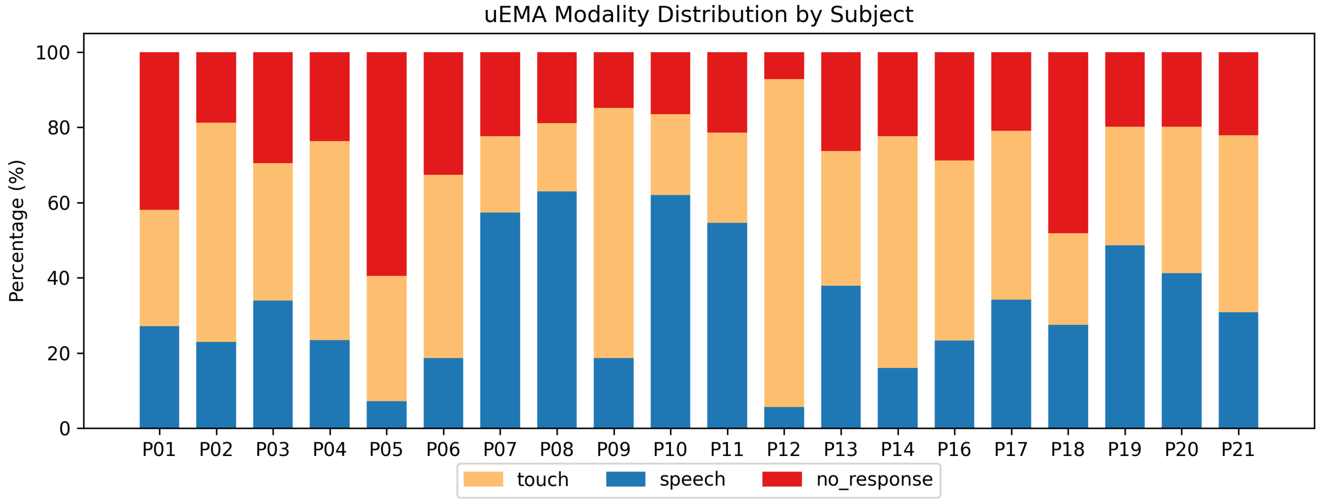


Figure 5: Distribution of modality choice and non-responses for all participants. Even though the ratio of speech to touch input for all participants was relatively balanced (54.7% touch, 45.3% speech), there was large variance between participants. We removed P15 since the participant withdrawn from the study during the consent period.

$$modality \sim heart\_rate + location\_ + phone\_usage + ambient\_noise$$
$$+time\_of\_day + day\_of\_week + day\_into\_study + (1|subject\_id). \quad (2)$$

For modality choice, the main effect of wrist AUC was significant, with $\beta = .19$ ($SE = .02, p < .001$). This reflects an increase in speech interactions under higher wrist movement. The main effect of heart rate was significant, with $\beta = .05$ ($SE = .02, p = .05$). This reflects an increase in speech interactions under higher heart rate. The effect of location is significant, with $\beta = .74$ ($SE = .05, p < .001$). This reflects an 75% increase in speech interactions when participants were at home compared to not at home. The effect of ambient noise (detecting "speech" in the environment) was significant, with $\beta = -.16$ ($SE = .03, p < .001$). This indicates lower speech interactions when detecting speech or conversation noise in the background, compared to silence. The effect of ambient noise (detecting noises other than "speech" in the environment) was significant, with $\beta = .01$ ($SE = .03, p = .02$). This indicates slightly higher chance of speech interactions when detecting noises other than speech or conversation noise in the background, compared to silence. This is because the majority of noises detected in this case are noises coming from activity of the wrist, such as hand washing. Lastly, the effect of day into study was significant, with $\beta = -.11$ ($SE = .02, p < .001$). This reflects a decrease in speech interactions as the study progresses. We observed no statistical significance for other variables.

For response rate, the main effect of heart rate was significant with $\beta = -.26$ ($SE = .09, p = .006$). This reflects a decrease in response under higher heart rate. The effect of location is significant ("Home"), with $\beta = .34$ ($SE = .16, p = .03$). This reflects an increase in response when participants were at home. The effect of afternoon $\beta = -.23$ ($SE = .1, p = .05$) and evening $\beta = -.36$ ($SE = .2, p = .001$) were significant, indicating that participants were more likely to respond to the prompt in the morning compared to later in the day. We observed no statistical significance for wrist AUC, phone usage, detecting speech or other noises in the background, weekday and during night time.

6.2.2 *How contextual variables and modality choice influences interruption and interaction burden of μEMA prompts.*

Based on the thematic analysis, we identified two major burdens associated with our system (Table 5): the **interruption burden** and the **interaction burden**. We further identified two different sub-categories associated with the interruption burden (*cognitive* and *social* burden) and three associated with interaction burden (*physical*, *cognitive*, and *social* burden) (Table 5).

Interruption burden refers to the burden produced by the μEMA reporting cue. While the interruption burden could remain the same for both modalities of μEMA, the level of burden depends on the in-the-moment context of the participants. There are two different aspects to the interruption burden. The *cognitive* interruption burden (`cognitive interruption`) occurs when participants receive a prompt during a cognitively engaging activity: (e.g.,"*I am in complete zone like for most of the [computer] games. For an example like when crucial [moments] and the watch vibrate, I actually get irritated.*" [P14]). The *social* interruption occurs when participants are in a public setting or around other people, and the prompts break the flow of the conversation or draw attention in a quiet public space (`disturbing others`).

Interaction burden refers to the burden perceived when participants interact with the μEMA system. The interaction burden can vary based on the participants' contextual state and modality choice. Participants experienced a *physical* interaction burden when they brought their hand close to their mouth when speaking to the watch, or when they used one hand to tap/draw on the watch face during

Table 4: Associations of contextual variables with modality choice and prompt response. We used a mixed-effects logistic regression with random intercept for both experiments. In the touch vs. speech model, we set "touch" to be the reference factor. In the response vs. non-response model, we set "non-response" to be the reference factor. Coefficient converted to an odds ratio (OR) shows how much the odds of an outcome change with a one-unit increase in the predictor, where OR > 1 means higher odds, and OR < 1 means lower odds. ** p-value < .001, * p-value < .05.

Variables		P_{speech}		$P_{responded}$	
		OR	95% CI	OR	95% CI
Heart rate		1.05 ** ▲	1.0-1.11	0.76 ** ▼	0.63-0.93
Movements	Wrist AUC	1.22 ** ▲	1.15-1.28	1.0	0.85-1.17
Location	Home	2.05 ** ▲	1.84-2.28	1.41 * ▲	1.01-1.98
Phone usage	Phone in use	1.06	0.95-1.17	1.13	0.8-1.6
Ambience noises	Speech	0.83 ** ▼	0.78-0.88	0.91	0.76-1.11
	Other noises	1.07 * ▲	1.01-1.13	1.0	0.83-1.22
Time of day	Afternoon	0.96	0.86-1.07	0.79 ** ▼	0.53-1.16
	Evening	0.93	0.81-1.05	0.69 ** ▼	0.45-1.06
	Night	2.04	0.57-7.34	0.03	0-inf
Day of week	Weekday	0.97	0.87-1.08	1.17	0.84-1.62
Day into study		0.89 ** ▼	0.84-1.93	0.83 ** ▼	0.74-0.92

