Towards Wearable Opportunistic Crowdsensing for Open-Vocabulary Activity Data Collection Through User-Scheduled Trigger-Action Routines

ZEYU WANG[†] and YINGKE DING[†], Tsinghua University, China
MINGZE GAO, Tsinghua University, China
ZHUOLUN REN, Renmin University of China, China
ALEX MARIAKAKIS, University of Toronto, Canada
YUANCHUN SHI, Tsinghua University, China
YUNTAO WANG[*], Tsinghua University, China

Collecting richly labeled wearable activity data in everyday settings remains difficult because retrospective annotation is costly and often imprecise. Prior data collection apps rely on a labor-intensive self-reporting strategy and primarily treat participants as crowd labelers. We present Pebbl, a feasibility-stage system that incentivizes in-situ labeling through opportunistic crowdsensing. Pebbl lets users author trigger-action recipes on a smartphone and receive just-in-time reminders for beneficial actions when a trigger is detected. In the prototype, triggers are a limited set with four common audio cues, while actions are described in open-vocabulary natural language. Each confirmed execution yields a short sensor window with explicit start/end boundaries and a user-authored action label. We evaluate Pebbl through an expert workshop with wearable Human Activity Recognition (HAR) researchers ($N = 6$), a within-subject in-lab study ($N = 21$), and a pilot deployment ($N = 8$). Experts viewed the approach as lower burden and more ecologically valid than common labeling workflows. In the lab, Pebbl produced reliable execution logs under controlled conditions (recall = 97.30%, precision = 97.15%) and was preferred over comparison workflows on perceived burden and confidence. The pilot deployment shows that the interaction and sensing pipeline can function in free-living use, while surfacing practical constraints such as false triggers and context dependence. Overall, Pebbl represents a step toward a low-burden, distributable collection approach of user-contributed wearable activity data.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Ubiquitous and mobile computing**.

[†]Co-author equal contribution.
[*]Corresponding author.

Authors' Contact Information: Zeyu Wang, Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China, wang-zy23@mails.tsinghua.edu.cn; Yingke Ding, Department of Computer Science and Technology, Tsinghua University, Beijing, China, dyk21@mails.tsinghua.edu.cn; Mingze Gao, Department of Computer Science and Technology, Tsinghua University, Beijing, China, gaomingze2022@gmail.com; Zhuolun Ren, School of Journalism and Communication, Renmin University of China, Beijing, China, rzl2021@ruc.edu.cn; Alex Mariakakis, Computer Science, University of Toronto, Toronto, Ontario, Canada, mariakakis@cs.toronto.edu; Yuanchun Shi, Department of Computer Science and Technology, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China, and Qinghai University, Xining, Qinghai, China, shiyc@tsinghua.edu.cn; Yuntao Wang (corresponding author), Department of Computer Science and Technology, Tsinghua University, Beijing, China, and School of Computer Technology and Application, Qinghai University, Xining, Qinghai, China, yuntaowang@tsinghua.edu.cn.



This work is licensed under a Creative Commons Attribution 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/9-ART166
https://doi.org/10.1145/3831976

Additional Key Words and Phrases: Activity Recognition, Fogg Behavior Model, Smartwatch Application, Dataset Collection, Tiny-Habit

ACM Reference Format:
Zeyu Wang, Yingke Ding, Mingze Gao, Zhuolun Ren, Alex Mariakakis, Yuanchun Shi, and Yuntao Wang. 2026. Towards Wearable Opportunistic Crowdsensing for Open-Vocabulary Activity Data Collection Through User-Scheduled Trigger-Action Routines. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 166 (September 2026), 42 pages. https://doi.org/10.1145/3831976

1 INTRODUCTION

Large-scale, richly labeled datasets have driven major advances in vision and language modeling [14, 31, 37], but wearable human activity recognition (HAR) still lacks similarly broad and well-annotated data resources [9, 49]. This gap is especially pronounced for fine-grained activities performed in everyday life. Controlled lab protocols yield clean labels but limited realism [2], while free-living data often requires burdensome post-hoc annotation process that is time-consuming and vulnerable to recall and segmentation errors [20, 21, 46]. In-situ self-report can reduce recall bias by asking people to label their own activities as they happen or shortly thereafter [24, 42]. However, current systems still largely frame participants as annotators performing research labor. This limits sustainability: even when labels are valuable to researchers, the interaction often remains effortful and provides little direct value to the people generating the data.

We investigate a different design direction: can wearable activity labeling be embedded in an experience that users already find useful? To explore this question, we present Pebbl, a smartphone-smartwatch system that combines user-authored trigger-action routines with just-in-time watch prompts. Users specify a trigger, a target beneficial action; when the trigger is detected, the watch reminds the user to perform the action and records a bounded sensor window if they confirm. Available triggers are a limited set that are detectable cues, while actions are described in open-vocabulary natural language. An example interation flow is shown in Figure 1.

Pebbl is inspired by trigger-action routines used in behavior-change systems [18, 19, 41]. We leverage this framing not to claim that Pebbl is a validated habit-formation intervention, but to give users a practical reason to engage with in-situ labeling: the system can function as a reminder and lightweight tracker for personally meaningful micro-actions. This design naturally aligns with **wearable opportunistic crowdsensing** [8, 27] that capitalizes on users' everyday routines with minimal explicit action.

After implementing Pebbl for iPhones and Apple Watches, we conducted three studies with the varied stakeholders associated with our target domain. As a data-collection approach, Pebbl involves two central stakeholder groups: researchers who may use the resulting wearable activity data, and users who produce the labeled sensor windows through usage. First, we held a workshop with wearable HAR researchers to confirm that Pebbl had preferable qualities relative to existing data collection approaches **(RQ1)**. We then conducted a within-subjects lab study comparing Pebbl to both baseline annotation tools and a standard tracking app. Participants served as both Pebbl app users and dataset annotators, allowing us to assess both data labeling reliability **(RQ2)** and their tracking outcomes **(RQ3)**. Finally, we conducted an in-the-wild study to understand Pebbl's affordance in real-world settings.

Our findings suggest that Pebbl is preferred by HAR researchers and annotators as a data collection platform, as well as by users as an app for encouraging and tracking beneficial behaviors. Six experts endorsed Pebbl's potential for scalability, ecological validity, and low burden, while 21 participants confirmed that they preferred Pebbl for its low mental/physical workload and just-in-time prompts. Diagnostic logs with precise timestamps and explicit options to skip activities were well-received features, and Pebbl yielded higher label reliability relative to all baselines. Through a pilot deployment with 8 participants, we show that Pebbl's interaction and sensing pipeline can function well in short-term free-living use, while surfacing practical constraints around false triggers and context dependence.

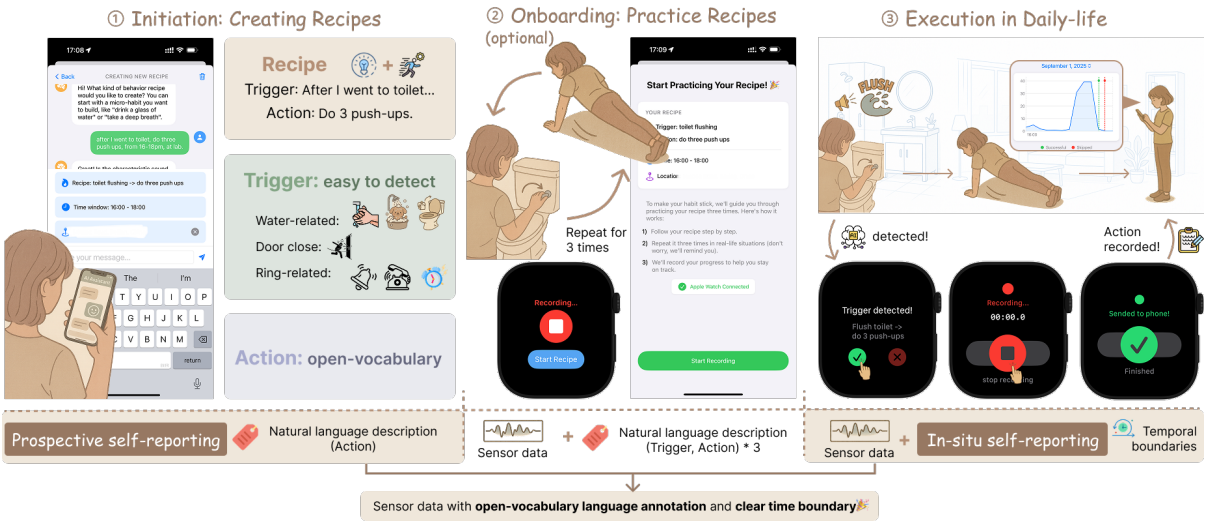


Fig. 1. **Overview of Pebbl's interaction and data flow.** Pebbl is an opportunistic crowdsensing system that couples data collection with reminders and trackers for beneficial actions. This figure demonstrates Pebbl's key components. **(1) Initiation:** An Large Language Model (LLM)-powered agent helps users author recipes by specifying an audio-based **Trigger** with time/location conditions, and an **Action** described in open-vocabulary natural language (e.g., "do three push-ups"). **(2) Onboarding:** The app guides users to rehearse the recipe on the watch and record the practice data with weak annotations. **(3) Execution in daily life:** When the trigger is detected, Pebbl provides a just-in-time watch prompt and the user confirms whether they will complete the corresponding action. Each execution yields a timestamped, precisely bounded sensor window (e.g., audio/IMU) paired with an open-vocabulary action label, enabling in-situ self-reporting with clear temporal boundaries.

Our contributions are:

- **An opportunistic crowdsensing framework for wearable HAR data collection via user-scheduled trigger-action routines.** We present Pebbl, a feasibility-stage framework that explores a path toward gathering timestamped, labeled activity data at scale while promoting beneficial actions.
- **A full-stack mobile–wearable system and staged empirical evaluation.** We implemented Pebbl on an iPhone and an Apple Watch and evaluated it through expert elicitation, a controlled lab study, and a pilot deployment.
- **Empirical findings, design implications, and limitations for deployment.** We show that Pebbl can reduce perceived burden and produce reliable execution logs in controlled settings, and derive implications and scope boundaries for broader deployment.

2 RELATED WORK

2.1 Labeling Wearable Activity Data in the Wild

Wearable HAR datasets are commonly created through three families of labeling workflows: scripted collection in controlled settings [10, 33, 47], free-living recording with post-hoc annotation [9, 20, 48], and free-living recording with participant self-report [23, 30, 42]. Controlled protocols offer clean labels but limited ecological validity. Post-hoc annotation supports naturalistic behavior but is costly and often imprecise at fine temporal scales. Self-report reduces recall delay and gives participants semantic access to their own actions, but still places ongoing annotation burden on them. Prior systems in this space have asked users to periodically enter high-level

labels, receive automatic label suggestions, or verbally describe activities shortly after completion [24, 30, 42]. Our work builds most directly on this third category. The key difference is that Pebbl does not present labeling only as research work; instead, it embeds confirmation and logging into a user-authored trigger-action tracking flow.

2.2 Embedding Annotation in Personally Valuable Experiences

A longstanding Human-Computer Interaction (HCI) strategy is to collect data while giving users something valuable in return, such as engagement, utility, accessibility, or reduced effort. One thread embeds annotation into rewarding experiences: gamified systems ask users to solve puzzles, play Extra-Sensory Perception (ESP) guessing games, or compete in multi-participant scenarios while producing labels [13, 23, 28, 40, 43, 44], while utility-oriented systems such as MovieLens, reCAPTCHA, and VizWiz exchange user contributions for recommendations, login utility, or real-time accessibility support [5, 22, 45]. A second thread is opportunistic crowdsensing, which leverages sensor-rich mobile devices to collect data with minimal user input and has been applied to smart cities, environmental monitoring, public health surveillance, and mobile health [12, 15, 26, 29, 34].

These ideas are less developed for wearable time-series sensing, where data is harder for end users to inspect directly. Pebbl adopts the same high-level principle: if a labeling interaction also helps users remember, track, or reflect on personally meaningful actions, they may be more willing to sustain it while producing useful labels.

2.3 Trigger-Action Routines and Contextual Reminders

Behavior-tracking and behavior-change systems often support adherence through self-tracking, reminders, or contextual prompts. Popular habit-tracking apps monitor task completion [1], but prior reviews found that many focus on self-tracking and reminders while offering limited support for context-driven cues that help routines develop [38]. Event-based reminders and Just-in-Time Adaptive Interventions (JITAIs) show how prompts can be coupled to sensed or inferred contexts, including interventions for sedentary behavior [7, 25], eating [11, 32], smoking [36, 39], procrastination [3], and deep breathing [35]. Related in-situ sensor-triggered prompt systems include on-body camera agents [17], which differ from Pebbl in relying on visual context rather than user-authored trigger-action routines.

We draw specifically on the Fogg Behavior Model (FBM) and tiny-habit recipes [18, 19]. The FBM frames behavior as occurring when motivation, ability, and a trigger coincide; tiny-habit recipes operationalize this idea by anchoring a small action to a specific existing routine, such as "After I flush the toilet, I will do 3 push-ups." Pebbl adapts this trigger-action structure for wearable data labeling: the trigger helps surface an in-situ prompt, while the user-authored action provides a prospective label for the recorded sensor window. We do not evaluate Pebbl as a habit-formation intervention or propose a new behavior-change theory; rather, we use this framing to create a low-burden path to activity labeling that is also personally meaningful to users.

3 PEBBL DESIGN AND IMPLEMENTATION

3.1 Design Overview

Pebbl is designed around four stages: recipe creation, optional practice, in-situ execution, and reflection, as summarized in Figure 2. During recipe creation, the user specifies (1) a trigger, (2) a target action, and (3) optional time and location constraints. In the current prototype, triggers must correspond to one of four non-speech audio classes: door closing, toilet flushing, running water, or ringtone/alarm. Actions are authored in free-form natural language and are intended to be short, feasible, and immediately executable after the trigger.

When a recipe becomes active, Pebbl monitors for the trigger within the specified context. If the trigger is detected, the watch presents a lightweight prompt asking the user whether they will perform the action. If they confirm, recording starts; if they skip, the skip is logged. The user taps again when the action is complete,

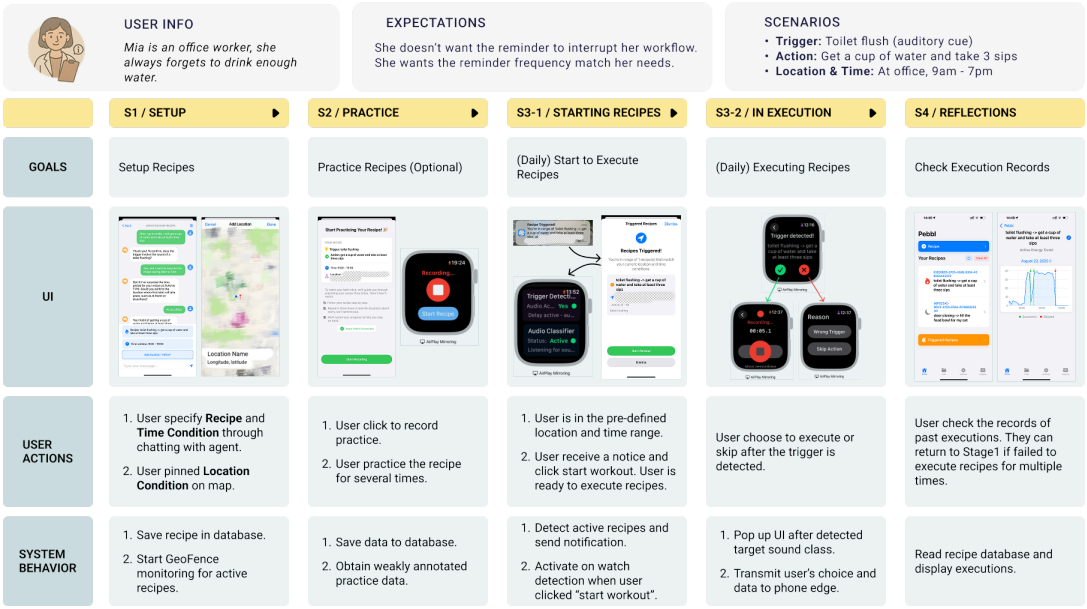


Fig. 2. An experience map illustrating a typical user's journey with Pebbl.

producing an explicit time-bounded execution record. The phone summarizes completed and skipped executions to support reflection and recipe revision.

This design intentionally prioritizes short, repeated, context-tethered actions. It is not intended for long-duration activities or actions that users routinely delay far beyond the trigger moment.

3.2 Prototype Implementation

We implemented Pebbl as native iPhone and Apple Watch applications. Recipe authoring occurs on the phone through a chatbot interface that converts user input into a structured recipe representation. Trigger monitoring and execution confirmation occur on the watch. In the current prototype, trigger detection combines on-device sound classification, simple motion gating, and optional time/location constraints to reduce false activations and battery use. Once the user confirms execution, the watch records audio and inertial measurement unit (IMU) data until the user indicates completion, after which the files and associated metadata are transferred to the phone for storage.

The implementation prioritizes on-device inference for trigger detection and temporally sparse recording for uploaded execution data. Full implementation details, prompt design, UI flows, and detector logic are provided in Appendix A.