a touch interaction. Participants commented that the physical interaction burden tended to be higher for touch input when they were moving (`movement/activity`), or when their hands were busy (`hand availability`) ("*since [...] my hands are busy, I won't tap*" [P1]). Participants' `reactivity` to the prompt can cause unnecessary physical interaction burden, which might warp the perceived interaction time of the prompt. Interestingly, we found that participants disagreed over the perceived interaction time between the speech and touch input. Six participants mentioned that speech is noticeably faster ("*I feel it's [speech] a faster process.*" [P20]). Yet, five participants believed one-tap interaction was faster ("*I click the button I'm standing or walking most of the time, so it's faster.*" [P14]), since they already have an intuition to look at the watch's notifications ("*when it [the watch] vibrates* [sic] *, it's human nature to then see the watch* " [P14]). The intuition of looking at the watch screen or bringing the watch close to the mouth when a prompt occurs can influence participants' perception of how long the interaction takes, which could affect modality choice. The *cognitive* interaction burden refers to the mental effort needed to respond to the prompt. Participants raised two major cognitive interaction burdens with μEMA. The first was their `mental bias/uncertainty` about what to report ("*I do agree that having the options [on the screen] and responding to that lessens the cognitive load because then I'm like, oh, these are my options.*" [P10]) and the quality of their self-report ("*I felt like what I said may not be caught by the system itself [because of] my accent cannot be caught by the system. That could potentially lead to a mislabeling.*"[P12]). The second cognitive interaction burden was the `repetition fatigue` produced by consecutively giving the same labels over a long period of time. While our daily burden survey did not capture the interruption burden from prompt frequency, our qualitative evaluation from the exit interviews show that 11 (55%) participants reported the repetitive labeling over time as the primary source of burden (not the prompt frequency), seven (35%) expressed a desire for longer intervals between prompts, and two (10%) reported no concerns about prompt frequency. This implies that the repetition fatigue could be related to (or potentially caused by) the intense prompting interval of our study ("*You have to give multiple answers of the same activity [...] For example, you are using your laptop for an hour straight, then once every 5 minutes, you have to give [...] 10 similar prompts*" [P8]). Finally, the *social* interaction cost refers to their `social discomfort` of the interaction, such as being considered rude to interact with the watch during a conversation or a meeting ("*In public, I feel like when you look at the watch and tap, it's sort of rude. Like it looks like you're responding to a message or something.*" [P8]; "*You know close [intimate] conversations, at that time I will tap instead of speak.*" [P1]).

6.3 RQ3: Examine the characteristics and potential utility of labels collected using multimodal μEMA

We collected 11,320 labels from the field study. For the speech self-reports, we manually listened to all audio recordings collected from the μEMA prompts and extracted the posture, activity, and context labels from each self-report. We grouped each self-report into one of the five (not mutually exclusive) categories:

- *Singleton posture:* self-report only containing posture label — no activity included
- *Singleton activity:* self-report only containing activity label — no posture included
- *Posture and activity:* self-report containing both posture and activity

Table 5: Factors influencing participants' response rate and modality choice. The full definition of the codes are available in Appendix D.

Type of burden			*Excerpts from exit interviews*
Interruption	Social	`Disturbing others`	*[P20] "When there are classes, it could be disturbing during class for others nearby, like sitting next to you, [. . .] probably the professor would also not be, you know, like willing to have a vibration during class."*
	Cognitive	`Cognitive interruption`	*[P21] "I used to try and just mute it in the class because it was just that I wanted to concentrate there. So every five minutes it's not just productive to [respond] between the lecture."*
Interaction	Physical	`Hand availability`	*[P16] "Whenever, let's say I'm cooking at that time my hands are dirty, or if I'm cleaning or something like that, then my hands are dirty. So a vibration just occurs on my hand and I would speak to the prompt like 'standing cooking.'"*
		`Movement/activity`	*[P7] "I think that plays into it for sure, like walking the dog or doing dishes or, you know, playing with the dog out in the yard or something like that. If I was up moving around then I would. It's easier. It's certainly a lot easier to respond verbally."*
		`Reactivity`	*[P10] "Intuitive like having a smartwatch and when you have notifications and you just want to look at the screen or respond or see what the notification is."*
	Cognitive	`Mental bias/uncertainty`	*[P2] "Oh, I'm concerned over the data variety basically. So I think from the start if I say like the confusion was like should I say a very detailed approach of like what I'm doing or the type option of sitting or something like that."* *[P12] "When I try to tap something wrong, I was told that if I speak to it before some bits, it might capture the second thing. So I did that, but I don't know which one it captured exactly. So there's no confirmation."*
		`Repetition fatigue`	*[P8] "Then you just feel a bit silly because you just say the same thing over and over."*
	Social	`Social discomfort`	*[P20] "If I'm talking to you, I cannot immediately break your flow or mine [to respond to the prompts]"*

- *Multiple activities:* self-report containing multiple activities — may or may not contain posture label
- *Context included:* self-report including contextual information (e.g. location)

We present the distributions of categorized self-reports in Table 6. Because μEMA is limited to a single-question, single-tap response, participants can only report activity *or* posture during touch input, not both. Quantitative results from the field study show that participants tended to choose to report posture over activity in touch responses (of all touch activity labels, 41% of them are "Walking"). From the qualitative findings, we identified three major reasons for this tendency. First, participants might perceive posture labels as more helpful to the researchers (`mental bias/uncertainty`) ("*I think the posture is given first priority more than the activity, right? Because that's how the audio cues [instructions for the speech input] are. First is the posture then activity.*" [P17]). Second, participants expressed hesitation in searching for activity because there was no guarantee that the label was present in the system (`mental bias/uncertainty`) ("*Regarding writing, the main issue was that I didn't know what are the exact all categories of things that's available to me.*" [P14]). Finally, participants reported tapping on whatever options were closest to their activity/posture that appeared on the first screen and could be answered with a simple tap (`reactivity`) ("*When you're in a hurry, you just look at the screen and then tap on whatever most relevant*" [P16]). These findings show that the interaction cost affects compliance, modality choice, *and* the content of participants' responses.

To further understand the characteristics of the labels collected from the field study, we manually categorized each label into 10 high-level categories. Additionally, for each self-report, we labeled it as either "macro-label" or "micro-label." "Macro-labels" are high-level posture and activity labels (e.g., "sitting," "standing," "cooking," "cleaning," "grooming"). "Micro-labels" are either 1) macro-labels with more context that could potentially influence the sensing signal (e.g., "lounging" (a form of "sitting"), "walking and carrying groceries") or 2) a micro-activity of a "macro-label" (e.g., "chopping vegetables" vs. "cooking," "applying lotion" vs. "grooming"). Different HAR datasets/models might focus on detecting or collecting data about macro-labels [56, 93, 115] or micro-labels [1, 2, 15, 100]. Micro-labels can also potentially help debug HAR models, by providing additional contexts that could influence the signal quality

Table 6: Distribution of self-reports (% (count)) categorized into five categories based on their contents/labels, sorted by modality. Participants can only report either posture or activity in the touch input.

Modality	*Singleton posture*	*Singleton activity*	*Posture and activity*	*Multiple activities*	*Context included*
touch	59.5% (2980)	39.5% (2130)	—	—	—
speech	19.6% (733)	19.4% (723)	46% (1717)	2.2% (84)	6.5% (241)

(e.g. "standing" vs. "standing and washing dishes"). Table 7 shows the distribution of reported macro-labels and micro-labels.

We categorized the self-reports after completing the data collection period. The categorization process required significant time to parse through and consolidate the list of labels. In HAR data collection studies, however, annotation schemes are likely determined beforehand. After-the-fact label cleaning does not enable real-time feedback or real-time training/monitoring of HAR models. To show the potential for automatic real-time label extraction, we report the results from two tasks: 1) automatic speech recognition (ASR) on the audio recordings collected from the field study, and 2) automatic mapping from users' self-reports to a pre-determined list of ADL labels used by two publicly available HAR datasets.

For the ASR tasks, we customized a commercial ASR model, Google Cloud speech-to-text v1. We used model adaptation to improve the accuracy of the model and tune the model to recognize targeted word/phrases (e.g., "sitting" is recognized more often than "setting" or "city")[3]. The watch sent the audio recordings to Firebase and retrieved the transcription on the watch. Due to the unpredictability of network/cellular connections, the ASR process was not guaranteed to be completed in real-time. If there was no network/cellular connection on the watch, the system would stop transcribing to avoid battery drain. To evaluate the usability of ASR, we followed a similar evaluation process as the prior work on audio-μEMA [59]. A human annotator transcribed the audio recording manually to extract the posture/physical activity labels from the recordings. If all posture/activity/context labels were presented in the ASR output, that result was classified as a "correct" ASR transcription. On average, the accuracy of the ASR was 85.9% (SD = 5.9 between subjects), significantly higher than the accuracy observed in prior work on audio-μEMA that used off-the-shelf ASR (20-25% accuracy) [59]. Although this can be attributed largely to the changes made to the ASR model, participants' bias and uncertainty about the audio quality could potentially lead to overall better audio recordings collected (e.g. participants not using speech input in noisy environment, or participants with heavy accents opting to use touch input).