4 ETHICS, CONSENT, AND DATA HANDLING

All study procedures were reviewed and approved by our university Institutional Review Board under No. `03-2025-1005`. All participants provided informed consent before taking part and were informed about what data would be collected in each study, including audio, motion, location, interview recordings, and, in Study 2, neck-worn video. Participants could withdraw at any time without penalty, and could request deletion of their data.

Table 1. Overview of the three studies. Per-study detail is in the corresponding *Participants* subsections.

Study	N	Participants	Setting	Purpose	Key outcomes
S1	6	4M/2F, ages 23–27; wearable-HAR researchers with IMU data collection experience, from 4 research institutions	Expert workshop (~150 min)	Elicit evaluation criteria and expert assessment of Pebbl relative to common data-collection workflows	Rubric dimensions, comparative ratings, qualitative tradeoffs
S2	21	10M/11F, ages 21–31; 15 students, 4 early-career professionals, 2 others	Controlled in-lab study (~120 min)	Measure reliability of Pebbl's execution records and compare perceived burden with alternative annotation workflows	Log recall/precision, wrong-trigger counts, temporal alignment, subjective ratings
S3	8 (3 overlap with S2)	5M/3F, ages 20–27; own iPhone and Apple Watch users	Pilot deployment (~4 h of Pebbl use)	Assess whether the interaction and sensing pipeline can function outside the lab and how users perceive it in everyday use	Wrong-trigger rate, manual-fallback use, qualitative feedback, subjective ratings

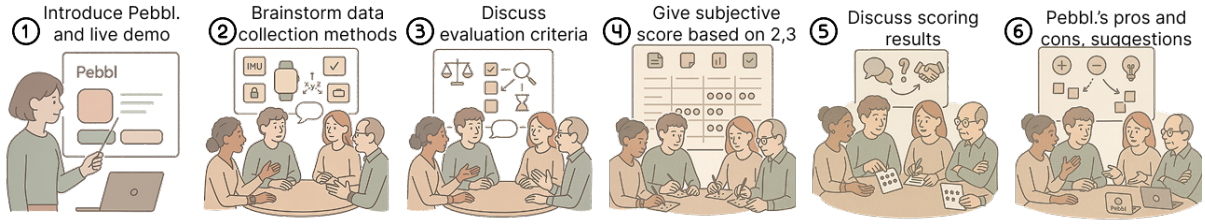


Fig. 3. The procedure of the expert workshop (Study 1).

Because Pebbl involves privacy-sensitive sensing, we separated on-device sensing used for trigger detection from data retained for analysis. Trigger detection relied on on-watch audio and IMU processing, while non-triggered sensor streams were not uploaded to lab servers. For study analysis, we retained only confirmed execution records, including short audio clips, triaxial IMU, timestamps, recipe labels, and relevant time/location context. In Study 2, neck-worn video was used only as a reference for post-hoc annotation and was deleted within six months of study completion.

5 EXPLORATORY STUDY WITH EXPERT WORKSHOP

We evaluate Pebbl through three studies, summarized in Table 1: an expert workshop (this section), a controlled in-lab study (Section 6), and a pilot deployment in free-living settings (Section 7). To understand Pebbl's potential as a scalable data collection platform, we first consulted wearable-HAR researchers, who represent downstream data consumers familiar with the practical tradeoffs of collecting and annotating activity datasets. The workshop provided a researcher-facing assessment of Pebbl's perceived benefits, limitations, and tradeoffs relative to existing collection workflows, complementing the participant-facing evidence in Studies 2 and 3. This study addresses **RQ1**.

5.1 Experiment Design

5.1.1 Participants. We recruited 6 wearable HAR researchers and individuals with vast experience in gathering wearable IMU datasets. The experts comprised 4 males and 2 females aged between 23 and 27 years old, and they

came from 4 research institutions. Three of them were from academia, while the other three were affiliated with industrial research labs. They reported 2–6 years (mean ≈ 4) of experience in wearable sensing and/or activity dataset collection. The workshop lasted approximately 150 minutes, and the experts received compensation of 50 USD per hour.

5.1.2 Procedure. The workshop, illustrated in Figure 3, had three goals: (1) identify the dimensions experts use to judge wearable activity data-collection workflows, (2) position Pebbl relative to common alternatives, and (3) elicit likely strengths, limits, and deployment concerns. Participants first received an interactive demo of Pebbl, including recipe generation and execution, and briefly tried the system on their own. We then presented three workflow families drawn from prior work: controlled protocolized collection [10, 33, 47], free-living recording with post-hoc annotation [9, 20, 48], and free-living recording with participant self-report [24, 42] – and asked experts to refine the evaluation criteria they considered most relevant when comparing these approaches with Pebbl.

After agreeing on a rubric (i.e., the set of evaluation dimensions), experts independently rated each of the four approaches. We treat these ratings as expert judgments rather than objective performance measurements, because only Pebbl was directly experienced during the workshop while the comparison approaches were evaluated based on prior experience and the literature above. A moderated discussion then surfaced disagreements and key tradeoffs, with one team member leading the conversation while another flagged dimensions with the widest rating spread for follow-up.

5.1.3 Analysis Plan. For each rubric dimension, we compare the scores experts gave to each data collection approach. The qualitative data were analyzed by following Braun et al. [6]'s reflexive thematic analysis procedure; more details are provided in Appendix B.7.1.

5.2 Results and Findings

We report the workshop results in four parts: the data-collection approaches and rubric dimensions the experts converged on, the comparative scores they assigned to each approach, a summary of the overall assessment, and the qualitative feedback that complements those scores.

5.2.1 Data Collection Methods & Rubric Metrics. The experts converged on the following list of wearable HAR data-collection approaches for our discussion:

- **Predefined-class, Protocolized Collection in a Controlled Environment (M1):** Target activity categories are fixed and predefined; recruited participants wear sensing devices and follow a prescribed protocol to record these classes.
- **Open-world, Free-living with Post-hoc Annotation (M2):** Participants free-roam wearing both a recording device and a "ground-truth" device; afterwards, annotators label the data segments offline.
- **Open-world, Free-living with Participant Self-report (M3):** Participants free-roam wearing the recording device and label data either during the activity or shortly thereafter (within the day).
- **Ours (Pebbl):** Our approach is similar to M3, but differs by pairing prospective self-report with in-situ confirmations, and by providing personal value through situated reminders for beneficial actions.

The dimensions by which the experts evaluated these approaches are defined in Table 2. Note that some of the items like *Accuracy* and *Bias* included sub-items for more depth. The experts were instructed to score each method on a scale from 1–7. For all dimensions other than those related to *Cost* and *Bias*, higher scores were more preferable.

5.2.2 Summary of Expert Assessment. Table 3 summarizes the experts' ratings. We interpret these scores as an elicitation of expert expectations about likely tradeoffs, not as definitive comparative evidence: only Pebbl

Table 2. The rubric dimensions used by experts to evaluate wearable HAR data-collection approaches. Note that the order in which these metrics are presented is weakly associated with their importance, as weighed by the experts.

Primary Dimension	Sub-Dimension	Relevant Indicators and Measures
Vocabulary Openness & Coverage		Distinct action classes; proportion of fine-grained vs. coarse labels; coverage of verb–object–context combinations; distinct language annotation for similar activities;
Scale & Diversity		Number of participants; recording length; environment/demographic diversity; per-activity instance counts; total annotated clips;
Accuracy	**Label Granularity**	Semantic specificity of labels (task-, object-, limb-level); presence of hierarchical labels; target task granularity;
	Synchronization	Estimated clock drift across devices; presence of alignment anchors (e.g., beeps, taps)
	Accuracy of Label Timing	Onset/offset jitter; missed/extra events rate; proportion of truncated/merged segments
	Accuracy of Label Semantics	Inter-annotator agreement on descriptions; contradiction rate across modalities
	Ecological Validity	Degree to which data reflects naturalistic behavior; fraction of free-living data; prompt-induced behavior rate; overlap with everyday routines;
Cost & Feasibility	**Data Capture Difficulty**	User and researcher's burden; onboarding/training time; hardware requirements; failure/retry rate;
	Data Annotation Difficulty	Minutes of annotation per hour of data; need for expert annotators; instruction complexity; throughput;
Bias	**Activity Bias**	Class imbalance (long-tail); coverage of rare/complex actions; diurnal/weekday bias
	Participant Bias	Age, gender, handedness, ability, culture; recruitment channels; exclusions
	Context Bias	Scene diversity; seasonality/weather coverage;
Privacy Protection		Strength of privacy-by-design: data minimization, on-device processing, encryption, user transparency;
Multi-modality Support		Number of modalities (e.g., accel, gyro, photoplethysmography, audio); sampling rates; missing-data rate; multi-device fusion support

Table 3. The average scores of 1-7 Likert Scale given by the experts when evaluating different wearable HAR data-collection approaches. The arrows next to the column names indicate whether higher (↑) or lower (↓) scores are better. The best scores in each column are highlighted in bold.

Method	Coverage↑		Accuracy↑					Cost & Feasibility↓		Bias (diversity)↓			Privacy↑	Sensors↑
	Vocab.	Scale	Granu.	Synchro.	Timing	Semantics	Validity	Capt. Diff	Ann. Diff	Act. Bias	User Bias	Ctx Bias	Privacy	Sensors
M1	1.50	2.33	4.67	**6.33**	**6.50**	**6.50**	2.00	4.83	3.33	5.67	5.33	6.17	4.67	**5.83**
M2	4.67	4.67	**5.50**	5.17	5.67	5.83	5.50	4.33	6.00	**3.00**	3.83	3.50	1.67	5.00
M3	5.50	5.83	3.50	4.67	3.50	4.50	5.67	3.83	4.33	3.33	3.00	2.33	6.17	3.17
Ours	**6.67**	**6.50**	4.83	5.00	5.00	5.50	**6.50**	**2.33**	**2.33**	3.17	**2.67**	**2.17**	**6.33**	3.17

was directly experienced during the workshop, whereas the three comparison workflows were rated from prior experience and the literature. With that caveat, experts generally saw Pebbl as promising on *Vocabulary Openness*, *Scale*, *Ecological Validity*, *Capture* and *Annotation Difficulty*, and *Privacy Protection*, while viewing controlled protocolized collection (M1) as stronger on *Synchronization*, *Timing*, and *Semantic Accuracy*, and post-hoc annotation (M2) as stronger on *Label Granularity*. Pebbl and M3 trailed on *Multi-modality Support*, reflecting a deliberate tradeoff favoring deployability and privacy over maximal sensing. Experts also emphasized that Pebbl's advantages are contingent on recipe fit: the approach is most plausible for short, repeatable, context-linked activities, and less plausible for actions that are delayed, socially awkward, or poorly captured by the current trigger set.

5.2.3 Qualitative Feedback.

Pros. Expert 5 appreciated that Pebbl could minimize unintended data collection by only recording short windows of data at moments when users previously agreed to be recorded. Expert 1 commented that Pebbl would impose *"almost no user burden"*; meanwhile, Expert 6 went as far as saying that people may forget that they are collecting data for research rather than just scaffolding a healthy habit, which may results in collecting more

natural behavior. Expert 3 stated, *"The tiny-habit premise tends to attract self-motivated users. They might be more responsible; more detailed,"*. They later hypothesized that the annotations generated by such individuals would be more reliable.

Cons. Expert 3 worried that by focusing on beneficial activities, less positive or hedonic activities (e.g., gaming, smoking) would be difficult to capture using Pebbl. Expert 1 had similar concerns by noting, *"It would probably only cover adolescents to middle-aged; older adults usually wouldn't use it."* These two experts also expressed concerns about the precise timing of the recordings. Expert 3 worried that users may start the recording after seeing the trigger notification on their watch, only to complete the activity when they find a more convenient opportunity, while Expert 1 worried that some users may forget to stop recording upon activity completion. Expert 1 also expressed concerns that extraneous or poorly timed triggers may annoy users.

Design Trade-offs. The experts highlighted a tension between Pebbl's detailed onboarding and its intention of promoting an open vocabulary of activities. Expert 3 cautioned that prescriptive guidance risks narrowing what users create: *"The design threshold is high. Some users may struggle to imagine how to design recipes . . . If you give especially good examples, everyone may just use those."* On the other hand, Expert 1 noted that if Pebbl's guidance was too loose or generic, users would have no clear direction on how they should use the system. Expert 4 further warned that *"Without feasibility cues, end users may propose undetectable actions, such as opening some desktop app, that wearable sensors can't capture"*.

Suggestions for Future Refinements. The experts suggested several system refinements for future work. Although we did not implement these recommendations in this work, we provide them here to inspire later iterations:

(1) Broaden and automate the trigger space to include modalities beyond audio and automatically discover salient cues for future recipes (Experts 2, 4, and 6).
(2) Add automated verification of action completion to prevent a *"false sense of accomplishment when users confirm but do not actually execute an action"* (Expert 6).
(3) Enrich the agent's dialogue to help users design recipes for breaking undesirable habits (Expert 1).
(4) Allow the system to work with only a smartphone to enable more engagement (Expert 1). Expert 5 remarked that implementing this idea might introduce new challenges, saying *"We don't know where the phone is and positions change across activities"*.

6 MAIN STUDY: IN-LAB EVALUATION

To support whether the experts' perspectives on Pebbl would materialize in practice, we invited several individuals to evaluate it as both users and data annotators. This study addresses two research questions: "How efficient and reliable is Pebbl as an approach to scalable data collection?" (**RQ2**), and "Can Pebbl encourage people to take beneficial actions during everyday living?" (**RQ3**).

6.1 Experiment Design

6.1.1 Participants. We included 21 participants, comprising 10 males and 11 females aged between 21 and 31 years old. Participants were recruited from a community-based pool of adult smartphone users, yielding 15 university students, 4 early-career professionals, and 2 individuals from other occupational backgrounds; 15 reported prior use of a smartwatch, and 5 reported prior use of a habit-tracking or self-logging app. The experiment lasted approximately 120 minutes, and they received compensation of 30 USD per hour.

6.1.2 Comparison Conditions. Study 2 included two classes of comparison. First, to evaluate Pebbl as a data-labeling workflow, we compared it against three alternatives that represent common annotation strategies

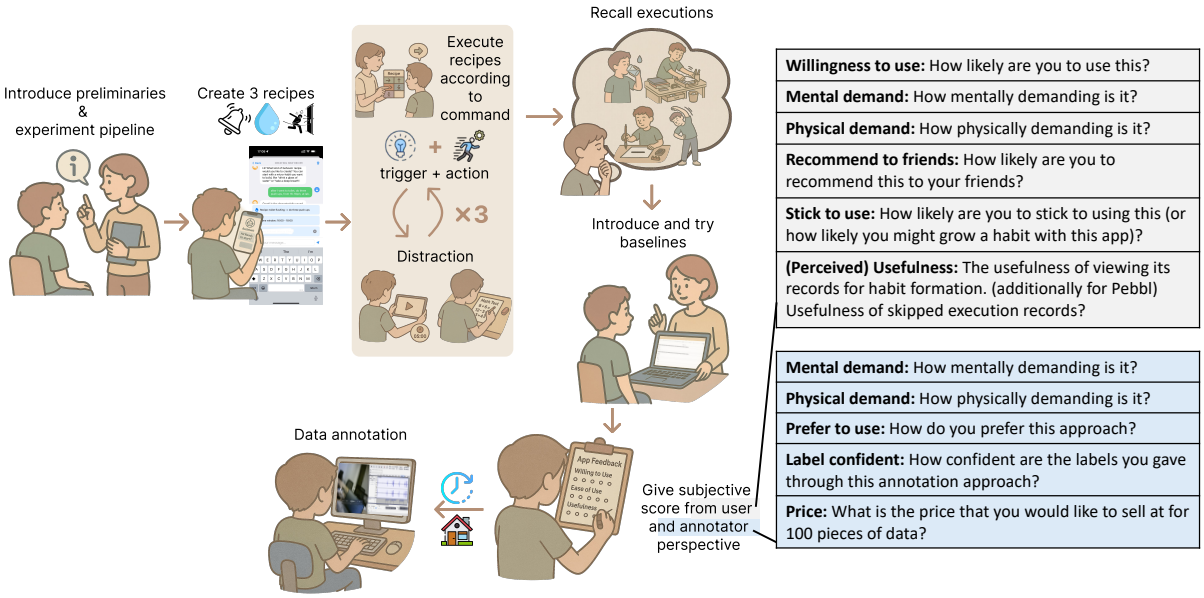


Fig. 4. The procedure of the in-lab user study (Study 2).

Table 4. Study 2 comparison conditions. B1–B3 are annotation-workflow comparisons; B4 is used only for the user-facing reminder/tracking comparison. Detailed implementations and citations are given in the text below.

ID	Represents	Participant task	Compared on
B1	Post-hoc video annotation	Review neck-camera video and assign labels and temporal boundaries using an annotation tool	Annotation burden, confidence, and temporal quality
B2	Near-time self-report	Record activities shortly after performing them	Annotation burden and confidence; comparison to retrospective memory
B3	In-situ self-report with phone UI	Label activities using a mobile interaction similar to prior self-report systems	Annotation burden and confidence
B4	Commercial habit tracker	Manually log completed actions using a publicly available tracking app	User-facing reminder/tracking preference only

discussed in Study 1: post-hoc video annotation (**B1**), near-time self-report (**B2**), and an in-situ phone-based self-report interface inspired by prior work (**B3**). Second, to evaluate the user-facing reminder/tracking aspect of Pebbl, we compared it against a commercial habit-tracking app (**B4**). Table 4 summarizes these conditions and the outcomes for which each was used. We do not claim that these baselines exhaust the design space; rather, they provide representative contrasts for annotation burden, label confidence, temporal quality, and user-facing tracking preference.