To evaluate the potential for automatic mapping of labels to a target set of labels, we used an open-source large language model (LLM), llama3-8b [105], and prompted it to map the participants' self-report open-ended labels to two different label lists used by large ADL HAR datasets. By leveraging the LLM's embedded common sense reasoning about relationship between concepts, we hope it can manage the variability in participants' self-reports and improve the mapping to structured label sets [80]. CAPTURE-24 is a large scale wrist-worn accelerometer activity dataset collected in-the-wild [14]. Pirsiavash and Ramanan (P&R) is an egocentric camera activity dataset collected in a lab-based setting [81]. The label list used in CAPTURE-24 consists primarily of high-level (macro-labels), while the annotation scheme in the P&R dataset is more detailed and descriptive (micro-labels) [4]. The P&R dataset, however, only contains home-based activities, while the CAPTURE-24 annotation scheme covers activities outside the home (e.g., "vehicle," "walking"). If the dataset label list did not contain basic postures ("sitting," "standing," "kneeling," "bending over," "lying") or "walking," we added those labels to the label list. We also added "other" as a category. We include the list of labels in both annotation schemes we used in our experiments in Table 8.

For each self-report collected from the participants, we ran a prompt through llama-3 (see Appendix B) to obtain the mapping of the raw self-report response to the respective annotation scheme. One research team member went through the same process manually and we compared the results to that of the LLM. Another research team member went through the same mapping process on a subset of the labels (400/11,320; 3.5%). The inter-rater reliability rate (Cohen's kappa) between the two annotators was $\kappa = 0.99$, which indicates substantial agreement between annotators. We identified three common mistakes made by the LLM:

- *Wrong mappings* refer to obvious mistakes made by the LLM (e.g. "standing" is mapped to a "walking" label).
- *Inferences* refer to instances when the LLM tries to infer the mapping from the self-report (e.g., "doing homework" is mapped to "sitting+using computer" (P&R)).
- *Made-up labels* refer to instances when the LLM mapped the self-report to a non-existant label (e.g., "grooming").

Figure 7 shows the distribution of correct mappings and mistakes made by the LLM in the automatic label mapping task. Compared to a human annotator, the LLM has an accuracy of 78.1% and 67.2% for the two annotation schemes. We found that that the distribution of errors between two data set is significantly different ($\chi^2(3) = 46.5, p < .001$). We noticed that the LLM made up significantly more labels using the P&R label list than the CAPTURE-24 list (8.6% vs. 0.8% made-up labels). We believe this was because the P&R label list only contains home activities, so the LLM made up new labels (hallucinations) for self-reports of activities outside the home. For both annotation schemes, the LLM was able to make logical inferences (9.9% and 15.6% inference errors). Even though the inferences might not be always correct in naturalistic settings (e.g., "using computer" might not always associated with "sitting"), these inferences can be useful for future system designs that allow follow-up questions.

[3] https://cloud.google.com/speech-to-text/ondevice/docs/model_adaptation

[4] In this experiment, we used the mapped labels that the authors of CAPTURE-24 used when training/evaluating HAR models [14].

Table 7: The distribution of reported macro-labels and micro-labels.

Category	*Macro-labels*	*Micro-labels*
POSTURE (5430)	sitting (3842), standing (1193), bend over (14), kneeling (3), crouching (7), lying (368)	lounging (1), hunching (3), ...
IN TRANSIT (1069)	walking (859), stairs (10), traveling (1), driving (26), ...	going to station (1), riding train (3), going to the station (1), riding electric scooter (1), biking (2), talking/driving (1), waiting for the bus (1), ride plane (1), ...
CHORES (264)	cooking (178), do chores (13), doing laundry (4), cleaning (47), ...	folding stuff (1), carry stuff (2), dusting (1), washing stuff (1), walking/cleaning kitchen (1), putting food away (1), pick up stuff (1), watering plants (2), ...
WORK/SCHOOL (279)	writing (6), doing work (2), meeting (152), working (84), studying (7)	using whiteboard (1), using the whiteboard (1), conversing/writing on whiteboard (1), writing on whiteboard (1), doing an assignment (1), ...
FOOD/DRINK CONSUMPTION (520)	drinking (13), eating (170)	walking/eating (1), eating dinner (2), eating breakfast (2), chopping vegetables (1), drinking water (13), ...
HOBBY (176)	doing crafts (2), shopping (32), reading (35), playing game (31), ...	playing guitar (3), playing frisbee (1), playing video games (15), reading book (18), ...
USE ELECTRONICS (1744)	using computer (94), using phone (85), watch contents (26), use tablet (12), ...	watch a movie (1), working on computer program (1), watch tv (67), use phone/waiting for train (1), typing (50), ...
SELF-CARE (87)	use bathroom (9), grooming (2), ...	wash hands (4), showering (4), washing hands (9), brush teeth (2), taking off socks (1), combing hair (1), wiping face (1), skincare (1), ...
PUTTER AROUND (34)	wake up (1), get ready to walk (1), ready for sleep (2), ...	scratching head (1), locking door (1), climb down ladder (1), charging tablet (2), opening drawer (1), unlocking door (1), ordering food (1), ...
LOCATION (93)	at park (4), around campus (3), to the bus (1), to college (2), in the kitchen (2), around the house (2), in the office (2), outdoors (1), bus connection (1), ...	

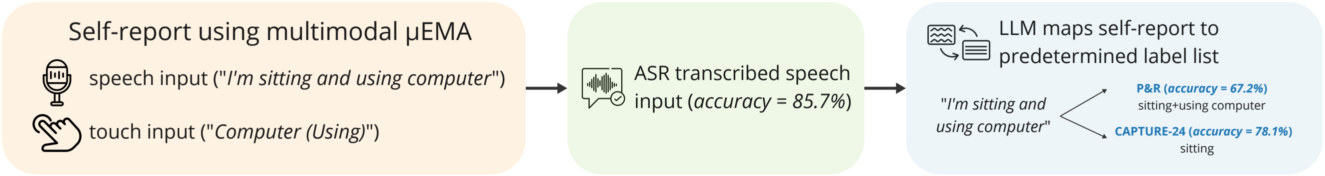

Self-report using multimodal µEMA
speech input ("I'm sitting and using computer")
touch input ("Computer (Using)")
ASR transcribed speech input (accuracy = 85.7%)
LLM maps self-report to predetermined label list
"I'm sitting and using computer"
P&R (accuracy = 67.2%)
sitting+using computer
CAPTURE-24 (accuracy = 78.1%)
sitting

Figure 6: Our proposed pipeline for real-time automatic label extraction.

7 DISCUSSIONS AND FUTURE WORK

We have extended the line of work in μEMA and audio-μEMA by combining two input modalities, touch and speech, into a novel type of μEMA: multimodal-μEMA. In this section, we discuss the implications of this work for future deployments of multimodal-μEMA; design implications for context-aware, multimodal prompting systems; and how our method can be used in a real-time human-in-the-loop activity recognition system.

7.1 Usability and Challenges of Deploying Multimodal μEMA in-the-wild (RQ1)

At 15-minute prompting intervals, prior works with touch-only μEMA on a smartwatch reported response rates of 80-90% [43, 85], and speech-only μEMA reported a response rate of 85-90% [38]. Recent work on speech-only μEMA, at a five-minute prompting interval, demonstrated response rates of 65-68% [59]. Despite disrupting participants approximately three times more than standard μEMA, our field deployment showed that participants were able to respond to our proposed multimodal μEMA prompts with a high response rate of 72.4%. Compared to EMA studies of similar duration (7 days) [111], which report an average response rate of 79%, we achieved a comparable level of participant engagement while delivering 10–20 times more prompts (6 vs. 80–120 prompts per day). This shows the promise of using multimodal μEMA implementation on the smartwatch to collect data at high temporal density while maintaining a good response rate.

The majority of participants (55%) expressed it was not the frequency of the prompts, but the repetition of reporting the same labels and the mental load to come up with a response that posed

Table 8: List of labels in datasets used for the automatic self-report to label mapping. Labels with an asterisk (*) were added by our research team.