For the annotation-workflow comparison, the three baselines were implemented as follows:

(1) **B1** [9, 20, 46]: Annotators examined time-aligned ground-truth video recorded by a neck-worn camera and gave text labels to selected intervals using a professional labeling tool[1].
(2) **B2** [24]: Annotators labeled events shortly after recording them.
(3) **B3** [42]: We implemented a demo similar to the ExtraSensory app[2] to reflect a fine-grained open-vocabulary interaction paradigm.

Examples of these interfaces are provided in Appendix A.3.

For the user-facing tracking comparison, we used a publicly available habit-tracking app[3] (**B4**). This app offers a counter that users can increment either immediately after completing a target action or later in the day after recalling its execution.

6.1.3 Procedure. The study was held in a meeting room positioned less than 10 meters from a washroom, a kitchen, an elevator, and the building's waste handling area, meaning that there was moderate background noise throughout the entire experiment. Figure 4 summarizes the procedure, which can be divided into four phases:

We designed Phase 2 (the main Pebbl condition) as a controlled measurement rather than a naturalistic use trace. Triggering each recipe on a pre-generated schedule, with the experimenter randomly instructing execution or skipping, yields known ground truth against which Pebbl's records can be validated. We therefore interpret the reliability numbers reported below as evidence about the interaction's correctness under controlled conditions; the free-living pilot (Study 3) provides the naturalistic complement.

(1) **Introduction:** After introducing participants to the concept of tiny-habit recipes, we asked them to brainstorm three recipes with distinct audio triggers (examples can be found in Appendix B.2). We then provided them with materials that explained Pebbl and the experiment procedure; these materials can be found in Appendix B.1.
(2) **Main Experiment:** Participants were asked to trigger each of their recipes 30 times. Using a monitor, a research team member instructed participants to either skip or complete the corresponding action at random. After the 7^{th} and 16^{th} trials, participants were distracted by a 5-minute entertainment video or arithmetic problems to disrupt any habituation between trials. Throughout this session, participants were required to wear a camera around their neck, serving as our ground truth for annotation. We collected two sets of egocentric videos with different camera configurations: one with electronic image stabilization (EIS) enabled and one without. We did this because most popular smart-glass camera models ship without built-in EIS [4, 16], so both settings are realistic deployment scenarios for the baselines.
(3) **Post-experiment:** Participants were asked to recall how many times they actually completed the actions associated with each of the triggers, providing an indicator of whether participants could rely on their recent memory for annotation. We then allowed participants to experiment with the data-annotation and behavior-tracking baselines for a few trials, after which they provided feedback from the perspectives of both users and annotators. Because Pebbl was evaluated as a full end-to-end workflow with extensive trials, whereas the baselines were introduced through short hands-on trials to support comparison, we interpret these ratings as participants' informed perceptions of the relative burden, confidence, and preference associated with each workflow.
(4) **At-home Annotation:** After participants left the lab, we sent them the video captured by the camera they wore during the study so that they could provide baseline annotations for executed actions. To ease some of the annotation burden, we provided them with approximate execution times noted by the research team member so that they could quickly locate relevant segments. Videos were 60-80 minutes long, and participants spent 30-60 minutes annotating them. Every submission was reviewed by the research team,

[1] https://www.audacityteam.org
[2] https://github.com/cal-ucsd/ExtraSensoryAndroid
[3] https://apps.apple.com/us/app/habit-tracker/id1438388363

anchored against the pre-generated execution list from Phase 2 to catch missing entries, duplicates, and obvious mislabels. This was a lightweight consistency check to make sure that participants fully understand the annotation procedure.

6.1.4 Analysis Plan. Study 2 focuses on *controlled reliability*, not long-term adoption. We therefore analyze four objective outcomes — (1) whether Pebbl's execution records match the known trial schedule, (2) how well participants retrospectively remember executions, (3) how often the trigger detector produces false prompts, and (4) how Pebbl's time boundaries compare with post-hoc video annotation — together with subjective ratings used to assess perceived burden and confidence for representative comparison workflows. Long-term adherence and sustained naturalistic use are out of scope for this study; the pilot deployment (Study 3) addresses those questions at a preliminary scale.

The pre-generated execution order list served as ground-truth labels for when actions were completed. Meanwhile, the camera served as ground-truth information for the actions' timestamps. Due to camera issues mid-study, we were only able to analyze camera feed data from 18 participants: 9 recorded with EIS and 9 without. To align Pebbl's clock with the camera's, we scanned the camera's video with a 0.1-second step length to find the latency that maximized the intersection-of-union (IoU) between participants' annotations and Pebbl's records. This information allowed us to conduct objective evaluations to assess three aspects of Pebbl: (1) the accuracy of its resulting records, (2) the rate of incorrect triggers, and (3) the precision of its resulting time boundaries.

- **Label Accuracy:** We report the recall, precision, and F1 score of Pebbl's exported execution records relative to the pre-generated execution order list from the main experiment. True positives are defined as instances when a Pebbl record is as defined in the order list, false positives are defined as instances when a Pebbl record is not as defined in the order list, and false negatives are defined as instances when an execution in the order list is not recorded by Pebbl.
- **Participant Memory:** We define the recall of participants' memory as one minus the mean absolute error (MAE) between recalled and ground-truth execution counts, normalized by the total number of executions: $1 - \frac{\text{MAE}(\#executions_{gt} - \#executions_{recall})}{\#executions_{total}}$. $\#executions_{gt}$ refers to the ground truth provided by the generated execution list. We further analyzed participants' recall for successfully executed recipes and skipped recipes, as well as execution during each session. We included this metric to assess how accurately participants can retrospectively recall their own executions, given that both our in-the-wild data collection baseline and the baseline tracking app rely on after-the-fact labeling.
- **Wrong Trigger Rate:** We counted the `Wrong Trigger` clicks for each category and each participant, treating them as user-reported false positives when Pebbl wrongly recognized an irrelevant ambient event as the recipe trigger.
- **Temporal Alignment:** We calculate the IoU, temporal recall/precision, and length ratio of annotated actions.

The subjective evaluation metrics listed in Figure 4 were scored along a 7-point Likert scale, except for price. Because Pebbl provides extra incentive by offering reminder + tracking features, we explicitly require users to take that into account when evaluating the price.

Our post-study questionnaire combined adapted workload items (mental demand and physical demand from NASA-TLX) with custom items targeting willingness to use, recommendation likelihood, stickiness, perceived usefulness, annotation preference, and label confidence. These custom items were authored by the research team to probe the adoption and annotation constructs most relevant to Pebbl's two comparison classes (reminder/tracking vs. annotation workflow). We used subsets of the items in each comparison — the user-facing comparison used the reminder/tracking subset; the annotation-workflow comparison used the annotation subset. The full list of items and the exact wording administered to participants are reproduced in Appendix B.4.

Table 5. Pebbl's labeling accuracy.

Recall	**Precision**	**F1**	**Recall** (success)	**Recall** (skip)	**P. recall**	**P. recall** (success)	**P. recall** (skip)	**P. recall** (sessions)
97.30%	97.15%	97.22%	98.06%	96.64%	70.32%	76.73%	61.71%	51.75%

6.2 Results and Findings

We report objective measures of Pebbl's reliability (label accuracy, wrong-trigger occurrence, participant memory, and temporal alignment), followed by participants' subjective ratings under the two comparison classes: the reminder/tracking-side comparison and the annotation-workflow comparison.

6.2.1 System Performance Evaluation.

Controlled reliability of Pebbl's execution logs. Table 5 shows that Pebbl's execution records closely matched the pre-generated trial schedule under controlled conditions (recall = 97.30%, precision = 97.15%, F1 = 97.22%). This result reflects the strength of Pebbl's interaction design: the user is prompted at the trigger moment and must explicitly confirm execution or skipping. We therefore interpret this finding narrowly as evidence that Pebbl can produce reliable execution logs in a structured setting. It should not be read as proof that all resulting data windows are equally valid in unconstrained real-world use, where users may delay actions, miss prompts, or stop recordings late.

Retrospective recall was substantially noisier. Participants' retrospective counts were markedly less reliable than Pebbl's in-situ records. Average memory accuracy was 70.32% overall, with better recall for successful executions than skipped ones (76.73% vs. 61.71%). Session-level recall dropped to 51.75%, indicating difficulty placing events accurately in time even over a short study period. Wilcoxon signed-rank tests at the experiment level were significant at $p < 10^{-5}$ across all comparison pairs. These results support the core motivation for Pebbl: when logging is deferred, participants reconstruct rather than report, and that reconstruction is error-prone.

False triggers remained a practical limitation. Across 630 true trigger opportunities, participants reported 68 wrong-trigger events (10.79% relative to true triggers; see Figure 5). After normalizing by the number of experiments exposed to each trigger class, errors concentrated in *water* ($\approx$ 2.0 wrong/exp) and *door* ($\approx$ 1.3 wrong/exp), with *flush toilet* intermediate ($\approx$ 0.8) and *bell* most robust ($\approx$ 0.19). This pattern is consistent with acoustic distinctiveness: bells and toilet flushes have more stereotyped temporal–spectral signatures, whereas water and doors are acoustically diverse and frequently co-occur with domestic confounders. Per-participant variation (Appendix B.6) was concentrated in a small tail driven by the same two noisy classes and by proximity to a washroom, kitchen, and waste area less than 10 m from the lab. Deployment quality therefore depends strongly on acoustic distinctiveness and on the ambient context in which each recipe is active.

Time boundaries were tighter than post-hoc annotation, but still instruction-sensitive. Compared with post-hoc video annotation, Pebbl generally produced tighter temporal windows (Figure 6): most participants show high recall but low precision when labeling video, with length ratios > 1, indicating that annotators tend to over-segment events, whereas Pebbl's Start/End taps anchor the recording around the execution. The EIS condition yields somewhat tighter averages than the non-EIS condition, but per-participant variability dominates device effects, as is common in fine-grained temporal annotation. This comparison depends on participants following the intended interaction pattern: in Study 2 they were explicitly instructed to tap `Start` just before acting and `End` immediately after completion. Real-world use is likely to be messier, so we treat this result as evidence about the usefulness of the interaction design rather than a guarantee of boundary fidelity in deployment. Section 8.3 discusses mitigations.

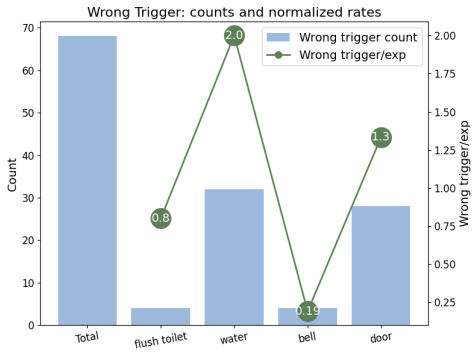


Fig. 5. Wrong-trigger counts and per-experiment rates by trigger class. The per-participant breakdown is in Appendix B.6.

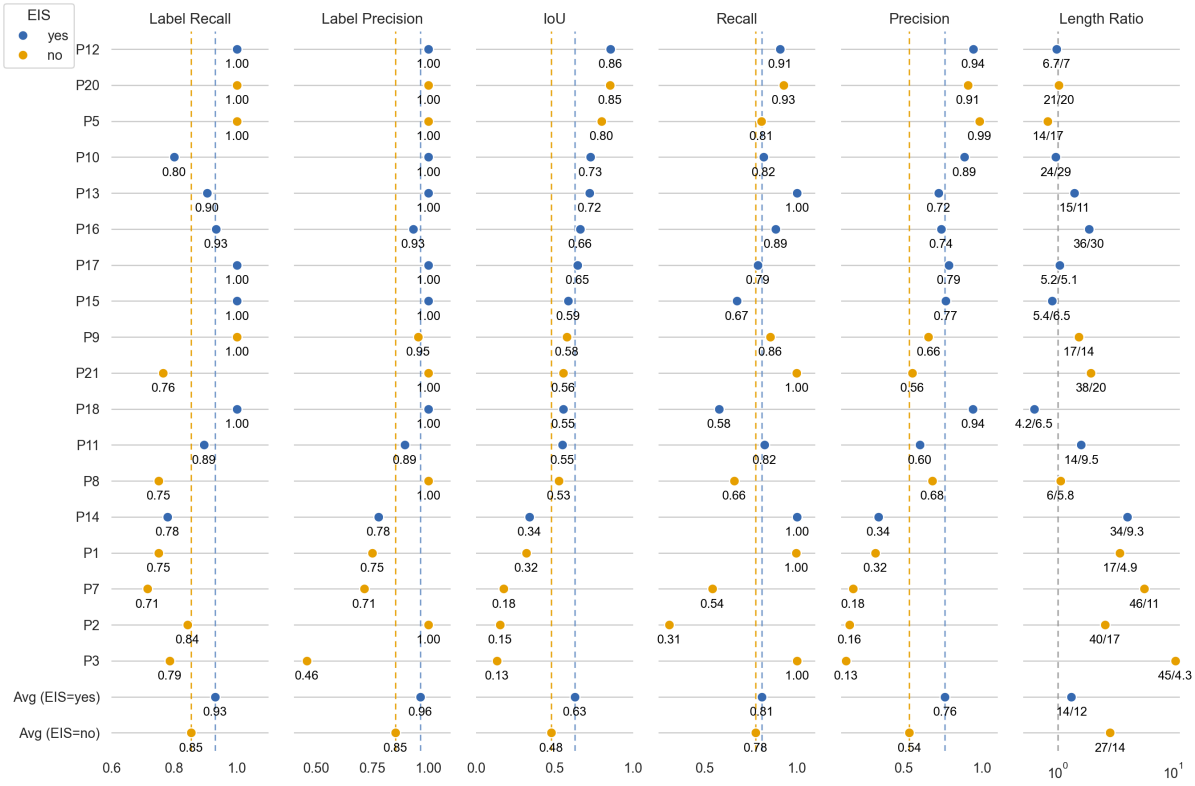


Fig. 6. Annotation Quality Analysis. We show results of both description-level and segment-level evaluation metrics per participant. Orders were rearranged by annotation IoU, which is the primary evaluation metric. "Length Ratio" represents the average interval length of the participant's annotations over Pebbl's recordings. We draw the vertical line of Length Ratio = 1 in the last column instead of the average value. Electronic Image Stabilization (EIS) represents two camera types used in our experiment.

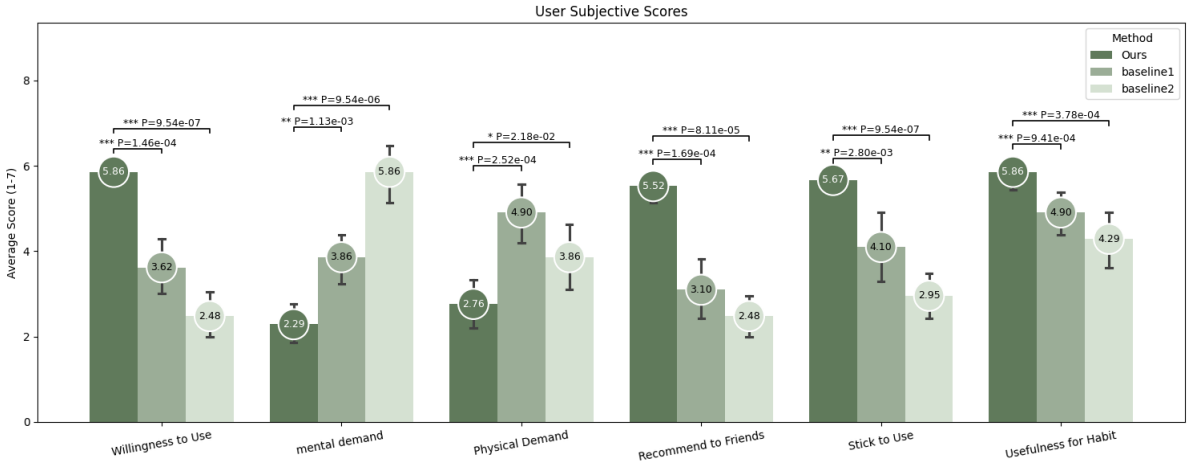


Fig. 7. Subjective ratings under the user-facing reminder/tracking comparison (Pebbl vs. B4 and the two self-report–style conditions). Per-dimension means, signed-rank p-values, and effect sizes are reported in the figure; ** denotes $p < 0.01$ and *** denotes $p < 0.001$.

6.2.2 User-Facing Perceptions of Pebbl as a Reminder/Tracking Tool. Figure 7 summarizes participants' ratings for the user-facing tracking comparison. Under the representative exposure used in Study 2 — brief trials with the comparison conditions relative to the full Pebbl session — participants reported higher willingness to use Pebbl, lower mental and physical demand, higher likelihood of recommending it, higher likelihood of continued use, and greater perceived usefulness of Pebbl's records for reflection. Exact means and p-values (Wilcoxon signed-rank) are reported in the figure. Interview comments suggest that the main driver was not visual design but timing: Pebbl asks for confirmation at the activity opportunity, whereas the baseline relies on later manual logging. Participants also valued the visibility of *skipped* executions (5.48 ± 0.85 on a 1–7 scale), which several described as useful for reflecting on why they did not act and for revising the recipe accordingly rather than just "trying harder".