CAPTURE-24 [14]	*Pirsiavash and Ramanan* [81]
sleep, sitting, standing, lying*, kneeling*, bend over*, household-chores, manual-work, walking, mixed-activity, vehicle, sports, bicycling, others*	sitting*,standing*, lying*, kneeling*, bend-over*, walking*, combing hair, make up, brushing teeth, dental floss, washing hands/face, drying hands/face', enter/leave room, adjusting thermostat, laundry, washing dishes, moving dishes, making tea, making coffee, drinking water/bottle, drinking water/tap, making hot food, making cold food/snack, eating food/snack, mopping in kitchen, vacuuming, taking pills, watching tv, using computer, using cell, making bed, cleaning house, reading book, using mouth wash, writing, putting on shoes/socks, drinking coffee/tea, grabbing water from tap, other*

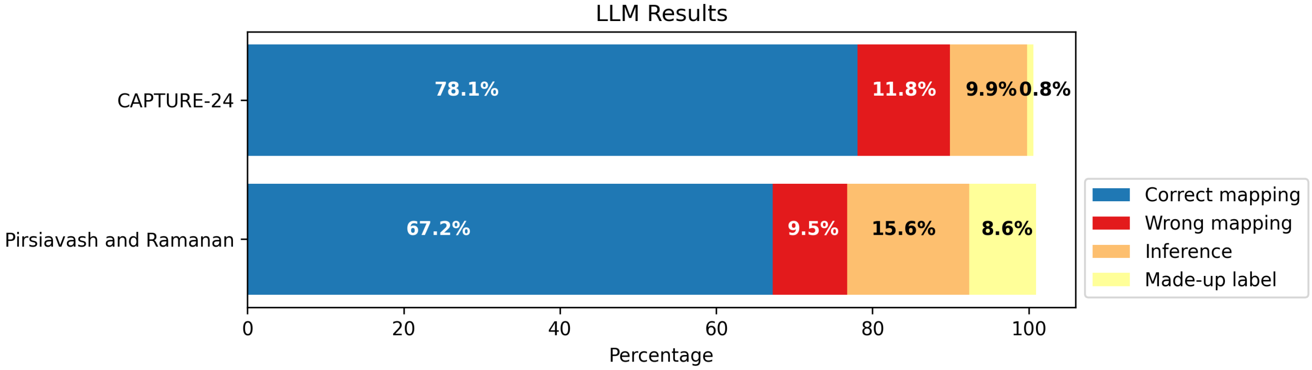


Figure 7: Distribution of correct mappings and mistakes made by the LLM.

significant cognitive burden [59, 65]. This suggests the possibility of a transition-based prompting mechanism where the system only prompts participants when it detects a change in movement/activity [5, 40, 62, 63, 76]. However, determining the optimal moment to prompt while balancing researchers' information needs [62] and participants' burden [71] requires further investigation.

Another major usability challenge we observed during our study was participants' reactivity to the watch prompt. While in prior work Ruan et al. showed that speech input tends to be faster than touch input [89], our qualitative findings show that participants disagreed about which interaction is faster – speech or touch. Many participants expressed that their ingrained automatic response to a watch prompt was to look at or touch the watch face, even in speech interaction where they were instructed not to. Training people out of this habit is difficult. One possible solution could be to design distinct haptic/auditory cues for the different μEMA prompts. Prior researchers have studied different types of haptic cues and tying them to specific actions/messages [3, 86]. Another option would be to add a new device form factor that *only* has an audio interface and has become quite popular – earables [91]; these devices may prevent participants from looking at the watch screen. Several prior studies have successfully deployed speech-based EMA on earbuds or headsets [9, 59]. Adding additional on-body devices that are not yet socially acceptable in all situations, however, could increase perceived burden by drawing unwanted attention from bystanders.

7.2 Disconnection between passive sensing data and perceived source of burden (RQ2)

From the quantitative analysis of our study (*H*4), we observed that higher heart rate, higher wrist movement, and location (at home) have a positive association with the use of speech interaction. Furthermore, presence of speech in the background has a negative association with speech interaction. These findings indicate that people were comfortable using the speech interaction while undergoing high physical activity and movement. Participants were, however, less likely to use the speech interaction if there was background speech (suggesting presence of other people). Our qualitative analyses further support these observations, where participants reported speech input having a lower physical burden, but higher social and cognitive burden compared to touch input. Furthermore, we found that higher heart rate, higher wrist movement, phone usage, weekday, and detecting noise aside from speech in the background are associated with higher μEMA response rate. Participants also are more likely to respond to the prompt in the morning compared to later during the day. These findings are consistent with prior work on contextual biases in μEMA non-response [72, 83].

We want to highlight, however, that there is still a disconnect between the passive sensing data and the interaction burdens raised by our participants [21, 35]. This semantic gap can hinder the deployment of future context-aware system that attempt to predict participants' receptivity to μEMA prompt [17, 33, 60, 63, 71]. The contextual variables used in our quantitative experiments are often proxies to detect physical and social burden (e.g. location and ambient noises can indicate potential `social discomfort`, phone usage might indicate `hand availability`, wrist movement can indicate `movement/activity`). These contextual variables, however, do not reveal information about potential cognitive interaction/interruption burden, or all situations where hands were unavailable (e.g., "carrying/moving stuff" or "typing"). Hand availability and repetition fatigue might be detected using fine-grained activity recognition models, and using physiological signals to predict `cognitive interruption` burden [31]. Burden may exist on a spectrum (e.g., a work meeting likely imposes a higher social burden than a casual conversation with a friend), which passive sensing data may not fully capture. Future systems might automatically measure or estimate burden and explore how it affects participants' decision-making. Future work should also investigate whether a threshold exists at which burden significantly influences participants' modality choices.

7.3 Towards real-time human-in-the-loop activity recognition systems (RQ3)

Our proposed system, multimodal μEMA, could be useful for real-time annotation and training of activity recognition models [16, 22, 70, 73]. Its flexible input modalities and open-ended responses permit participants to naturally enter what they are doing via speech in a way that allows researchers to gather temporally dense and high-quality labels and subsequently use them to define and tune personalized HAR models. We explored the characteristics of the labels for the behavior labeling task and the possibility of future automation of label extraction. Our findings show that participants were able to report various activity labels under different contexts using our system. We also show that by tuning a commercial ASR and using an open-source LLM we can build a pipeline for automatic label extraction. The LLM performance, however, depends on the list of labels defined by the researchers. Furthermore, participants may have their own mental biases regarding what qualifies as an "activity" or how to search for one using touch input, highlighting the need to co-develop inclusive activity labeling schemes and feedback mechanisms to guide participants in providing useful information (which can also reduce their mental load). In this work, we demonstrate viability of mapping what people report using multimodal μEMA onto desired labels using zero-shot prompting of the LLM. Future work might improve on this method by fine-tuning an LLM model to increase accuracy or by including a re-prompting mechanism to avoid the LLM making up random labels beyond the pre-defined corpus [23].

It is important to note that participants' mental biases could lead to lower compliance or undesirable reporting patterns for speech interaction (such as overly verbose self-reports or avoiding using speech input), resulting in lower quality of the labels. A real-time feedback loop to combat this bias would depend on the accuracy of the ASR model. Despite the successful use of model adaptation to increase ASR accuracy (86% accuracy), there is still a lot of room for improvement, and the system still requires human supervision to extract all the labels post-hoc. Furthermore, in our current implementation, the ASR model needs a network connection to run, and the network is not always available. We want to emphasize that adding the same confirmation screen used for touch input may not be the optimal solution. Even with a stable network to run the ASR model, the latency of ASR could be too long for the interaction to be considered a "microinteraction." If the ASR misrecognizes input, it can further increase participant frustration. Instead of using in-the-moment feedback, an end-of-day summary report [42] or making the screen disappear upon detecting the "end-of-speech" could serve as sufficient confirmation mechanism. Future works could further explore the effectiveness of different feedback mechanisms.

8 LIMITATIONS

There are some limitations of this work. Many EMA studies only run for a week [20, 97], and thus the seven-day study results reported here provide a baseline for multimodal μEMA use. In this pilot study, a member of the research team checked in daily with participants using text messages. These daily messages built rapport with participants in this short study period, which can increase response rate and compliance with the system [29, 74, 107]. Future studies should assess longer-term, and even longitudinal, use and compliance. The second limitation is that our study populations is skewed towards young male adults who are facile with technology (55% participants self-reported they were very familiar with and regularly used tracking applications on a smartwatch). Our sample size also skewed towards non-Hispanic Asian students, which limits the generalizability of our findings. Future research should study reactions to multimodal μEMA among other groups. A another limitation of this work resulted from hardware restrictions of the Pixel Watch 2; due to the computationally-intensive on-device audio processing and five-minute prompting intervals, the smartwatches running software had a battery life of 12-13 hours before a recharge was required. Future studies could turn off the ambient noise classifier to extend battery life. Furthermore, newer watches continue to have better battery capacity [5]. Due to a restriction from Google, our watch application required a network connection to run the ASR model; this requirement prevented us from implementing real-time feedback to participants during a speech input. We did not find a significant quantitative association between perceived burden and the day in study (*H2*). This suggests participants might sustain our system over time without increased burden, though our use of unvalidated burden scale may have influenced the results. Future work should look into using other additional measurement for burden/workload, for example the user burden scale or NASA-TLX scale [90, 102]. Our qualitative findings suggest that desire for feedback influenced modality preferences for some participants. A useful future extension to the system might add feedback to the speech responses, either in real-time or via an end-of-day report. Finally, due to the scope of this paper, we did not assess the validity of the labels collected by the system. Given our promising feasibilty

[5] https://blog.google/products/pixel/google-pixel-watch-3/

results, in future work we seek to investigate the usability and validity of the annotations collected using μEMA system, either through an egocentric camera, or using data from the watch sensors [10, 32, 84, 109].

9 CONCLUSION

In this paper, we present a novel data collection method, multimodal μEMA, by combining speech and touch input on a smartwatch. We conducted a seven-day free-living study and examined the usability and feasibility of an in-the-wild deployment of our system. Despite the temporal density of the prompts (once every 5 minutes), participants were highly engaged with our system, with an average response rate of 72.4%. We quantitatively and qualitatively identified different factors affecting response rate and modality choice, investigating the characteristics and usefulness of the labels recorded from the field study study. Our field deployment shows the potential of leveraging multimodal μEMA for collecting useful, rich posture and physical activity labels, which can potentially be integrated within a real-time activity recognition system.

Acknowledgments

Research reported in this publication was supported, in part, by the National Cancer Institute of the National Institutes of Health under award number R01CA252966. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Ali Abbas, Michael Haslgrübler, Abdul Mannan Dogar, and Alois Ferscha. 2021. Micro activities recognition in uncontrolled environments. *Applied Sciences* 11, 21 (Nov. 2021), 10327. doi:10.3390/app112110327

[2] Mohammed A. A. Al-qaness. 2019. Device-free human micro-activity recognition method using WiFi signals. *Geo-spatial Information Science* 22, 2 (April 2019), 128–137. doi:10.1080/10095020.2019.1612600

[3] Stephen A. Alexander, Roderico Garcia, and Marcia K. O'Malley. 2021. Enhancing multi-sensory cue salience and perceptual identification in a wearable haptic device. In *2021 IEEE World Haptics Conference (WHC)*. IEEE, 900–905. doi:10.1109/WHC49131.2021.9517130

[4] Pengcheng An, Jiawen Zhu, Zibo Zhang, Yifei Yin, Qingyuan Ma, Che Yan, Linghao Du, and Jian Zhao. 2024. EmoWear: Exploring emotional teasers for voice message interaction on smartwatches. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3613904.3642101

[5] Christoph Anderson, Clara Heissler, Sandra Ohly, and Klaus David. 2016. Assessment of social roles for interruption management: A new concept in the field of interruptibility. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1530–1535. doi:10.1145/2968219.2968544

[6] Gentry Atkinson and Vangelis Metsis. 2020. Identifying label noise in time-series datasets. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. ACM, New York, NY, USA, 238–243. doi:10.1145/3410530.3414366

[7] Aaron Bangor. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies* 4, 3 (2009), 114–123. doi:10.5555/2835587.2835589

[8] Ling Bao and Stephen S. Intille. 2004. Activity recognition from user-annotated acceleration data. In *Pervasive Computing*, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Alois Ferscha, and Friedemann Mattern (Eds.). Vol. 3001. Springer Berlin Heidelberg, 1–17. doi:10.1007/978-3-540-24646-6_1 Series Title: Lecture Notes in Computer Science.

[9] Tao Bi, Temitayo Olugbade, Akhil Mathur, Catherine Holloway, Aneesha Singh, Enrico Costanza, and Nadia Berthouze. 2022. A taxonomy of noise in voice self-reports while running. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Cambridge United Kingdom, 229–232. doi:10.1145/3544793.3563421

[10] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. 2024. WEAR: An outdoor sports dataset for wearable and egocentric activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (Nov. 2024), 1–21. doi:10.1145/3699776

[11] Mehdi Boukhechba, Lihua Cai, Philip I. Chow, Karl Fua, Matthew S. Gerber, Bethany A. Teachman, and Laura E. Barnes. 2018. Contextual analysis to understand compliance with smartphone-based ecological momentary assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, New York NY USA, 232–238. doi:10.1145/3240925.3240967

[12] John Brooke. 1996. SUS - A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).

[13] Min Chul Cha and Yong Gu Ji. 2023. Context matters: Understanding the effect of usage contexts on users' modality selection in multimodal systems. *International Journal of Human–Computer Interaction* (Aug. 2023), 1–16. doi:10.1080/10447318.2023.2250606

[14] Shing Chan, Hang Yuan, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. 2024. CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. http://arxiv.org/abs/2402.19229 arXiv:2402.19229 [cs].

[15] Soumyajit Chatterjee, Bivas Mitra, and Sandip Chakraborty. 2022. AmicroN: Framework for generating micro-activity annotations for human activity recognition. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, Helsinki, Finland, 26–31. doi:10.1109/SMARTCOMP55677.2022.00019

[16] Heng-Tze Cheng, Feng-Tso Sun, Martin Griss, Paul Davis, Jianguo Li, and Di You. 2013. NuActiv: Recognizing unseen new activities using semantic attribute-based learning. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 361–374. doi:10.1145/2462456.2464438

[17] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-stage receptivity model for mobile just-in-time health Intervention. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (June 2019), 1–26. doi:10.1145/3328910

[18] Mathias Ciliberto, Vitor Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. 2021. Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science* 3 (Dec. 2021), 792065. doi:10.3389/fcomp.2021.792065

[19] S. Consolvo and M. Walker. 2003. Using the experience sampling method to evaluate Ubicomp applications. *IEEE Pervasive Computing* 2, 2 (April 2003), 24–31. doi:10.1109/MPRV.2003.1203750

[20] Delphine S. Courvoisier, Michael Eid, and Tanja Lischetzke. 2012. Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics. *Psychological Assessment* 24, 3 (Sept. 2012), 713–720. doi:10.1037/a0026733

[21] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. 2022. Semantic gap in predicting mental wellbeing through passive sensing. In *CHI Conference on Human Factors in Computing Systems*. ACM, 1–16. doi:10.1145/3491102.3502037

[22] Shizhuo Deng, Chuangui Yang, Zhubao Guo, Boqian Lin, Dongyue Chen, Tong Jia, and Botao Wang. 2023. Fast personalized human activity recognition on heuristic parameter estimation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 606–611. doi:10.1109/ICME55011.2023.00110

[23] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 5, 3 (March 2023), 220–235. doi:10.1038/s42256-023-00626-4

[24] Aiden R. Doherty, Niamh Caprani, Ciarán Ó Conaire, Vaiva Kalnikaite, Cathal Gurrin, Alan F. Smeaton, and Noel E. O'Connor. 2011. Passively recognising human activities through lifelogging. *Computers in Human Behavior* 27, 5 (Sept. 2011), 1948–1958. doi:10.1016/j.chb.2011.05.002

[25] Genevieve Fridlund Dunton, Yue Liao, Keito Kawabata, and Stephen Intille. 2012. Momentary assessment of adults' physical activity and sedentary behavior: Feasibility and validity. *Frontiers in Psychology* 3 (2012). doi:10.3389/fpsyg.2012.00260

[26] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of Advanced Nursing* 62, 1 (April 2008), 107–115. doi:10.1111/j.1365-2648.2007.04569.x

[27] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. 2015. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 10. doi:10.1145/2702123.2702154

[28] Elliot G. Mitchell, Elizabeth M. Heitkemper, Marissa Burgermaster, Matthew E. Levine, Yishen Miao, Maria L. Hwang, Pooja M. Desai, Andrea Cassells, Jonathan N. Tobin, Esteban G. Tabak, David J. Albers, Arlene M. Smaldone, and Lena Mamykina. 2021. From reflection to action: Combining machine learning with expert knowledge for nutrition goal recommendations. In *Proceedings of*

the 2021 CHI Conference on Human Factors in Computing Systems. ACM, 1–17. doi:10.1145/3411764.3445555

[29] Helge Giese and Laura M König. 2024. The impact of incentivization on recruitment, retention, data quality, and participant characteristics in Ecological Momentary Assessments: Experimental study. *Journal of Trial and Error* (2024). doi:10.36850/28b4-4f59

[30] Google Research. 2017. YAMNet: Yet another multilabel neural network for audio event recognition. https://www.tensorflow.org/hub/tutorials/yamnet

[31] Nitesh Goyal and Susan R. Fussell. 2017. Intelligent interruption management using electro dermal activity based physiological sensor for collaborative sensemaking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 1–21. doi:10.1145/3130917