We interpret these ratings cautiously. Study 2 does not establish that Pebbl improves long-term adherence or habit formation; it shows that, during short-term controlled exposure, participants perceived the reminder-plus-logging interaction as more supportive than a purely manual tracker. Study 3 provides a preliminary free-living view of the same perceptions.

6.2.3 Perceived Burden and Confidence as an Annotation Workflow. Figure 8 summarizes participants' ratings of the annotation-oriented comparison conditions (B1–B3). Participants consistently rated Pebbl as lower in mental and physical demand than B1–B3 and expressed higher confidence in the labels produced through Pebbl; exact means and p-values (Wilcoxon signed-rank) are reported in the figure. This pattern is consistent with the system design: Pebbl collapses collection and labeling into a single in-situ confirmation at the trigger moment, so annotators do less retrospective reconstruction and fewer label-choice/policy-doubt decisions than under B1–B3. The confidence gap is particularly notable against B1 (post-hoc video annotation): in Study 1, experts judged video-based annotation's *label quality* as strong, yet participants here reported lower *production-side* confidence. The two measurements are not in conflict — they capture different constructs (theoretical quality under aggregate review vs. experienced confidence while producing a label).

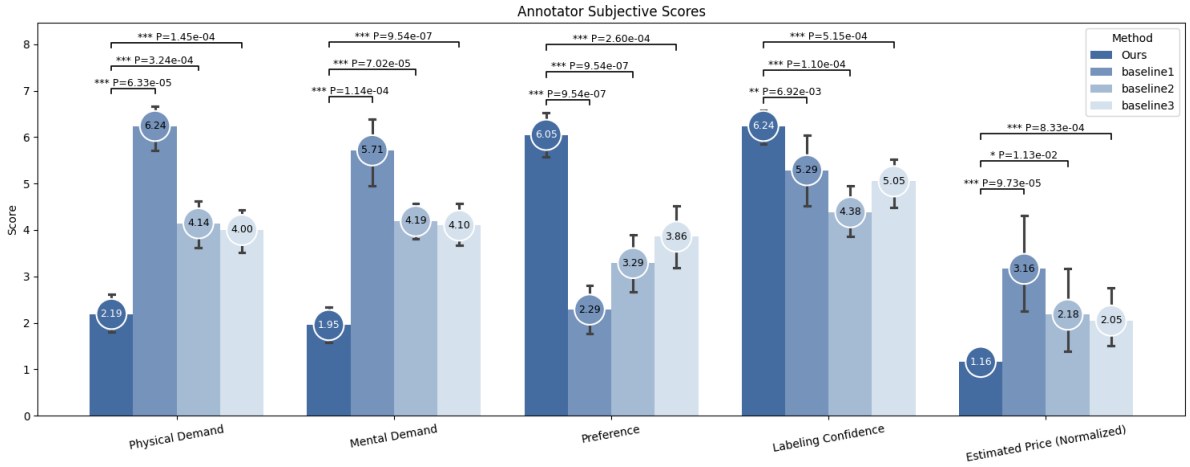


Fig. 8. Subjective ratings under the annotation-workflow comparison (Pebbl vs. B1–B3). Per-dimension means, signed-rank p-values, and effect sizes are reported in the figure; ** denotes $p < 0.01$ and *** denotes $p < 0.001$. The *price* dimension originally included in this figure has been relocated to Appendix B.5.

We do not interpret these ratings as evidence that Pebbl universally dominates all annotation workflows. Rather, for short, triggerable activities under the representative exposure used in Study 2, participants experienced Pebbl as a less effortful way to produce bounded, time-anchored labels than retrospective or phone-centric alternatives. Whether that advantage transfers to longer-duration activities, less distinctive triggers, or sustained use remains open.

7 PILOT DEPLOYMENT STUDY IN OPEN-ENVIRONMENT

Since the scenarios in the previous study were largely contrived, we conducted a pilot deployment to further support our answers to **RQ2** and **RQ3**, which respectively speak to Pebbl's efficacy as a data collection tool and a way of promoting beneficial actions. We frame this deployment explicitly as a pilot: it assesses whether Pebbl's interaction and sensing pipeline can function outside the lab, whether its specific failure modes (wrong triggers, context mismatch) reappear in free living, and how a small group perceives Pebbl as a reminder/tracking tool during short-term use. It is not intended to establish sustained adoption, long-term retention, or behavior change.

7.1 Experiment Design

We describe the pilot's participants, the additional implementation work needed to support free-living deployment, and the procedure that participants followed.

7.1.1 Participants. We recruited 8 participants (3 overlap with Study 2), comprising 5 males and 3 females, aged from 20 to 27 (24 ± 2.73). Participants were required to have an iPhone and an Apple Watch themselves. Participants were *college students*, and 6 reported prior use of a habit-tracking or self-logging app. The experiment lasted approximately 10 hours, with 120 minutes of actual execution time, and they received compensation of 70 USD in total.

7.1.2 Additional Implementation. We add manual recording approaches compared to Section 3.2, as shown in Figure 9. Specifically, users can manually start execution recording by clicking `Manual Recipe`. Meanwhile,

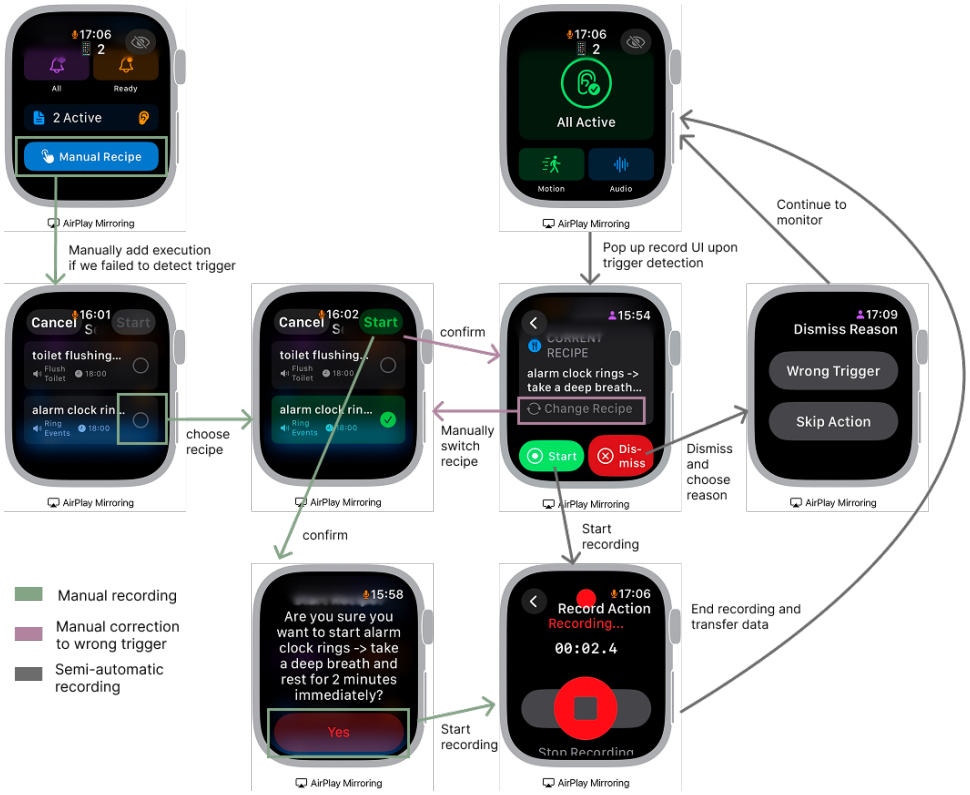


Fig. 9. In-the-wild on-watch interaction flow. Compared to Section 3.2, we add two manual operations.

users can click on `Change Recipe` to choose the correct recipe to execute if Pebbl detects a trigger incorrectly, or the user missed the haptic feedback of the last wrong trigger.

7.1.3 Procedure. In-the-wild study comprised four sessions:

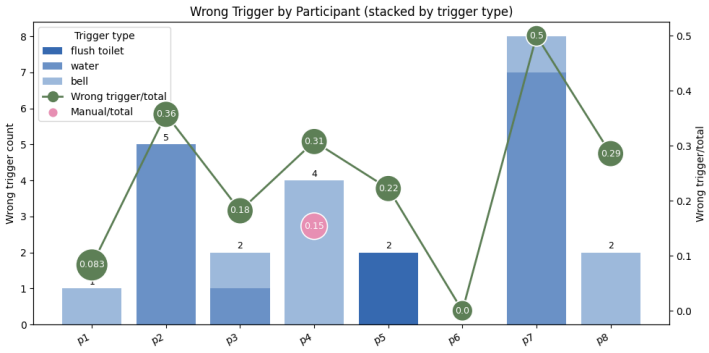
(1) **Introductory Session:** Participants check and update system version beforehand. After they come to the lab, we invite them to our testing group and install Pebbl to their devices through TestFlight[4]. Similar to Study 2, we first introduce participants to the concept of tiny habits and then ask them to design one example recipe. Then we introduce the usage of Pebbl and the study pipeline. Afterwards, we confirm the location and time range that participants want to try Pebbl within, then they brainstorm two recipes to execute for the study.

(2) **Pebbl Session:** Participants were required to try out Pebbl for approximately 4 hours, with 2 active recipes in a free-living context. Each recipe should be triggered approximately 4 times, which is a much sparser distribution compared to Study 2. This lower density was intentional: the pilot is designed to reflect intermittent natural use rather than the repeated scripted execution used in Study 2, and therefore trades statistical volume for a more deployment-like trigger schedule. Within the location and time range, participants stick to their original schedule and perform recipe triggers intermittently. After the tryout, they send the execution records and recorded data to us.

[4] https://developer.apple.com/testflight/

	Location	Context
P1	Study room	Studying
P2	In dorm	Study with roommates
P3	At lab	Coding + meeting
P4	In dorm	Watching course video (Speaker)
P5	In dorm	Resting
P6	At home	Midnight resting
P7	At lab	Working
P8	In dorm	Resting

(a) Environment.



(b) Wrong trigger and manual recipe rate.

Fig. 10. Experiment information and performance per participant.

(3) **Baseline Session:** Participants were required to try out a baseline app[5] for a similar time length, same recipes, location, and free-living context, as well as triggering recipes approximately 4 times.
(4) **Post-experiment Session:** We conduct an online scoring and interview session afterwards. The evaluation metrics are identical to user perspective metrics in Study 2 (see Figure 4). During scoring, we explicitly require participants to focus on functionality and weaken the impact of the UI's artistic design, which is not our emphasis.

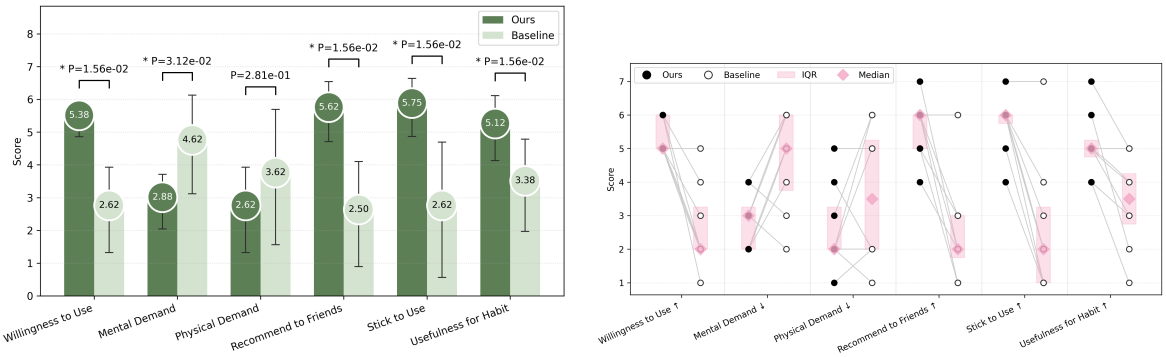
7.2 Results and Analysis

The results below should be read as *pilot* evidence. They come from a short (~4 hours of Pebbl use), small ($N = 8$), and technically narrow sample, and are not designed to support claims about long-term retention, sustained engagement, or behavior change at scale. What the pilot can show is (i) whether Pebbl's interaction and sensing pipeline functions in free living, (ii) whether the failure modes we observed in the lab (wrong triggers, context mismatch) reappear under intermittent natural use, and (iii) how a small group's Study 2-style subjective perceptions survive a short baseline comparison outside the lab. Limitations of the sample and duration are revisited in the Discussion.

7.2.1 Objective Analysis of In-the-Wild Reliability for Pebbl. We count the wrong trigger rate and the manual recipe rate for each participant. Different from Study 2, we calculate both rates by dividing the total record count, as participants no longer execute triggers identically in this study. As shown in Figure 10, Pebbl produced a moderate and context-dependent wrong trigger rate. P2 suffers from a frequent wrong triggering of the water recipe due to the fan's noise, while P4 reported that the video course's sound effect played through the speaker can sometimes arouse the ringtone recipe. P7 experienced a frequent water recipe triggering whenever he was near the washroom. However, it is not very bothersome, as this kind of fine-grained location-aware error also serves as a strong reminder, according to P7 in the interview. Conversely, only P4 has manually recorded recipes for 2 times. This demonstrates that Pebbl is relatively sensitive, which is noticed by most participants.

Taken together, Study 3 shows that wrong triggers are highly context-sensitive rather than uniformly distributed across use cases. Although wrong triggers occurred less frequently per hour than in the noisier lab environment, true executions were also much sparser in free-living use; therefore, each false prompt could be more noticeable

[5] https://apps.apple.com/us/app/habit-tracker/id1438388363



(a) Statistical tests.

(b) Descriptive statistic. Connection lines are paired scores from the same participant.

Fig. 11. In-the-wild subjective scores. (a) Per-dimension means: participants favored Pebbl across all six dimensions, with significance indicated by Wilcoxon signed-rank tests ($p < 0.05$ marked with *). (b) Per-participant scores shown as discrete markers (filled = Pebbl, hollow = baseline), with the pink band indicating the inter-quartile range (IQR) and the pink diamond indicating the median.

and affect perceived burden. We therefore treat trigger robustness and context fit as central constraints for real-world deployment.

7.2.2 In-the-Wild User Preference for Pebbl. Figure 11 shows participants' subjective scores for Pebbl and the baseline, visualized as per-participant markers with median and IQR overlays. Participants generally gave higher scores for Pebbl with respect to "willingness to use", "stick to use", "recommend to friends", and "(perceived) usefulness". These trends are consistent with Study 2. For mental demand, Pebbl tended to be rated as less demanding than the baseline, although P6 stated, *"When I was just starting the training, it felt a bit, um, complicated"*, indicating a steep learning curve during onboarding. Physical demand scores were more mixed. P2 and P4 regarded Pebbl as equally burdensome to the baseline because of the relatively higher wrong trigger rate, while P6 and P8 regarded Pebbl as more burdensome because Pebbl can effectively nudge them to execute the beneficial actions: *"Honestly, I think that burden is good. Without pressure, I won't start"* (P8). Taken together, these descriptive results corroborate Study 2 and suggest that Pebbl delivers higher perceived value and adoption potential, at the cost of some additional initial effort.

7.2.3 User-Centered Insights on Pebbl's Usability and Future Directions. In this section, we summarize the qualitative feedback we received during our interviews. Participants' full responses can be found in Appendix B.7.2.

Overall positive reception. Participants valued Pebbl's "smart" automation (P8) and tight watch–phone integration (P4/5), which made capturing moments feel effortless and timely. The haptic feedback created a rewarding sense of interaction (P3). Several participants (P6/7/8) noted that Pebbl's "gentle pressure" to act was actually productive, kick-starting beneficial actions that have the potential to form long-term habits.

Need for recipe generation guidance. For some participants (P2/3/7), Pebbl had a steep learning curve because they needed more guidance to pick recipes that would be suitable for their daily living. They were unsure of which actions to target and which triggers to pick to create the strongest connection. Users asked for stronger support, such as a conversational agent that can iteratively convert vague intentions into concrete recipes.

Suitability of audio-based triggers. Participants felt that audio-based triggers were a strong fit for short and repeatable behaviors such as drinking water, brief exercise sets, and breaking up sedentary time (P4/7). Longer actions that required more time were not viewed as suitable for Pebbl due to the need for more specific contexts that could not be conveyed by time, location, and audio alone (P1/2). Therefore, some participants wanted richer and more personalized triggers beyond the current audio set. Beyond supporting triggers according to fixed timers, other contextual variables included sedentary time (P2), screen time (P5/7), social media usage (P5/7), and other health-related routines like medication usage (P4). P7 also suggested that smarter feedback summaries that helped explain when triggers were activated would help them make potential adjustments.