[32] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4D: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 18973–18990. doi:10.1109/CVPR52688.2022.01842

[33] Wayne D. Gray and Deborah A. Boehm-Davis. 2000. Milliseconds matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied* 6, 4 (2000), 322–335. doi:10.1037/1076-898X.6.4.322

[34] Cathal Gurrin, Zhengwei Qiu, Mark Hughes, Niamh Caprani, Aiden R. Doherty, Steve E. Hodges, and Alan F. Smeaton. 2013. The smartphone as a platform for wearable cameras in health research. *American Journal of Preventive Medicine* 44, 3 (March 2013), 308–313. doi:10.1016/j.amepre.2012.11.010

[35] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom. 2006. Mind the gap: Another look at the problem of the semantic gap in image retrieval, Edward Y. Chang, Alan Hanjalic, and Nicu Sebe (Eds.). 607309. doi:10.1117/12.647755

[36] Sourav Bhattacharya Henrik Blunck. 2015. Heterogeneity activity recognition [dataset]. doi:10.24432/C5689X

[37] Stephen D. Herrmann, Erik A. Willis, Barbara E. Ainsworth, Tiago V. Barreira, Mary Hastert, Chelsea L. Kracht, John M. Schuna, Zhenghua Cai, Minghui Quan, Catrine Tudor-Locke, Melicia C. Whitt-Glover, and David R. Jacobs. 2024. 2024 Adult Compendium of Physical Activities: A third update of the energy costs of human activities. *Journal of Sport and Health Science* 13, 1 (Jan. 2024), 6–12. doi:10.1016/j.jshs.2023.10.010

[38] Jack Hester, Ha Le, Stephen Intille, and Meier Erin. 2023. A feasibility study on the use of audio-based ecological momentary assessment with persons with aphasia. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23)*. 7. doi:doi/10.1145/3597638.3608419

[39] Julia Himmelsbach, Markus Garschall, Sebastian Egger, Susanne Steffek, and Manfred Tscheligi. 2015. Enabling accessibility through multimodality?: Interaction modality choices of older adults. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 195–199. doi:10.1145/2836041.2836060

[40] Joyce Ho and Stephen S. Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 909–918. doi:10.1145/1054972.1055100

[41] Alexander Hoelzemann and Kristof Van Laerhoven. 2024. A matter of annotation: An empirical study on in situ and self-recall activity annotations from wearable sensors. *Frontiers in Computer Science* 6 (July 2024), 1379788. doi:10.3389/fcomp.2024.1379788

[42] Gary Hsieh, Ian Li, Anind Dey, Jodi Forlizzi, and Scott E. Hudson. 2008. Using visualizations to increase compliance in experience sampling. In *In Proceedings of the 10th international conference on Ubiquitous computing (UbiComp '08)*. ACM, 164–167. doi:10.1145/1409635.1409657

[43] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. micro-EMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, NY, USA, 1124–1128. doi:10.1145/2971648.2971717

[44] Denis Ivanko, Alexey Kashevnik, Dmitry Ryumin, Andrey Kitenko, Alexandr Axyonov, Igor Lashkov, and Alexey Karpov. 2022. MIDriveSafely: Multimodal interaction for drive safely. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 733–735. doi:10.1145/3536221.3557037

[45] Zhang Jiahao, Gould Stephen, and Ben-Shabat Itzik. 2020. Vidat—{ANU} {CVML} Video Annotation Tool. https://github.com/anucvml/vidat

[46] Paul Kelly, Simon J. Marshall, Hannah Badland, Jacqueline Kerr, Melody Oliver, Aiden R. Doherty, and Charlie Foster. 2013. An ethical framework for automated, wearable cameras in health behavior research. *American Journal of Preventive Medicine* 44, 3 (March 2013), 314–319. doi:10.1016/j.amepre.2012.11.006

[47] Jacqueline Kerr, Simon J. Marshall, Suneeta Godbole, Jacqueline Chen, Amanda Legge, Aiden R. Doherty, Paul Kelly, Melody Oliver, Hannah M. Badland, and Charlie Foster. 2013. Using the SenseCam to improve classifications of sedentary behavior in free-living settings. *American Journal of Preventive Medicine* 44, 3 (March 2013), 290–296. doi:10.1016/j.amepre.2012.11.004

[48] Alireza Khanshan, Pieter Van Gorp, Raoul Nuijten, and Panos Markopoulos. 2021. Assessing the influence of physical activity upon the experience sampling response rate on wrist-worn devices. *International Journal of Environmental Research and Public Health* 18, 20 (Oct. 2021), 10593. doi:10.3390/ijerph182010593

[49] N. W. Kim, S. C. Joyner, A. Riegelhuth, and Y. Kim. 2021. Accessible visualization: Design space, opportunities, and challenges. *Computer Graphics Forum* 40, 3 (June 2021), 173–188. doi:10.1111/cgf.14298

[50] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating older adults to collect in-situ activity labels on a smartwatch with speech. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. doi:10.1145/3491102.3517457

[51] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@Hand: Fostering visual exploration of personal data on smartphones leveraging speech and touch interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, 1–17. doi:10.1145/3411764.3445421

[52] Zachary D. King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. micro-Stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant mothers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1–22. doi:10.1145/3351249

[53] Mandy Korpusik, Salima Taylor, Sai Krupa Das, Cheryl Gilhooly, Susan Roberts, and James Glass. 2019. A food logging system for iOS with natural spoken language meal descriptions. *Current Developments in Nutrition* 3 (June 2019). doi:10.1093/cdn/nzz041.P21-009-19

[54] Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. 2014. A multimodal in-car dialogue system that tracks the driver's attention. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 26–33. doi:10.1145/2663204.2663244

[55] Hyeokhyen Kwon, Gregory D. Abowd, and Thomas Plötz. 2019. Handling annotation uncertainty in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*. ACM, New York, NY, USA, 109–117. doi:10.1145/3341163.3347744

[56] Paula Lago, Shingo Takeda, Sayeda Shamma Alia, Kohei Adachi, Brahim Bennai, Francois Charpillet, and Sozo Inoue. 2020. A dataset for complex activity recognition with micro and macro activities in a cooking scenario. http://arxiv.org/abs/2006.10681

[57] Rithika Lakshminarayanan, Arushi Uppal, Ha Le, James Spilsbury, and Stephen Intille. 2024. Detecting sleep disruptions in adolescents using context-sensitive ecological momentary assessment: A feasibility study. In *Proceedings of the 18th EAI International Conference on Pervasive Computing Technologies for Healthcare*. ACM, New York, NY, USA, 1–12.

[58] Pontus Larsson. 2016. Speech feedback reduces driver distraction caused by in-vehicle visual interfaces. In *Proceedings of the Audio Mostly 2016*. ACM, Norrköping Sweden, 7–11. doi:10.1145/2986416.2986435

[59] Ha Le, Rithika Lakshminarayanan, Jixin Li, Varun Mishra, and Stephen Intille. 2024. Collecting self-reported physical activity and posture data using audio-based ecological momentary assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–35. doi:10.1145/3678584

[60] Matthew L. Lee and Anind K. Dey. 2014. Real-time feedback for improving medication taking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2259–2268. doi:10.1145/2556288.2557210

[61] Saija Lemmelä, Akos Vetek, Kaj Mäkelä, and Dari Trendafilov. 2008. Designing and evaluating multimodal interaction for mobile contexts. In *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, New York, NY, USA, 265–272. doi:10.1145/1452392.1452447

[62] Jixin Li, Aditya Ponnada, Wei-Lin Wang, Genevieve Dunton, and Stephen Intille. 2024. Ask less, learn more: Adapting ecological momentary assessment survey length by modeling question-answer information gain. *Proceedings of the ACM*

on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, 4 (Nov. 2024), 1–32. doi:10.1145/3699735

[63] Jieun Lim, Youngji Koh, Auk Kim, and Uichin Lee. 2024. Exploring context-aware mental health self-tracking using multimodal smart speakers in home environments. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. doi:10.1145/3613904.3642846

[64] Quanfeng Luo, Jiaji Zhou, Fei Wang, and Liping Shen. 2009. Context aware multimodal interaction model in standard natural classroom. In *Hybrid Learning and Education*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Fu Lee Wang, Joseph Fong, Liming Zhang, and Victor S. K. Lee (Eds.). Vol. 5685. Springer Berlin Heidelberg, 13–23. doi:10.1007/978-3-642-03697-2_2 Series Title: Lecture Notes in Computer Science.