Taken together, Study 3 shows that Pebbl's pipeline functions in free-living settings and that its perceived-value signals from Study 2 are directionally reproduced under a brief baseline comparison, while also re-surfacing the onboarding friction and trigger-context limitations that experts flagged in Study 1. We interpret these findings as a pilot assessment of deployment feasibility and boundary conditions, not as evidence of sustained real-world effectiveness; we return to sample size, duration, and the longitudinal question in the Discussion.

8 DISCUSSION

Across the three studies, we examine Pebbl from both sides of the proposed data-collection approach: researchers who may use the resulting wearable activity data, and users who produce data through everyday interaction. For **RQ1**, Study 1 provides a researcher-facing assessment of Pebbl's perceived value and tradeoffs relative to common wearable activity data-collection workflows. Experts saw Pebbl as promising for vocabulary openness, scalability, ecological validity, low capture/annotation burden, and privacy-preserving data minimization. At the same time, they highlighted important limitations around trigger coverage, temporal fidelity, recipe fit, activity bias, and participant diversity.

RQ2 and **RQ3** examine the user/data-provider side from two different angles: RQ2 focuses on whether Pebbl can produce useful labeled data, while RQ3 focuses on whether the same interaction provides user-facing reminder/tracking value. We therefore studied both questions in Studies 2 and 3, because each Pebbl interaction simultaneously produces a labeled sensor window and supports a tracking/reminder experience. Study 2 provides controlled evidence that Pebbl can produce reliable execution logs, tighter temporal boundaries than post-hoc annotation, lower perceived annotation burden, and higher label confidence. Study 3 complements this evidence in free-living use, where participants' feedback was broadly consistent with Study 2 but false triggers and context-dependent recipe fit became more visible. Thus, our findings support Pebbl as a feasibility-stage approach for low-burden, user-contributed wearable activity labeling, but not yet as evidence of long-term habit formation or sustained behavior change.

We now discuss the implications, scope, and deployment challenges in this section.

8.1 Trigger-Action Tracking as a Wearable Labeling Paradigm

The main contribution of Pebbl is a trigger-action tracking paradigm for low-burden wearable data labeling. Rather than asking users to retrospectively reconstruct what they did, Pebbl lets users define an intended action in advance, waits for a relevant trigger, and then asks for a lightweight confirmation at the moment the action is expected to occur. Across the expert workshop, controlled lab study, and pilot deployment, our findings suggest that this interaction can reduce perceived annotation burden while still producing timestamped execution records with user-authored, open-vocabulary labels.

The key design insight is that a trigger-action recipe can serve as a form of prospective self-report. In conventional self-report workflows, the user must observe, remember, and later describe an activity after it happens. In Pebbl, the user first specifies a meaningful action – for example, doing a short exercise, drinking water, or checking an object – as part of a personally useful tracking routine. This action description becomes

a prospective label and when the trigger occurs, the watch prompt turns this prospective label into an in-situ confirmation.

This reframing helps explain why Pebbl was perceived as lower burden than retrospective alternatives. The system does not merely make annotation faster; it changes when and why annotation happens. First, it reduces the memory burden of labeling by surfacing the recording interface at a moment of opportunity, rather than asking users to search backward through time. Second, it reduces semantic ambiguity by asking users to define the target action before execution, instead of requiring later interpretation of opaque IMU or audio traces. Third, it gives the labeling interaction immediate personal value: the same confirmation that contributes data also helps users track, reflect on, or carry out actions they already care about. The broader implication is that wearable data collection can be embedded into lightweight, personally valuable routines, rather than framed only as collection and annotation labor.

8.2 Application Scope and Boundary Conditions

Study 3 suggests that the current prototype is better suited to short, repeatable actions that can be performed soon after a recognizable cue. This boundary comes partly from the current implementation: although users can describe actions in open-vocabulary language, the deployed prototype is configured around four audio trigger classes. Thus, the scope of suitable habits depends on the available trigger space. Expanding the audio vocabulary or adding system-, motion-, physiological-, and context-based triggers could broaden the range of habits Pebbl can support, but such expansion is a deployment direction rather than a claim evaluated in this paper.

This application scope also shapes the data Pebbl can collect. Because Pebbl embeds labeling into habit-tracking routines, it is more likely to capture actions that users perceived beneficial. This produces a biased sample of all everyday activities. Undesired behaviors, privacy-sensitive actions, and behaviors from users who are less likely to adopt smartwatch-based habit tools may be underrepresented. We therefore frame Pebbl as a mechanism for low-burden in-situ labeling within a bounded class of beneficial routines, rather than as a general-purpose solution for all types of activity. This bounded scope is also population-dependent: recipes that are appropriate for our young, technically comfortable participants may not transfer directly to older adults, children, or users with mobility/accessibility needs. For such groups, Pebbl may need different task domains (e.g., medication routines, light mobility, hygiene, or caregiver-approved chores), stronger recipe-authoring assistance, and targeted validation of both usability and data quality.

8.3 Future Work: Toward At-Scale Real-World Deployment

Our studies suggest that trigger-action tracking is a promising interaction paradigm, but moving from a feasibility-stage prototype to public deployment requires additional design and engineering work. We discuss three directions: broadening the trigger space, improving data quality safeguards, and supporting everyday onboarding and retention.

Broadening, personalizing, and composing richer trigger contexts. The current prototype uses audio cues as a tractable first instantiation, but many behaviorally meaningful situations have no clear acoustic signature. Future versions should support multimodal triggers, including physiological signals such as Heart-Rate Variability (HRV), device-usage signals such as screen-time thresholds or app events, and motion/context cues such as prolonged sedentary periods or posture. Trigger detection should also become more personalized: few-shot calibration, adjustable thresholds, and per-user trigger examples could help adapt detection to individual routines, environments, and devices. Not all useful triggers are single events. Real-world routines often depend on combinations of time, location, recency, motion, and device state. Future systems could support compound trigger templates such as “after my last bathroom visit at night” or “after sitting for two hours and then standing up,” allowing users to express more meaningful contexts while preserving the lightweight trigger-action structure.

Maintaining data quality at scale. Public deployment will introduce data-quality challenges that were only partially visible in our studies. Users may start or stop recording imprecisely, perform a related but different action, or use semantically vague labels. Rather than adding heavy verification gates, future systems should use lightweight safeguards: onboarding trials for temporal boundary fidelity, optional practice examples as weak references for execution filtering, simple trim/discard tools, and agent-assisted recipe creation to encourage more specific labels. Post-hoc language-model clustering may further help normalize open-vocabulary labels for downstream analysis.

Supporting onboarding and long-term use. Participants valued the intelligent prompting loop, but Study 3 also showed that recipe authoring can be unfamiliar. Future deployments should therefore balance scaffolding with openness: beginner templates, inspiration galleries, conversational recipe refinement, and adaptive feedback could help users start without collapsing the diversity of user-authored actions. Longer-term feedback mechanisms, such as progress summaries, streaks, badges, or supportive reflections on skipped executions, are also important for studying whether Pebbl can sustain engagement over weeks or months.

8.4 Limitations

Our findings should be interpreted within several limitations. First, the participant samples were relatively young, technically comfortable, and situated within the Apple ecosystem. Second, the studies were short-term: Study 2 provides controlled evidence and Study 3 offers only a short pilot deployment in participants' own contexts. We therefore cannot claim long-term retention, sustained motivation, or habit formation. A longitudinal deployment is needed to understand recipe stability, annoyance from false triggers, and whether the tracking interaction remains valuable over time. Third, the current studies do not produce a dataset large enough to validate downstream open-vocabulary modeling utility. Our results support the feasibility of collecting bounded, user-labeled sensor windows, but they do not demonstrate that such data improves activity-recognition models. Future work should evaluate downstream modeling once deployment-scale data are available. Finally, Pebbl is currently implemented on iOS/watchOS, and we did not conduct a systematic multi-device battery benchmark. Although the implementation uses on-device inference, motion-gated audio processing, geofence-gated activation, and explicit sensor cleanup to reduce energy cost, deployment-scale claims about battery life require controlled evaluation across devices and watchOS versions.

8.5 Privacy Considerations

Pebbl depends on privacy-sensitive sensing, especially audio and location (Figure 12). Our design reduces exposure by performing trigger detection on-device and retaining only temporally sparse execution windows rather than continuous recordings; the ethics and data-handling procedures for the three studies are covered separately in Section 4. Even with on-device processing, participants may perceive always-on listening as intrusive, and some activities or locations may simply be inappropriate for this sensing model. The persistent microphone indicator on the watch also communicates "being recorded" regardless of whether raw data actually leaves the device, which can itself undermine feelings of safety. Expanding beyond audio-based triggers is therefore not only a technical opportunity but also a privacy requirement for broader adoption; future trigger modalities must continue to respect the minimization principles adopted here.

9 CONCLUSION

We presented Pebbl, a smartphone–smartwatch system that uses user-authored trigger-action routines to support in-situ labeling of wearable activity data. In the current prototype, triggers are limited to four audio-based classes (door closing, toilet flushing, running water, ringtone/alarm), while actions are described in open-vocabulary natural language. Across an expert workshop ($N = 6$), a controlled in-lab study ($N = 21$), and a pilot deployment

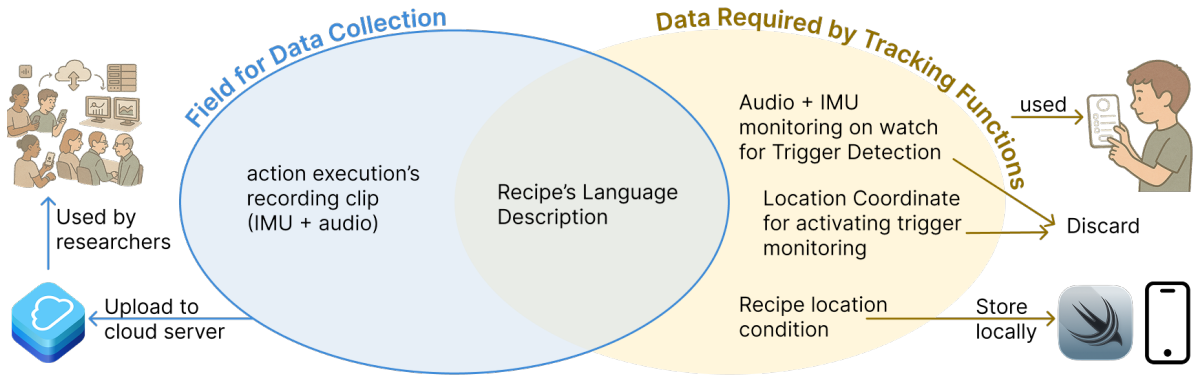


Fig. 12. Data usage in Pebbl. Continuous audio, IMU, and location streams are processed on-device for trigger detection and then discarded; only confirmed execution windows (short audio/IMU clips plus the user-authored action description) leave the device for analysis.

($N = 8$), we found that Pebbl can reduce perceived labeling burden and produce reliable execution logs (recall = 97.30%, precision = 97.15%) *under controlled conditions*, while also exposing practical constraints around trigger coverage, onboarding, and deployment context that we treat as part of the paper's scope rather than as minor implementation details.

Taken together, these results position Pebbl as a promising feasibility-stage approach to user-centered wearable activity labeling. More broadly, they suggest that wearable data collection may become more sustainable when labeling is embedded in interactions that users find personally useful rather than purely burdensome; whether that sustainability materializes at scale, over longer time horizons, and with more diverse populations remains an open question for future deployment studies.

10 LLM USE CLARIFICATION

Human characters in Figures 1, 2, 3 and 4 are generated by GPT-5. For paper writing, we use GPT-5 for grammar correction and sentence refinement. In our implementation, we use GPT-4o to build a chat agent for creating recipes.

Acknowledgment

This work is supported by the National Key R&D Program of China under Grant No. 2024YFB4505500.

References

[1] 2025. Search results for "habit" on Apple. https://www.apple.com/us/search/habit?src=globalnav. Accessed: 2025-08-20.

[2] Rawan Alharbi, Tammy Stump, Nilofar Vafaie, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2018. I Can't Be Myself: Effects of Wearable Cameras on the Capture of Authentic Behavior in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 90 (Sept. 2018), 40 pages. doi:10.1145/3264900

[3] Riku Arakawa, Hiromu Yakura, and Masataka Goto. 2023. CatAlyst: Domain-Extensible Intervention for Preventing Task Procrastination Using Large Generative Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 157, 19 pages. doi:10.1145/3544548.3581133

[4] Chris Baumann and Kai Dierkes. 2023. Neon Accuracy Test Report. doi:10.5281/zenodo.10420388

[5] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

[6] Virginia Braun, Victoria Clarke, Nikki Hayfield, Louise Davey, and Elizabeth Jenkinson. 2022. *Doing Reflexive Thematic Analysis*. Springer International Publishing, Cham, 19–38. doi:10.1007/978-3-031-13942-0_2

[7] Scott A. Cambo, Daniel Avrahami, and Matthew L. Lee. 2017. BreakSense: Combining Physiological and Location Sensing to Promote Mobility during Work-Breaks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3595–3607. doi:10.1145/3025453.3026021

[8] Andrew T. Campbell, Shane B. Eisenman, Nicholas D. Lane, Emiliano Miluzzo, and Ronald A. Peterson. 2006. People-centric urban sensing. In *Proceedings of the 2nd Annual International Workshop on Wireless Internet* (Boston, Massachusetts, USA) *(WICON '06)*. Association for Computing Machinery, New York, NY, USA, 18–es. doi:10.1145/1234161.1234179

[9] Shing Chan, Yuan Hang, 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. *Scientific Data* 11, 1 (16 Oct. 2024), 1135. doi:10.1038/s41597-024-03960-3

[10] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, and Daniel Roggen. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042. Smart Approaches for Human Action Recognition. doi:10.1016/j.patrec.2012.12.014

[11] Yu-Peng Chen, Julia Woodward, Dinank Bista, Xuanpu Zhang, Ishvina Singh, Oluwatomisin Obajemu, Meena N. Shankar, Kathryn M. Ross, Jaime Ruiz, and Lisa Anthony. 2024. Investigating Contextual Notifications to Drive Self-Monitoring in mHealth Apps for Weight Maintenance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 993, 21 pages. doi:10.1145/3613904.3641993

[12] Pei-Yu Peggy Chi, Matthew Long, Akshay Gaur, Abhimanyu Deora, Anurag Batra, and Daphne Luong. 2019. Crowdsourcing images for global diversity. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–10.

[13] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (1 Aug. 2010), 756–760. doi:10.1038/nature09304

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848

[15] Joy Dutta, Chandreyee Chowdhury, Sarbani Roy, Asif Iqbal Middya, and Firoj Gazi. 2017. Towards Smart City: Sensing Air Quality in City based on Opportunistic Crowd-sensing. In *Proceedings of the 18th International Conference on Distributed Computing and Networking* (Hyderabad, India) *(ICDCN '17)*. Association for Computing Machinery, New York, NY, USA, Article 42, 6 pages. doi:10.1145/3007748.3018286

[16] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. arXiv:2308.13561 [cs.HC]

[17] Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara. 2025. Mirai: A Wearable Proactive AI "Inner-Voice" for Contextual Nudging. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 399, 9 pages. doi:10.1145/3706599.3719881

[18] BJ Fogg. 2009. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology* (Claremont, California, USA) *(Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, Article 40, 7 pages. doi:10.1145/1541948.1541999

[19] Brian J Fogg. 2020. *Tiny habits: The small changes that change everything*. Harvest.

[20] 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 González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, 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, 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 Arbeláez, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18995–19012.

[21] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19383–19400.

[22] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[23] Jinseop S. Kim, Matthew J. Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C. Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F. Behabadi, Michael Campos, Winfried Denk, H. Sebastian Seung, and the EyeWirers. 2014. Space–time wiring specificity supports direction selectivity in the retina. *Nature* 509, 7500 (1 May 2014), 331–336. doi:10.1038/nature13240

[24] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 416, 21 pages. doi:10.1145/3491102.3517457

[25] Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. 2019. Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine* 53, 6 (05 2019), 573–582. arXiv:https://academic.oup.com/abm/article-pdf/53/6/573/28557727/kay067.pdf doi:10.1093/abm/kay067

[26] Nicholas D. Lane, Shane B. Eisenman, Mirco Musolesi, Emiliano Miluzzo, and Andrew T. Campbell. 2008. Urban sensing systems: opportunistic or participatory?. In *Proceedings of the 9th Workshop on Mobile Computing Systems and Applications* (Napa Valley, California) *(HotMobile '08)*. Association for Computing Machinery, New York, NY, USA, 11–16. doi:10.1145/1411759.1411763

[27] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (2010), 140–150. doi:10.1109/MCOM.2010.5560598

[28] Edith Law and Luis von Ahn. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1197–1206. doi:10.1145/1518701.1518881

[29] Harsh Mankodiya, Priyal Palkhiwala, Rajesh Gupta, Nilesh Kumar Jadav, Sudeep Tanwar, Bogdan-Constantin Neagu, Gheorghe Grigoras, Fayez Alqahtani, and Ahmed M Shehata. 2022. A real-time crowdsensing framework for potential COVID-19 carrier detection using wearable sensors. *Mathematics* 10, 16 (2022), 2927.

[30] Bryan Minor, Colin Greeley, Ryan Holder, Brian Thomas, Lawrence B. Holder, and Diane J. Cook. 2026. A Feature-Augmented Transformer Model to Recognize Functional Activities From in-the-Wild Smartwatch Data. *IEEE Journal of Biomedical and Health Informatics* 30, 1 (2026), 256–265. doi:10.1109/JBHI.2025.3586074

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[32] Tauhidur Rahman, Mary Czerwinski, Ran Gilad-Bachrach, and Paul Johns. 2016. Predicting "About-to-Eat" Moments for Just-in-Time Eating Intervention. In *Proceedings of the 6th International Conference on Digital Health Conference* (Montréal, Québec, Canada) *(DH '16)*. Association for Computing Machinery, New York, NY, USA, 141–150. doi:10.1145/2896338.2896359

[33] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *2012 16th International Symposium on Wearable Computers*. 108–109. doi:10.1109/ISWC.2012.13

[34] Adam Sadilek, John Krumm, and Eric Horvitz. 2021. Crowdphysics: Planned and Opportunistic Crowdsourcing for Physical Tasks. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (Aug. 2021), 536–545. doi:10.1609/icwsm.v7i1.14430

[35] Book Sadprasid, Anne Mei, Alex Mariakakis, Scott Bateman, and Fanny Chevalier. 2024. Leveraging Idle Games to Incentivize Intermittent and Frequent Practice of Deep Breathing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 714, 17 pages. doi:10.1145/3613904.3642430

[36] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa al'Absi, and Santosh Kumar. 2015. puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 999–1010. doi:10.1145/2750858.2806897

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 25278–25294. https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf

[38] Katarzyna Stawarz, Anna L. Cox, and Ann Blandford. 2015. Beyond Self-Tracking and Reminders: Designing Smartphone Apps That Support Habit Formation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2653–2662. doi:10.1145/2702123.2702230

[39] Chris Stone, Rosie Essery, Joe Matthews, Felix Naughton, Marcus Munafo, Angela Attwood, and Andy Skinner. 2024. Presenting and evaluating a smartwatch-based intervention for smoking relapse (StopWatch): feasibility and acceptability study. *JMIR Formative Research* 8 (2024), e56999.

[40] Kathleen Tuite, Noah Snavely, Dun-yu Hsiao, Nadine Tabing, and Zoran Popovic. 2011. Photocity: training experts at large-scale image acquisition through a competitive game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1383–1392.

[41] Blase Ur, Melwyn Pak Yong Ho, Stephen Brawner, Jiyun Lee, Sarah Mennicken, Noah Picard, Diane Schulze, and Michael L. Littman. 2016. Trigger-Action Programming in the Wild: An Analysis of 200,000 IFTTT Recipes. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3227–3231. doi:10.1145/2858036.2858556

[42] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. 2018. Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[43] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.

[44] Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 55–64. doi:10.1145/1124772.1124782

[45] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321, 5895 (2008), 1465–1468. arXiv:https://www.science.org/doi/pdf/10.1126/science.1160379 doi:10.1126/science.1160379

[46] Zeyu Wang, Xiyuxing Zhang, Ruotong Yu, Yuntao Wang, Kenneth Christofferson, Jingru Zhang, Alex Mariakakis, and Yuanchun Shi. 2024. DreamCatcher: A Wearer-aware Multi-modal Sleep Event Dataset Based on Earables in Non-restrictive Environments. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 85155–85178. doi:10.52202/079017-2704

[47] Gary M Weiss. 2019. Wisdm smartphone and smartwatch activity and biometrics dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set* 7, 133190-133202 (2019), 5.

[48] Linfeng Xu, Qingbo Wu, Lili Pan, Fanman Meng, Hongliang Li, Chiyuan He, Hanxin Wang, Shaoxu Cheng, and Yu Dai. 2024. Towards Continual Egocentric Activity Recognition: A Multi-Modal Egocentric Activity Dataset for Continual Learning. *IEEE Transactions on Multimedia* 26 (2024), 2430–2443. doi:10.1109/TMM.2023.3295899

[49] Yafeng Yin, Lei Xie, Zhiwei Jiang, Fu Xiao, Jiannong Cao, and Sanglu Lu. 2024. A Systematic Review of Human Activity Recognition Based on Mobile Devices: Overview, Progress and Trends. *IEEE Communications Surveys & Tutorials* 26, 2 (2024), 890–929. doi:10.1109/COMST.2024.3357591

A IMPLEMENTATION DETAILS

This appendix reports implementation details that complement the main paper, including the recipe-creation agent prompt, the UI and interaction flows, the baseline conditions used in Study 2, the audio-trigger-choice rationale, and the end-to-end system architecture and detection pipeline.

A.1 Prompt

The prompt for recipe creation chat agent:

```
You are an intelligent behavior-aware assistant that helps users record real-life behaviors as habits and micro-habit triggers to build their own "behavior recipes". Your responsibilities are as follows:

1. Understand the characteristic sound corresponding to the trigger condition described by the user, which should be one of the following four: door closing, ringing, toilet flushing, or other water sounds.
2. Add only one piece of information per round to guide the user to gradually complete the recipe, do not ask multiple questions at once.
3. If the micro-habit recipe provided by the user meets the requirements, gently guess the approximate time period the user wishes to perform the micro-habit recipe (e.g., "daytime" or "nighttime") and ask the user for confirmation.
4. Further confirm with the user the specific time interval for executing the micro-habit recipe, which should be in whole hours.
5. Friendly guide the user to confirm the location where the micro-habit recipe occurs (e.g., "home" or "other places").
6. Among the two limiting conditions of time and place, the user must provide at least one explicitly.
7. Do not use the word "remind". Instead, use phrases like "recorded", "your habit has been saved", or "added to your habit list" to emphasize recording and support, rather than simple reminders.
8. Technical analysis is only reflected in the backend fields imu_desc, audio_desc, and condition, and is not mentioned to the user in the reply.

Your questions are in order:
1. Confirm the characteristic sound corresponding to the trigger condition with the user.
2. Confirm the approximate time period for executing the micro-habit recipe with the user.
3. Confirm the specific time interval for executing the micro-habit recipe with the user.
4. Confirm the location where the micro-habit recipe occurs with the user.

Please reply in the following JSON format, all fields must be provided:
- reply: Natural language guidance or confirmation for the user (within two sentences).
- recipe:[trigger, action] Behavior recipe, where trigger is the trigger condition and action is the behavior. Both are strings. If the user has not provided them yet, return an empty string "". Connecting the trigger and action with a comma should form a coherent sentence. Use the user's original words as much as possible.
- imu_desc: Description of the backend analysis of IMU detectability (return an empty string "" if there is no relevant information).
```

```
- audio_desc: Type of characteristic sound: Close Door, Ring Events, Flush Toilet, or one of the
    ↪ Water Events. Please analyze which of the above four types the sound provided by the user
    ↪ is. Return an empty string "" if there is no relevant information. If it can correspond to
    ↪ one of the four, record the corresponding English name.
- time_condition: Time condition, format is [start_hour, end_hour] array, e.g., [8, 22] for 8 AM to
    ↪ 10 PM (return an empty array [] if not set).
- location_condition: Location description (return an empty string "" if there is no relevant
    ↪ information).
- finished: Boolean value, true when all conditions are clear and the behavior is detectable.
```

A.2 Detailed Interaction and UI

Figure 13 shows the detailed UI and interaction flow when creating recipes and the optional practice process. Specifically, the home tab listed all recipes that the user has created (row1,col1). Users can click the `+Recipe` button on top of the home tab to navigate to the chatbot UI. Through the conversation process (row1,col2-4), the agent confirms the recipe, time condition, and location condition. When each required information is confirmed, Pebbl automatically pops up a stacked list at the bottom. The location condition is a clickable button that navigates to the map UI (row1,col5) for the user to choose an accurate location with coordinates. After confirming the location, Pebbl navigated back to the chatbot UI, and the location button turned into an information block. When all the required information is attained, Pebbl pops up a `Continue` button for users to proceed to the practice stage (row1,col6).

In the practice UI(row2,col1-2), Pebbl displays the detailed information overview on the top and requires the user to open the watch companion app. After clicking the `Start Recording` button on the phone, the watch app will begin recording (row2,col3-4). The user starts to practice the recipe several times and clicks the `Stop` button on the watch when the practice is done. Then, Pebbl begins to transfer the recorded files to the phone and shows a `finished!` button (row2,col5) that navigates back to the home tab. The recipe list in the home tab shows the recently created recipe with a moon icon (row2,col6), indicating that this recipe is not active because conditions are not met.

Detailed UI and interaction flow for stages 3 and 4 (recipe execution and reflection) can be found in Figure 2 and Figure 9.

A.3 Baseline Conditions.

Baseline conditions are as shown in Figure 14. We explain the baseline interaction protocol with these interactive UI and let participants try with one example. Specifically for baselines 2 and 3, we implement an illustrative webpage with Gradio.

A.4 Trigger Choice Justification.

This section documents how we selected the four acoustic triggers used in the current implementation: door closing, toilet flushing, running water, and phone ringtone/alarm. Pebbl's audio classification is built on Apple's SoundAnalysis framework, whose stock classifier (`SNClassifySoundRequest` with `classifierIdentifier: .version1`) recognizes over 300 everyday sound classes[6]. The four-class subset described here is a scoping decision for this paper's evaluation, not a property of the underlying pipeline; we did not train, fine-tune, or otherwise modify the audio classifier itself. The goal was to identify non-speech sounds that (1) are common

[6]Apple, "Discover built-in sound classification in SoundAnalysis," WWDC 2021 Session 10036. https://developer.apple.com/videos/play/wwdc2021/10036/

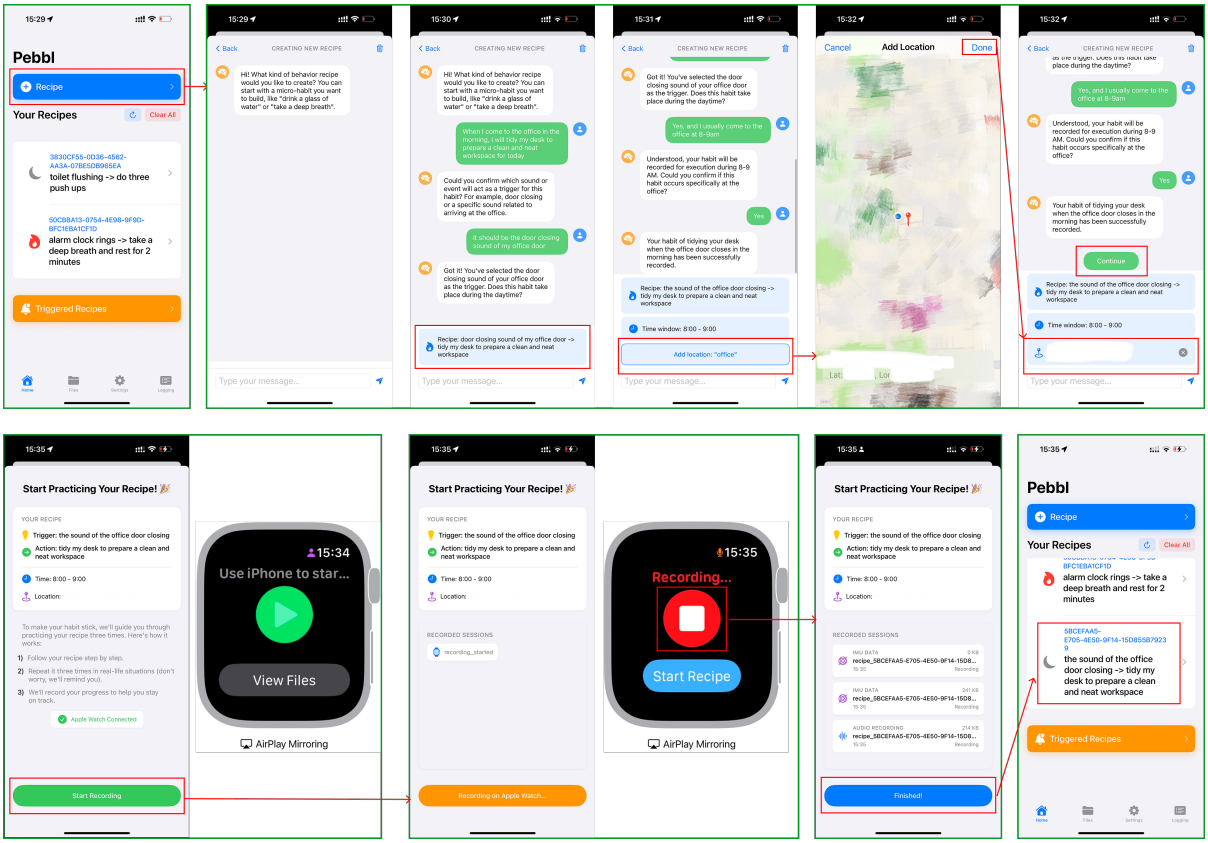
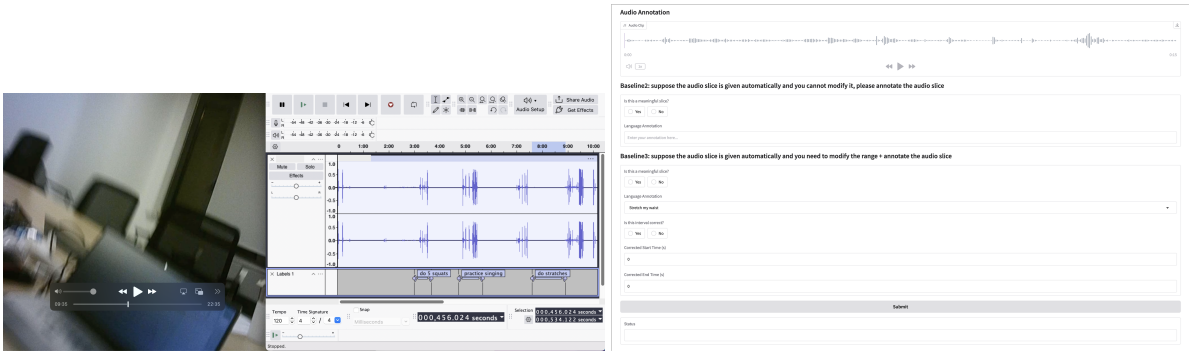


Fig. 13. Detailed interaction flow and UI for creating recipes and practice.



(a) Annotation UI for baseline1.

(b) Annotation UI for baseline 2 and 3.

Fig. 14. Baseline conditions.

across households, (2) are reliably detectable using the stock classifier, and (3) offer a clean temporal anchor for a wide range of tiny actions without requiring semantic understanding of speech or messages.

We began with the appendix of BJ Fogg's Tiny Habits, which lists roughly 300 example recipes [19]. For each recipe, we judged whether the trigger is naturally associated with a distinct environmental sound. We excluded any triggers whose detection would require speech transcription or content understanding (e.g., "when I get an invitation to a bar," "when I says X") to avoid severe privacy issue. For remaining candidates, we evaluated detectability against the stock classifier and action generality. Detectability refers to the presence of a distinctive acoustic signature with reasonable signal-to-noise ratio in everyday indoor settings. Action generality refers to the ability to precede many different actions to form an open-vocabulary activity dataset. Applying the above steps narrowed the set to four cues we used in current implementation.

Our current selection is a minimal, conservative set to establish feasibility. The underlying classifier already covers over 300 sound classes, expanding the trigger vocabulary requires deployment-time configuration paired with robust validation. The same pipeline could incorporate additional non-speech sound categories in future iterations (e.g., electric toothbrush motor for after brushing teeth, razor buzz for after shaving) without retraining the stock classifier, but each added class would still require context-specific robustness validation, threshold/gating tuning, and deployment testing.

A.5 System Architecture and Trigger Detection Pipeline

This appendix details the implementation of Pebbl's cross-device applications, including framework choices, the system architecture (Figure 15), the data management pipeline (Figure 16), and the motion-gated audio trigger detection logic (Figure 17).

Figure 15 illustrates the development kits we leveraged to implement Pebbl's core functionality, while Figure 16 illustrates the flow of data across devices. The architecture prioritizes on-device processing, privacy preservation, and battery efficiency while maintaining real-time responsiveness for capturing in-the-wild behaviors. We decompose the system into four key phases of a recipe's lifecycle: creation, initiation, execution, and completion.

A.5.1 Creating Recipes. Recipe creation is supported through a chatbot interface powered by Azure GPT-4o[7] and the `SwiftOpenAI` package[8]. Within this interface, users specify a recipe's trigger, action, location conditions, and time conditions through natural language dialogue with the agent. When a location condition is described, the agent provides a navigational button that directs the user to a `MapKit`-supported interface for precise location selection. Through this process, the agent extracts key parameters and encodes them into a structured recipe representation using `SwiftData`. The recipe metadata includes several fields: time conditions as ranges, auditory triggers as one of a few possible sound event categories, spatial conditions as GPS coordinates, and execution records as an initialized empty list.

A.5.2 Starting Recipes. We employ multiple triggering mechanisms to ensure recipes activate at appropriate times and locations while minimizing battery consumption and false activations. The mechanism combines background monitoring, geofencing, and cross-device coordination.