[65] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. FoodScrap: Promoting rich data capture and reflective food journaling through speech input. In *Designing Interactive Systems Conference 2021*. ACM, New York, NY, USA, 606–618. doi:10.1145/3461778.3462074

[66] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: Shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3313831.3376616

[67] Lucas M. Silva and Daniel A. Epstein. 2021. Investigating preferred food description practices in digital food journaling. In *Designing Interactive Systems Conference 2021*. ACM, 589–605. doi:10.1145/3461778.3462145

[68] Jaclyn P. Maher, Amanda L. Rebar, and Genevieve F. Dunton. 2018. Ecological momentary assessment Is a feasible and valid methodological tool to measure older adults' physical activity and sedentary behavior. *Frontiers in Psychology* 9 (Aug. 2018), 1485. doi:10.3389/fpsyg.2018.01485

[69] Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: Investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 477–486. doi:10.1145/1357054.1357131

[70] Alan Mazankiewicz, Klemens Böhm, and Mario Berges. 2020. Incremental real-time personalization in human activity recognition using domain adaptive batch normalization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 1–20. doi:10.1145/3432230

[71] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting receptivity for mHealth interventions in the natural environment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (June 2021), 1–24. doi:10.1145/3463492

[72] Varun Mishra, Byron Lowens, Sarah Lord, Kelly Caine, and David Kotz. 2017. Investigating contextual cues as indicators for EMA delivery. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 935–940. doi:10.1145/3123024.3124571

[73] Lingfei Mo, Zengtao Feng, and Jingyi Qian. 2016. Human daily activity recognition with wearable sensors based on incremental learning. In *2016 10th International Conference on Sensing Technology (ICST)*. IEEE, Nanjing, China, 1–5. doi:10.1109/ICSensT.2016.7796224

[74] Aja L. Murray, Tong Xie, Luke Power, and Lucy Condon. 2024. Recruitment and retention of adolescents for an ecological momentary assessment measurement burst mental health study: The MHIM engagement strategy. *Health Expectations* 27, 3 (June 2024), e14065. doi:10.1111/hex.14065

[75] National Institutes of Health (NIH). 2015. *Racial and Ethnic Categories and Definitions for NIH Diversity Programs and for Other Reporting Purposes*. Notice NOT-OD-15-089. https://grants.nih.gov/grants/guide/notice-files/not-od-15-089.html

[76] Tadashi Okoshi, Julian Ramos, Hiroki Nozaki, Jin Nakazawa, Anind K. Dey, and Hideyuki Tokuda. 2015. Reducing users' perceived mental effort due to interruptive notifications in multi-device mobile environments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Osaka Japan, 475–486. doi:10.1145/2750858.2807517

[77] Sharon Oviatt. 1999. Ten myths of multimodal interaction. *Commun. ACM* 42, 11 (Nov. 1999), 74–81. doi:10.1145/319382.319398

[78] Sharon Oviatt. 2000. Taming recognition errors with a multimodal interface. *Commun. ACM* 43, 9 (Sept. 2000), 45–51. doi:10.1145/348941.348979

[79] Sharon Oviatt, Rachel Coulston, and Rebecca Lunsford. 2004. When do we interact multimodally?: Cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 129–136. doi:10.1145/1027933.1027957

[80] Alina Petukhova, João P. Matos-Carvalho, and Nuno Fachada. 2024. Text clustering with LLM embeddings. http://arxiv.org/abs/2403.15112 arXiv:2403.15112 [cs].

[81] H. Pirsiavash and D. Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2847–2854. doi:10.1109/CVPR.2012.6248010

[82] Aditya Ponnada, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille. 2017. Microinteraction ecological momentary assessment response rates: Effect of microinteractions or the smartwatch? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 1–16. doi:10.1145/3130957

[83] Aditya Ponnada, Jixin Li, Shirlene Wang, Wei-Lin Wang, Bridgette Do, Genevieve F. Dunton, and Stephen S. Intille. 2022. Contextual biases in microinteraction ecological momentary assessment (micro-EMA) non-response. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (March 2022), 1–24. doi:10.1145/3517259

[84] Aditya Ponnada, Binod Thapa-Chhetry, Justin Manjourides, and Stephen Intille. 2021. Measuring criterion validity of microinteraction ecological momentary assessment (Micro-EMA): Exploratory pilot Study with physical activity measurement. *JMIR mHealth and uHealth* 9, 3 (March 2021), e23391. doi:10.2196/23391

[85] Aditya Ponnada, Shirlene Wang, Daniel Chu, Bridgette Do, Genevieve Dunton, and Stephen Intille. 2022. Intensive longitudinal data collection using microinteraction ecological momentary assessment: Pilot and preliminary results. *JMIR Formative Research* 6, 2 (Feb. 2022), e32772. doi:10.2196/32772

[86] Huimin Qian, Ravi Kuber, and Andrew Sears. 2014. Supporting the mobile notification process through tactile cues selected using a paired comparison task. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1741–1746. doi:10.1145/2559206.2581133

[87] Tiago Reis, Marco De Sá, and Luís Carriço. 2008. Multimodal interaction: Real context studies on mobile digital artefacts. In *Haptic and Audio Interaction Design*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Antti Pirhonen, and Stephen Brewster (Eds.). Vol. 5270. Springer Berlin Heidelberg, Berlin, Heidelberg, 60–69. doi:10.1007/978-3-540-87883-4_7 Series Title: Lecture Notes in Computer Science.

[88] Aki Rintala, Martien Wampers, Inez Myin-Germeys, and Wolfgang Viechtbauer. 2020. Momentary predictors of compliance in studies using the experience sampling method. *Psychiatry Research* 286 (April 2020), 112896. doi:10.1016/j.psychres.2020.112896

[89] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 1–23. doi:10.1145/3161187

[90] Susana Rubio, Eva Díaz, Jesús Martín, and José M. Puente. 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology* 53, 1 (Jan. 2004), 61–86. doi:10.1111/j.1464-0597.2004.00161.x

[91] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with earables: A systematic literature review and taxonomy of phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 1–57. doi:10.1145/3550314

[92] Ayshwarya Saktheeswaran, Arjun Srinivasan, and John Stasko. 2020. Touch? Speech? or Touch and speech? Investigating multimodal interaction for visual network exploration and analysis. *IEEE Transactions on Visualization and Computer Graphics* 26, 6 (June 2020), 2168–2179. doi:10.1109/TVCG.2020.2970512

[93] Markus Scholz, Till Riedel, Mario Hock, and Michael Beigl. 2013. Device-free and device-bound activity recognition using radio signal strength. In *Proceedings of the 4th Augmented Human International Conference*. ACM, 100–107. doi:10.1145/2459236.2459254

[94] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. Association for Computing Machinery, New York, NY, USA, 415–422. doi:10.1145/258549.258821

[95] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology* 4, 1 (April 2008), 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415

[96] Lucas M. Silva, Elizabeth A. Ankrah, Yuqi Huai, and Daniel A. Epstein. 2023. Exploring opportunities for multimodality and multiple devices in food journaling. *Proceedings of the ACM on Human-Computer Interaction* 7, MHCI (Sept. 2023), 1–27. doi:10.1145/3604256

[97] Paul J. Silvia, Thomas R. Kwapil, Kari M. Eddington, and Leslie H. Brown. 2013. Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review* 31, 4 (Aug. 2013), 471–481. doi:10.1177/0894439313479902

[98] Joshua M. Smyth and Author A. Stone. 2003. Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies* 4, 1 (2003), 35–52. doi:10.1023/A:1023657221954

[99] A. W. Sokolovsky, R. J. Mermelstein, and D. Hedeker. 2014. Factors predicting compliance to ecological momentary assessment among adolescent smokers. *Nicotine & Tobacco Research* 16, 3 (March 2014), 351–358. doi:10.1093/ntr/ntt154

[100] Georgios Sopidis, Michael Haslgrübler, Behrooze Azadi, Bernhard Anzengruber-Tánase, Abdelrahman Ahmad, Alois Ferscha, and Martin Baresch. 2022. Micro-activity recognition in industrial assembly process with IMU data and deep learning. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 103–112. doi:10.1145/3529190.3529204

[101] Arjun Srinivasan and John Stasko. 2018. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 511–521. doi:10.1109/TVCG.2017.2745219

[102] Hyewon Suh, Nina Shahriaree, Eric B. Hekler, and Julie A. Kientz. 2016. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3988–3999. doi:10.1145/2858036.2858448