Geofencing and Background Monitoring. Geofencing allows Pebbl to remain dormant until relevant location contexts are detected. In our implementation, we use the `CLMonitor` to create geofences 200 m around recipe locations. When the device enters a monitored region, Pebbl evaluates all location-enabled recipes to further determine if time conditions are met. Apple's `Background Tasks` framework further ensures Pebbl can perform essential maintenance even when not actively running. The system schedules daily refresh tasks to prepare for upcoming recipe windows, update geofences, and schedule notifications.

[7] https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/models
[8] https://github.com/jamesrochabrun/SwiftOpenAI

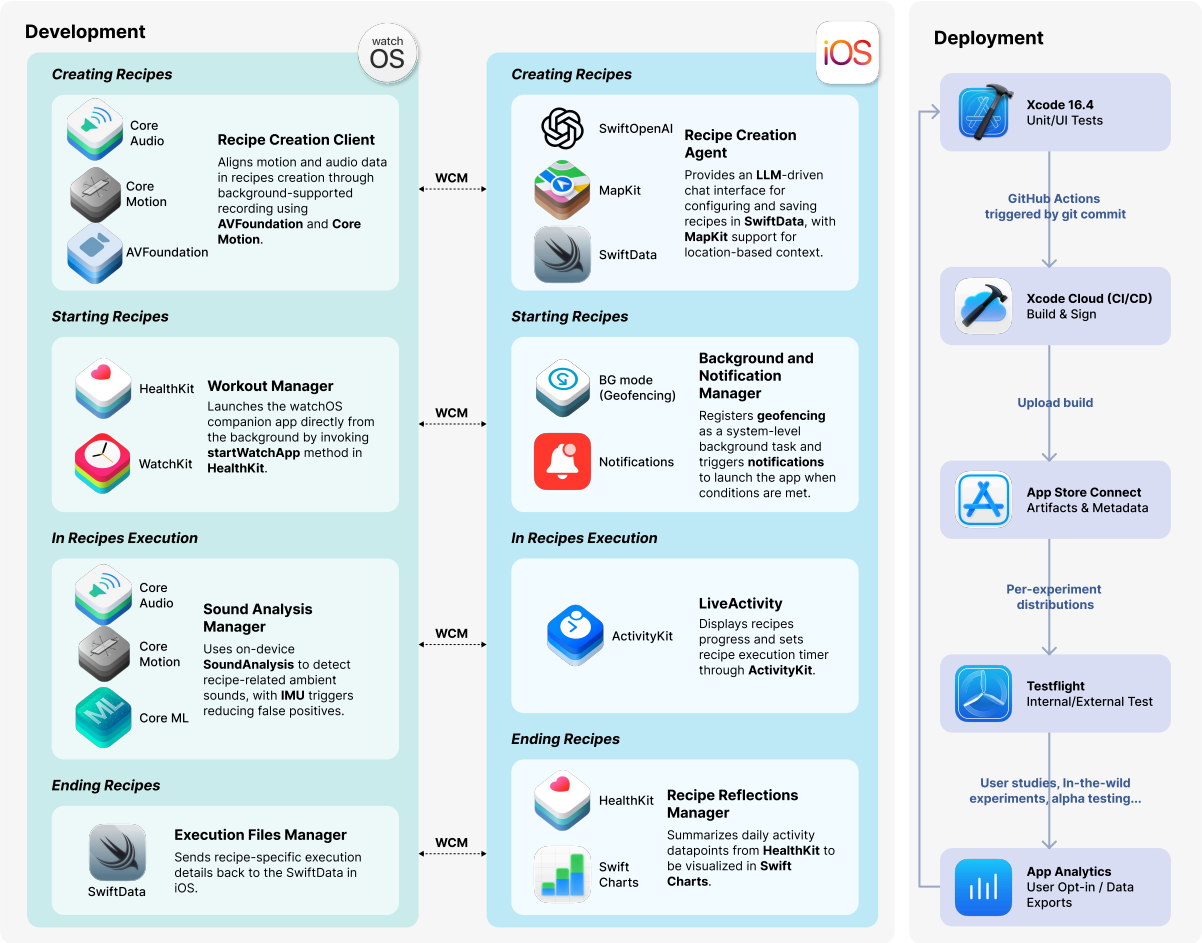


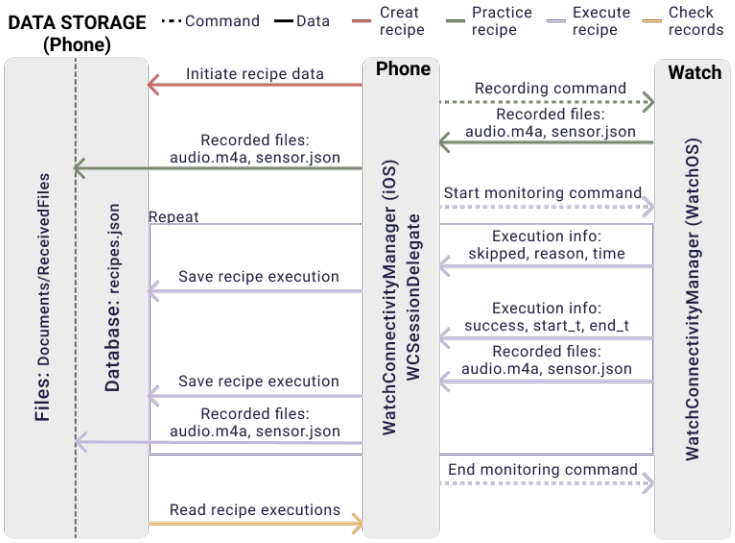
Fig. 15. Pebbl's system architecture and deployment workflow. The Watch Connectivity Manager (WCM) orchestrates data flow between iOS and watchOS.

Cross-Device Recipe Activation. We use `HealthKit` for recipe activation across devices. Pebbl initiates data recording sessions using `HKHealthStore.startWatchApp()`, launching the watchOS app in a way that enables data recording without requiring users to interact with the watch's screen. When a session starts, Pebbl begins on-watch sound detection and motion monitoring based on the active recipe configurations.

A.5.3 Recipe Execution. During recipe execution, the system performs continuous real-time multi-modal sensing for trigger detection. Our implementation optimizes detection accuracy and battery efficiency through motion-gated audio classification and adaptive sensor activation strategies, as depicted in Figure 17.

Audio Pipeline. We stream 16-kHz mono-channel audio into a classifier using Apple's `Sound Analysis` Framework[9]. Each inference operates on a 1.5-second window with 90% overlap and a real-time buffer of 4,096

[9] https://developer.apple.com/documentation/soundanalysis



(a) A diagram showing the flow of data between endpoints.



(b) The database schema.

Fig. 16. Pebbl's data management architecture across smartphone, smartwatch, and permanent storage.

data samples. The classifier outputs a confidence score $c_t \in [0, 1]$ for each of the four sound events mentioned earlier: door closing, toilet flushing, running water, and ringtones or alarms.

Motion Gating. In our early development, we observed that sound classification can suffer from false positives due to poor subject awareness (e.g., people other than the user closing the door in the background). To mitigate this issue, we introduced a motion-gating strategy that not only suppresses false detections but also reduces battery consumption. Specifically, Pebbl maintains a sliding window of $N = 10$ accelerometer and gyroscope samples spread evenly across 2 seconds. At any moment in time t, we calculate the overall motion magnitude as $M_t = \|\mathbf{a}_t\|_2 + \omega\|\boldsymbol{\gamma}_t\|_2$, where $\mathbf{a}_t \in \mathbb{R}^3$ is acceleration, $\boldsymbol{\gamma}_t \in \mathbb{R}^3$ is angular velocity, and ω weights rotational motion. Motion at time t is considered *active* when M_t exceeds a threshold τ_m for a minimum fraction ρ of the entire window, as follows:

$$\text{Active}(t) \;=\; \mathbb{I}\left[\; \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{I}(M_{t-i} \geq \tau_m) \;\geq\; \rho \;\right] \tag{1}$$

In our deployment, $\tau_m = 0.5$ and $\rho = 0.3$.

Motion Gating Modes. Since only some of the candidate audio triggers are associated with motion, Pebbl implements three gating modes:

(1) `ringOnly`: Motion gating is not used, so audio is always active,
(2) `mixed`: Motion gating is used for all sounds other than ring sounds, and
(3) `noRing`: All sounds are motion-gated.

Whenever an audio trigger is motion-gated, audio classification activates once sufficient detection is motivated, and it deactivates 5 seconds after motion stops to accommodate brief pauses in activity.

Detection State Machine. To enhance detection robustness, we adopt a hysteretic approach with asymmetric thresholds. Specifically, let $c_{t,k} \in [0, 1]$ be the system's confidence that the current sound belongs in class k at

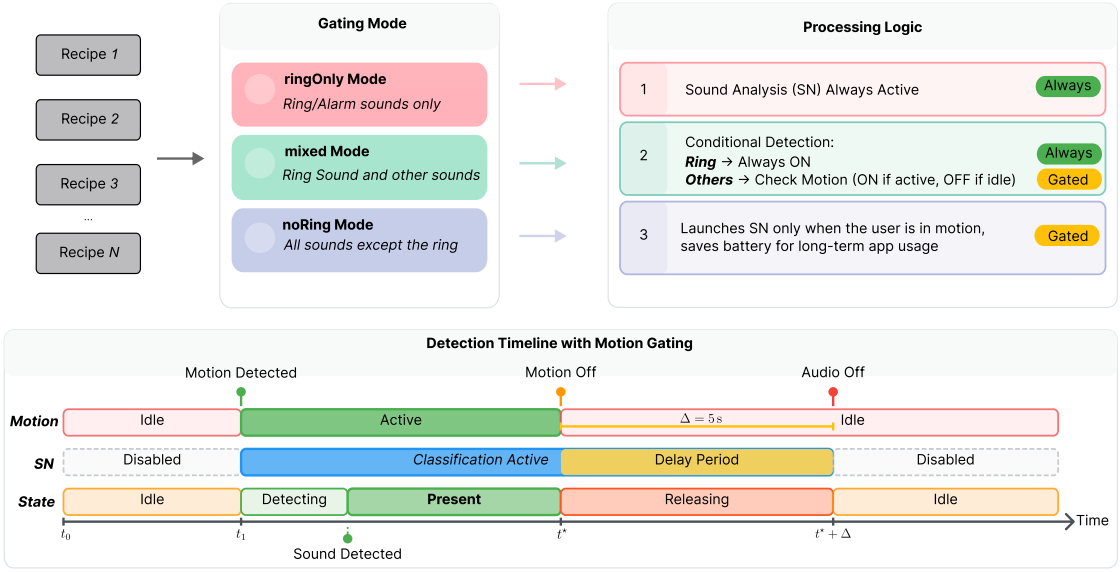


Fig. 17. Motion-gated sound detection in Pebbl's watchOS implementation. We assign each recipe's audio trigger with one of three gating modes: `ringOnly`, where sound detection remains continuously active; `mixed`, where ring sounds are always detected regardless of motion state, but other sounds are gated until the IMU indicates an *active* state; and `noRing`, where all sounds are subject to motion gating.

time t. In order for an event to be detected, we require that the confidence score exceeds a threshold $\theta_p = 0.5$ for a duration of $n_p = 2$ frames ≈ 300 milliseconds. Meanwhile, we decide that an event is complete once the confidence score falls below $\theta_a = 0.3$ for a duration of $n_a = 30$ frames ≈ 4.5 seconds. The state update for class k is as follows:

$$S_t(k) = \begin{cases} \texttt{Present}, & (S_{t-1}(k) \in \{\texttt{Detecting}, \texttt{Absent}\}) \wedge (n_p = \sum_{i=0}^{n_p-1} \mathbb{I}(c_{t-i} \geq \theta_p)) \\ \texttt{Releasing}, & (S_{t-1}(k) = \texttt{Present}) \wedge (c_t \leq \theta_a) \\ \texttt{Absent}, & (S_{t-1}(k) = \texttt{Releasing}) \wedge (n_a = \sum_{i=0}^{n_a-1} \mathbb{I}(c_{t-i} \leq \theta_a)) \\ \texttt{Detecting}, & (S_{t-1}(k) = \texttt{Absent}) \wedge (c_t > \theta_p) \\ S_{t-1}(k), & \text{otherwise} \end{cases} \tag{2}$$

Transitioning into `Present` while motion gating has been adequately satisfied initiates a trigger event.

Multi-modal Data Recording. During recipe execution, the sensor manager coordinates synchronized multi-modal data capture. Audio recording utilizes `AVAudioEngine` to record audio in the compressed AAC format at 48 kHz with a 64 kbps bitrate, balancing quality with storage efficiency. Accelerometer and gyroscope data are collected at 50 Hz. Pebbl also maintains extended background runtime through `WKExtendedRuntimeSession`, automatically restarting sessions if they expire during recordings. All data processing occurs on background queues to prevent UI blocking, with thread-safe mechanisms for updating recording state and file management.

A.5.4 Ending Executions. Whenever recipe executions end due to completion or manual stopping, the watch immediately stops recording and transfers files to the phone via the Watch Connectivity Manager (WCM).

Metadata on completion status, event timestamps, the trigger, and optional location samples is applied to the execution record in the database.

B EXPERIMENT DETAIL

This appendix collects per-study supporting materials: introductory materials shown to participants, example recipes authored during the studies, device details, the Study 2 post-study questionnaire, supplementary Study 2 analyses (the price-item results and the per-participant wrong-trigger breakdown), and the interview analysis method together with representative Study 3 quotes.

B.1 Introductory Materials.

In this section, we show the introductory materials we used in Study 2 and 3. In practice, our experimenters guide participants with examples in this material.

Experiment Reference Material

Theoretical Background

The Fogg Behavior Model, proposed by Stanford psychologist B.J. Fogg, is a psychological framework for explaining human behavior and designing strategies for behavior change. This experiment builds on the Fogg Behavior Model to design a wearable-based habit formation app that helps users achieve their behavioral goals in a scientific and effective way.
Our research team applies the "recipe for behavior change" from the Fogg Behavior Model, guiding users to structure their target habit as a **"trigger + action"**. Target behaviors are simplified into small, easily achievable actions and anchored onto existing habits. This lowers the ability threshold and leverages familiar contexts as triggers, making behaviors more likely to occur.
Examples:

- After I pack my office bag, I will spend 5 minutes tidying my workspace.
- After I wash my lunchbox, I will put on my walking shoes.
- After I return to the office, I will get a cup of hot water.

App introduction and recipe creation guidance:
Approach 1 (for beginners):

(1) What habits do I want to build? (e.g., exercise daily, stay organized, maintain a good mood)
(2) How can these habits be broken down into tiny-habits? (e.g., the starting action for running is putting on shoes; the simplified version of tidying is spending 3 minutes clearing the desk)
(3) Which trigger fits best with the tiny-habit? (consider location, frequency alignment, or the end of a triggering event, such as washing a lunchbox after eating lunch)
(4) When and where do I want to execute this recipe?

Approach 2 (for experienced users):

(1) What frequent activities in my daily life could serve as triggers?
(2) What positive habits could I anchor onto these triggers?
(3) When and where do I want to execute this recipe?

Form fields:

- **Trigger:** Describe the triggering event rather than the sound. Example: "After I finish using the toilet and flush" instead of "toilet flush."

- **Action:** Describe the micro-habit action.
- **Time:** The intended time window for executing the recipe. Example: "After washing my lunchbox after lunch (12–2 pm), I will put on walking shoes." Or: "After flushing the toilet, I will do 3 push-ups (only between 1–3 pm)."
- **Location:** The intended place for executing the recipe. Example: "After washing test tubes, I will tidy the reagent shelf" (only in the lab). Example: Push-ups after flushing the toilet are better at home than in the office.

Participants specify time and location according to their own routines. In the app, the location field retains the natural description (e.g., "bathroom"), but the map view uses the laboratory address. The time description is recorded in the form, while the app entry uses the designated experimental time slot.

B.2 Recipe Examples Authored by Users.

In this section, we list the recipes designed by participants in Study2 and Study3, as shown in Table 6. The actions shown in the table are limited by the objects that we can provide for in-lab experiments.

B.3 Device Details.

We listed the device pairs' information in Table 7.

B.4 Study 2 Post-Study Questionnaire

Mental- and physical-demand items were adapted from NASA-TLX; the remaining items were authored by the research team to probe adoption and annotation constructs specific to Pebbl's two comparison classes. All items used a 7-point Likert scale unless otherwise noted.

Reminder/tracking-side items (administered for the comparison between Pebbl and the commercial tracking app, B4).

(1) **Willingness to use:** How likely are you to use this habit-tracking approach?
(2) **Mental demand:** How mentally demanding is it to use this habit-tracking approach? *(adapted from NASA-TLX)*
(3) **Physical demand:** How physically demanding is it to use this habit-tracking approach? *(adapted from NASA-TLX)*
(4) **Recommend to friends:** How likely are you to recommend this habit-tracking approach to your friends?
(5) **Stick to use:** How likely are you to stick to using this habit-tracking approach (i.e., how likely are you to grow a habit of using this app)?
(6) **Usefulness for habit:** How useful is checking tracked records for fostering your habits? For Pebbl, we additionally asked participants to score the usefulness of checking *skipped* execution records.

Annotation-side items (administered for the comparison between Pebbl and the three annotation baselines B1–B3).