[103] Jessie Sun, Mijke Rhemtulla, and Simine Vazire. 2021. Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin* 47, 11 (Nov. 2021), 1535–1549. doi:10.1177/0146167220964639

[104] Patrick Tchankue, Dieter Vogts, and Janet Wesson. 2010. Design and evaluation of a multimodal interface for in-car communication systems. In *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*. ACM, 314–321. doi:10.1145/1899503.1899538

[105] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. http://arxiv.org/abs/2302.13971

[106] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. 2018. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (Jan. 2018), 1–22. doi:10.1145/3161192

[107] Niels Van Berkel, Jorge Goncalves, Simo Hosio, Zhanna Sarsenbayeva, Eduardo Velloso, and Vassilis Kostakos. 2020. Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies* 134 (Feb. 2020), 1–12. doi:10.1016/j.ijhcs.2019.10.003

[108] Niels Van Berkel, Jorge Goncalves, Peter Koval, Simo Hosio, Tilman Dingler, and Vassilis Kostakos. 2019. Context-informed scheduling and analysis: Improving accuracy of mobile self-reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3290605.3300281

[109] Mitchell Webber and Raul Fernandez Rojas. 2021. Human activity recognition with accelerometer and gyroscope: A data fusion approach. *IEEE Sensors Journal* 21, 15 (Aug. 2021), 16979–16989. doi:10.1109/JSEN.2021.3079883

[110] Jing Wei, Benjamin Tag, Johanne R Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What could possibly go wrong when interacting with proactive smart speakers? A case study using an ESM application. In *CHI Conference on Human Factors in Computing Systems*. ACM, 1–15. doi:10.1145/3491102.3517432

[111] Cornelia Wrzus and Andreas B. Neubauer. 2023. Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment* 30, 3 (April 2023), 825–846. doi:10.1177/10731911211067538

[112] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2022. GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (Dec. 2022), 1–34. doi:10.1145/3569485

[113] Xinghui Yan, Yuxuan Li, Bingjian Huang, Sun Young Park, and Mark W Newman. 2021. User burden of microinteractions: An in-lab experiment examining user performance and perceived burden related to in-situ self-reporting. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. ACM, 1–14. doi:10.1145/3447526.3472046

[114] Xinghui Yan, Shriti Raj, Bingjian Huang, Sun Young Park, and Mark W. Newman. 2020. Toward lightweight in-situ self-reporting: An exploratory study of alternative smartwatch interface designs in context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 1–22. doi:10.1145/3432212

[115] Yu Gu, Lianghu Quan, and Fuji Ren. 2014. WiFi-assisted human activity recognition. In *2014 IEEE Asia Pacific Conference on Wireless and Mobile*. IEEE, 60–65. doi:10.1109/APWiMob.2014.6920266

[116] Katrin Ziesemer, Laura Maria König, Carol Jo Boushey, Karoline Villinger, Deborah Ronja Wahl, Simon Butscher, Jens Müller, Harald Reiterer, Harald Thomas Schupp, and Britta Renner. 2020. Occurrence of and reasons for "missing events" in mobile dietary assessments: Results from three event-based ecological momentary assessment studies. *JMIR mHealth and uHealth* 8, 10 (Oct. 2020), e15430. doi:10.2196/15430

A Appendix A: Activity level estimation

To decide which activities to display on the screen, we use a rule-based, heuristics algorithm. We assigned each activity in the list into one level: *Sedentary, Moderate and Vigorous*. The categorization and the list of activities are included in the supplementary materials. We determined participant's activity level using heart rate and accelerometer data before a prompt popped up (see Table 9). We filtered the list of activities that fell into the detected physical activity level, and we displayed the three activities that were ***most reported*** by the participants. If there were ties, we displayed the activities based on alphabetical order.

Table 9: How we determined physical activity levels using heart rate (HR) and wrist accelerometer data (AUC unit).

Activity Level	*Heart rate (HR) and AUC values*
Vigorous	HR > 150 OR AUC >= 6000
Moderate	NOT Vigorous AND (HR >= 105 OR AUC >= 2000)
Sedentary	HR < 105 AND AUC < 2000

B Appendix B: Prompts used for LLM experiment

To run the LLM experiment, we used the following prompt, which is tailored for the CAPTURE-24 label list. To run the experiment with the P&R label list, we simply replaced the list of labels (see the *italicized text*).

```
This is what the person reported doing:
[participants' self-report].
Map their self-report to one or more of the
following labels:
sleep,
sitting,
standing,
lying,
kneeling,
bend over,
household-chores,
manual-work,
walking,
mixed-activity,
vehicle,
sports,
bicycling,
others.
If the self-report fits multiple labels,
seperate them with a backslash (such as
sitting/household-chores, standing/sports,
sitting/vehicle, ...). If the self-report doesn't
fit any of the labels, return 'others'. Give a
single answer (no explanation, limited prose). Do
not invent new labels.
```

The prompt for the P&R experiment is identical, except for the activity list is replace with these activities: sitting,standing, lying, kneeling, bend-over, walking, combing hair, make up, brushing teeth, dental floss, washing hands/face, drying hands/face', enter/leave room, adjusting thermostat, laundry, washing dishes, moving dishes, making tea, making coffee, drinking water/bottle, drinking water/tap, making hot food, making cold food/snack, eating food/snack, mopping in kitchen, vacuuming, taking pills, watching tv, using computer, using cell, making bed, cleaning house, reading book, using mouth wash, writing, putting on shoes/socks, drinking coffee/tea, grabbing water from tap, other

C Appendix C: Results of individual items in the SUS survey

Table 10 shows the results of individual SUS items.

D Appendix D: Themes, Codes and definitions (qualitative analysis for participant's perceived burden)

Table 11 shows the definitions of the themes mentioned in the qualitative analysis (Section 6.2.2), as well as exampled generative codes.

Table 10: Individual SUS questions with mean and standard deviation of the score. The scale for grading is from 0-4. The table shows the raw scores recorded in the SUS response ("Strongly Agree" is graded as 4, and "Strongly Disagree" is graded as 0). For the even numbered questions (marked with * in the table), the raw scores are reverse graded (by subtracting the score from 4) — so lower score on these question implies high usability. The final SUS score is calculated by summing up all the grade for individual questions, and multiplying by 2.5.

Question	*Mean*	*SD*
I think that I would like to use this system frequently.	2.73	1.10
I found the system unnecessarily complex*	1.2	0.94
I thought the system was easy to use.	3.27	0.80
I think that I would need the support of a technical person to be able to use this system.*	0.47	0.74
I found the various functions in this system were well integrated.	3.00	1.00
I thought there was too much inconsistency in this system.*	1.13	0.92
I would imagine that most people would learn to use this system very quickly.	3.33	0.82
I found the system very cumbersome to use.*	1.07	0.88
I felt very confident using the system.	3.27	0.70
I needed to learn a lot of things before I could get going with this system.*	0.6	0.74

Table 11: Definitions of the codes about participant's perceived source of burden emerged from the qualitative analysis.

Type of burden			*Example codes (count)*	*Definition*
Interruption	Social	`Disturbing others`	high perceived social disruption (10)	Situations where the prompting cue causes disruptions for people around the participants
	Cognitive	`Cognitive interruption`	prompt during activity transition is cognitively burdensome (3), annoyed with prompts during focus time (5)	Situations where participants claimed the prompts interrupted them from a cognitively engaging task
Interaction	Physical	`Hand availability`	speech preferred when hands are busy/moving around (14)	Situations where participants' hand movement affected their modality choice or non-response
		`Movement/activity`	speech was helpful to avoid looking at watch screen during activity (2)	Situations where participants' activity level affected their modality choice or non-response
		`Reactivity`	instinct to look at screen during watch prompt (7), tendency to bring watch hand near mouth for speech (2)	Situations where participants' reaction to the prompt affected their modality choice or non-response (e.g., a change in body movement, bringing their hand close to their mouth, looking/glancing at the watch)
	Cognitive	`Mental bias/uncertainty`	hard to come up with response for speech (6), audio quality concerns (20)	Situations where participants expressed doubt/uncertainty about what to report or the quality of the recorded label
		`Repetition fatigue`	repetition is burdensome (11), increase time between prompts (7)	Situations where participants complained about the repetition of the same label
	Social	`Social discomfort`	study participation details not shared in professional settings (2), non speech inputs preferred around others (5)	Situations where answering to the prompting caused participants to be uncomfortable around other people