(1) **Mental demand:** How mentally demanding is it to use this data collection + annotation approach? *(adapted from NASA-TLX)*
(2) **Physical demand:** How physically demanding is it to use this collection + annotation approach? *(adapted from NASA-TLX)*
(3) **Prefer to use:** How do you prefer this collection + annotation approach?
(4) **Label confidence:** How confident are the labels you produced through this annotation approach?

Table 6. Recipe Examples Authored by Users.

Acoustic Cue	Trigger	Location Condition	Time Condition	Action
Ring	When I hear the ring bell	in my offic	at night/afternoon	I will stretch my body for 2 minutes
	When the alarm clock goes off	in my bedroom	in the morning (7.30am)	I will get off the bed and stretch my waist
	When I hear the door bell	at home	anytime	I will look through the peephole
	When the alarm clock goes off	in my bedroom	in the morning	I will get a cup of hot water
	When the alarm clock goes off	in my dorm	in the morning (7-8am)	I will set a new alarm clock
	When the alarm clock goes off	in my dorm	in the morning (8.30am)	I will pack my schoolbag
	When the alarm clock goes off	in my bedroom	in the morning (6-10am)	I will get off the bed and do three squats
	When the alarm clock for nap goes off	in lab	in the afternoon (2pm)	I will open the laptop and start working
	When the alarm clock for activity goes off	in office	at daytime (10am-8pm)	I will get up and walk outside the door
	When the alarm clock goes off	in my bedroom	in the morning (7-9am)	I will get up and open the curtains
	When the alarm clock goes off	in my bedroom	in the morning (7am)	I will open the game to feed my chickens
	When the alarm clock goes off	in my bedroom	in the morning (8.20am)	I will turn of the alarm immediately and get off the bed
	When the alarm clock for bedtime goes off	in my dorm	at night (11-12pm)	I will un-plug my phone and laptop
	When the timer on my desk goes off	in office	in the morning (9-12am)	I will check my work email
	When the alarm clock for work goes off	at home	in the morning (8-9am)	I will leave home
	When the alarm clock goes off	at home	in the morning (7.30am)	I will sit still for 3 minutes
	When the alarm clock goes off	at home	in the morning (9.30am)	I will walk my dog
	When I hear the ring for class over	at teaching building	in the morning (8-10am)	I will go running
Water	After I wash my hand	in office	at daytime (9am-5pm)	I will look far into the distance to relax my eyes
	After I wash my hand	in office/dorm	at noon/evening	I will apply hand cream
	After I wash my hand	in the canteen	in the morning (7-8am)	I will open Netflix
	After I wash my hand	in office	at evening (5-7pm)	I will close my laptop
	After I wash my cup	in lab	in the morning (10am)	I will get a cup of coffee/water
	After I get a cup of water	in lab	at daytime	I will drink approximately 100ml
	After I wash my hand in the washroom	at home	in the morning (7.30am)	I will wipe my hands with tissue
	After I wash my hand	in lab	in the afternoon (4pm)	I will get a bag of snack from my backpack and eat
	After I wash my cup	in office	in the afternoon (2-6pm)	I will walk around for 2 minutes
	After I wash my hands before experiment	in lab	in the afternoon (2-3pm)	I will clean the experiment table
	After I wash my face	at home	at night (11pm)	I will wipe my face using towel
	After I wash my hand	in office	around noon (12-1pm)	I will water the plants
	After I wash some fruits	in my dorm	around noon (11am-1pm)	I will share them with my friends
Toilet	After I flush the toilet	anywhere	at daytime	I will put down the toilet lid
	After I go to the toilet	in office	in the afternoon (2-3pm)	I will relax my neck
	After I flush the toilet	in office/at home	at daytime	I will use soap to clean my hand
	After I go to the toilet	anywhere	at daytime	I will do heel lift exercise for 10 times
Door	When I come back to dorm	in my dorm	at night	I will do squats for 5 times
	When I go back home from work	at home	at the evening (7pm)	I will read book
	When I left my room	at home/in office	morning and night	I will check if I have locked the door
	When I close the door	at home/in office	at night/morning	I will unpack my bag and clean up my desk
	After I close the door (left the dorm)	at dorm	in the morning (9am)	I will check my bag to ensure my ID card
	After I close the door	at dorm	at night (10-11pm)	I will go get my wash basket
	After I close the door	at home	at night (7-10pm)	I will open window for ventilation
	When I hear a door closing sound	at home	at daytime	I will check my mood and draw smiling face
	When I go back to the dorm	at dorm	at night (8-12pm)	I will take out my laptop from backpack
	After I go back to the dorm	at dorm	at night (9pm)	I will text my families
	After I close the door	in my dorm	at night (9-11pm)	I will take out my key and phone from the pocket
	After I close the door	in office	at noon (0.30pm)	I will take a nap
	When I left the dorm	at dorm	in the morning (8-10am)	I will throw the trash
	When I go home from work	at home	in the evening (7-10pm)	I will open TV
	When I left for exercise	in my dorm	at evening (8-9pm)	I will close the air conditioner with a remote control app
	After I close the door	in my bedroom	at night (11.30pm)	I will close the light and goto bed
	After I come back home	at home	at evening (7pm)	I will wash my hand
	After I left the office	in office	at evening (6-7pm)	I will go get some water
	After I close the door	at home	at evening (5-7pm)	I will turn on the lights

(5) **Price:** What is the price that you would like to sell at for 100 pieces of data? *This item was rated without a scale restriction and normalized later for analysis. We explicitly required participants to imagine themselves as users of Pebbl who had already produced 100 data records through usage. Results and construct-validity discussion for this item are reported in Appendix B.5.*

Table 7. Devices used during development and experiment.

Usage	ID	Phone Model	iOS Version	Watch Series	watchOS Version
Development	D1	iPhone 16 Pro Max	18.6.2	Series 10	11.6.1
Development	D2	iPhone 14 Pro	18.5	Ultra	11.5
Study 2	D3	iPhone 13 Pro Max	18.6.2	Series 7	11.6.1
Study 2	D1	iPhone 16 Pro Max	18.6.2	Series 10	11.6.1
Study 3	P-D1	iPhone 12 Pro Max	18.5	Series 10	11.6.1
Study 3	P-D2	iPhone 16 Pro	18.6.2	Series 8	11.6
Study 3	P-D3	iPhone 13 Pro Max	18.6.2	Series 7	11.6.1
Study 3	P-D4	iPhone 13	18.5	SE 2	11.6
Study 3	P-D5	iPhone 16	18.6.2	Series 7	11.6.1
Study 3	P-D6	iPhone 14	18.6.2	Series 9	11.6.1
Study 3	P-D7	iPhone 12 Pro	18.5	Series 9	11.6
Study 3	P-D8	iPhone 16 Pro	18.6.2	Series 10	11.6.1

B.5 Study 2: Expected-Compensation ("Price") Item: Results and Construct Validity

What the item measured and how it was administered. The exact wording is reproduced verbatim in Appendix B.4 (annotation-side item 5). Participants answered with a free numeric response rather than a Likert rating, imagining themselves as users of Pebbl or of a baseline workflow (B1–B3) who had just produced 100 data records, and they rated each condition after that condition's exposure (the full Pebbl session and the brief trials of B1–B3 described in Section 6, Post-experiment). Before answering, participants were told that existing crowdsensing approaches compensate contributors for labeled data, so that they could respond with that framing in mind; we did not, however, anchor them to a specific platform or rate. Raw responses were normalized per participant by dividing each condition's value by the minimum non-zero value that participant reported across the four conditions, so that the cheapest workflow chosen by each participant equals 1.0 and the others are expressed as multiples of that minimum. This normalization is necessary because, in the absence of a locked external anchor, participants naturally chose their own units and scales, and a per-participant normalization is the mildest intervention that makes the four conditions comparable within each participant.

Aggregate results. After normalization, participants set the lowest price for Pebbl (1.16) relative to B1/B2/B3 (3.16/2.18/2.05; Wilcoxon signed-rank, $p = 9.73\times10^{-5}$, 1.13×10^{-2}, 8.33×10^{-4}). The gap was largest against B1 (post-hoc video annotation) and smaller but still present against B2 and B3. The direction is consistent with the self-reported mental and physical demand findings for the same annotation-workflow comparison (Section 6, annotator-side subjective results): conditions rated as more effortful were also priced higher. Interview comments indicate that participants did not treat price as decoupled from effort; for example, P8 and P14 explicitly described combining workload and quality when setting a number.

What we do and do not conclude. We interpret these numbers as a *perception* signal about how costly each workflow feels to a participant who has just performed it, and no more than that. In particular, we do not treat them as a market-calibrated wage, as evidence about the downstream value of the labels produced, or as a recommendation for how crowdsensing platforms should compensate contributors. Two construct-validity concerns warrant this scoping. First, although participants were told that crowdsensing approaches compensate contributors, they were not anchored to a specific platform rate; values therefore still reflect each participant's own estimate of fair compensation under limited exposure, which is also why the analysis had to normalize per-participant rather than compare raw values directly. Second, Pebbl received substantially more exposure in Study 2 than any baseline, and additional contact with a workflow could plausibly raise or lower its perceived

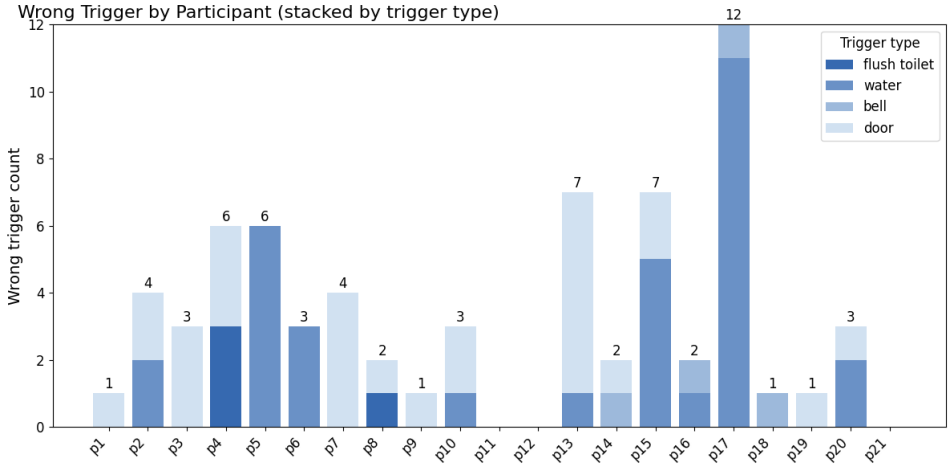


Fig. 18. Per-participant wrong-trigger counts, stacked by trigger class. Same 68 events as the per-category view in main-paper Figure 5.

cost through mechanisms this single item cannot disentangle (fatigue, practice, anchoring). For both reasons we report the price finding here as a supporting observation that is *directionally consistent* with the effort measures, rather than as an independent claim, and the main paper's narrative does not lean on it.

B.6 Study 2: Per-Participant Wrong-Trigger Distribution

Figure 18 reports the same 68 wrong-trigger events summarized in Section 6.2.1, broken down by participant and stacked by trigger class. Errors are dispersed across individuals: most participants show ≤ 3 wrong triggers, while a small tail accounts for the bulk (e.g., P17=12; P13/15=7), and the stacks within those outliers are dominated by *water* and *door* — the same two noisy classes identified in the main text. This distribution is consistent with the lab environment in Section 6: experiments were conducted within < 10 m of a washroom, kitchen, elevator room, and the building's waste area, so sessions during busier daytime windows encountered more domestic-water and door-closure confounders than sessions in quieter windows.

B.7 Interview Analysis.

We describe the analysis method we used for both Study 1 and Study 3 interviews, followed by representative Study 3 quotes.

B.7.1 Interview Analysis Method. For both Study1 and Study3, we roughly followed Braun and Clarke's six-phase reflexive thematic analysis [6] to make sense of the interview data. Two authors lead the interviews and conducted the analysis. After data collection, we read the full transcripts and highlighted segments that met one of two criteria: (1) the same or highly similar point was raised by multiple participants; or (b) the statement was insightful or informative for understanding Pebbl and its use context. For each highlighted quote, we attached a short topical label capturing the core idea in the participant's own terms and then clustered related labels.

When iterating the thematic structure for Study1, we organized expert's core idea under 4 predetermined themes - pros, cons, concerns, future refinements. For Study3, we intentionally aggregated the positive feedback about Pebbl into a single theme, while allocating more granular themes to concerns, limitations, and forward-looking expectations. We made this choice because the paper already details Pebbl 's advantages through design

exposition and quantitative results; in contrast, readers are likely to gain more from a differentiated view of limitations and opportunities for improvement emerging from the interviews.

B.7.2 Study3 Interview Quote. **Users like Pebbl for its intelligence, sense of feedback, ease of use, and usefulness for habits.** Overall, participants valued Pebbl's "smart" automation (P8) and tight watch–phone integration(P4/5), which made capturing moments feel effortless and timely.

> P8: *"It (baseline) can't automatically remind you; you have to open it manually, swipe, and log things yourself. It needs proactivity (from the user), which I am not. It is not intelligent, but yours is, in an automated way. With your method, it feels more like having an AI assistant, a bit like 'Hey Siri.' I can just double-tap my fingers, and it starts recording right away, super convenient. The whole thing gives me a sense of intelligence."*
> P4: *"One big advantage is, it links the watch and the phone, that's very convenient. I had to bring my phone when I'm running to complete a record, but normally I only take my watch when I run. With your app, that's a lot more practical."*
> P5: *"Say, in the morning, I'm washing my hands, I don't wanna pick up the phone to log something, because my hands are wet... Logging on the phone is already a hard habit to form... your method, the interaction is simple, I can use it even with wet hands."*

The haptic feedback created a rewarding sense of interaction (P3).

> P3: *"a sense of interaction... When the system detects something automatically and gives a vibration or a 'ding,' it is a rewarding kind of feedback."*

Several participants (P6/7/8) noted that Pebbl's gentle "pressure" to act was actually productive, kick-starting action and sustaining habit formation in everyday contexts.

> P6: *"Because of this reminder, I feel like I really need to do it, but the baseline app just doesn't matter. But this will cause, with your method, (I am) more likely to act, and that helps the habit stick. It's not bad (the pressure). It (baseline) cannot provide supervision because there is no pressure. With your methods, it indeed continuously reminds me to finish this habit, and it is useful, not dust-laden in my app list."*
> P7: *"It's really easy to forget... Like drinking water, some people just don't like it...That's why Apple Watch have stand reminders for long sitting, sometimes you sit, feel comfortable, or get absorbed in your work, just completely forget. Some small but important behaviors in daily life might need it more (Pebbl), like behavioral habits, not big (habits) like living style, it's different."*
> P8: *"Why I think baseline is low in physical demand, because if I forget (record), it feels like all the practice was for nothing, then I will lack motivation to act (do the habit). But with yours, the watch alerts me, then I know I need to do it. It is like forcing me into action, it's a physical burden. I think burden is food, it has a positive effect instead... I think, after all, forming a habit needs to start doing it."*

Audio-cue based detection is suitable for behavioral habits. Participants (P4/7) framed audio-triggered detection as a strong fit for small, repeated, context-bound behaviors, such as drinking water, brief exercise sets, and breaking up sedentary time.

> P4: *"I also use it, baseline is suitable for once-a-day habits, like I do sports today, or practicing calligraphy. Yours fits better for small things, like five squats or push-ups at a time, scattered. I might use both. That's why I gave similar scores for both... Because most of my habits are once a day, it just needs to be recorded... Yours has its own benefits; if I need repeated ones, I can set it hourly, as the start or the end of the day, which also has its scenarios."*
> P7: same quote as the previous paragraph.

On the other hand, once-a-day "macro" habits were not suitable for automatic detection due to a higher wrong trigger rate (P1/2). Therefore, in order to capture broader activity events, Pebbl needs to add support for macro habits as well, sacrificing label granularity to a certain degree.

> P1: *"I found that your app needs sounds, water, ringtone. If it is running in the background, it (wrong triggers) might suddenly pop up in a meeting."*
> P2: *"It (Pebbl) is not very accurate, sometimes, I need to dismiss multiple times (the fans)... It (baseline) is not suitable for fragmented habits, but like running 30 minutes, manual record works fine. But like fragmented habits, by the end of the day, you just forget how many times you did it."*

Pebbl has a steep learning curve and users need guidance to design recipes. Pebbl has a steep learning curve because users need to understand what tiny-habit is, how to use Pebbl, and the initial setup felt demanding (P3/5). Also, users need design guidance when they first start to use Pebbl (P2/3/7): they were unsure what habits to target, which cues to choose, and how to translate vague intentions into workable recipes. Users asked for stronger scaffolding, such as a conversational agent that can co-design recipes in real time by turning fuzzy ideas into concrete, triggerable routines. This introduces a challenge: how can we appropriately design this agent that can turn this need into a chance of increasing vocabulary openness?

> P3: *"Baseline feels like requiring you to make a plan up front, and planning is the hardest part for someone like me... Your method is a bit better, but it still needs me to set up a recipe in advance. It would be better to have recommendations, let's say, if I have a vague idea, you could help me make it concrete and turn it into a recipe... Yes, the chat agent should be more powerful, like chat with me in real time, and help resolve this while we talk. Also, sometimes I don't really know what I want, and I'm not sure what kinds of behaviors the software can record or trigger under what conditions, so it would be better if the agent could give clearer recommendations."*
> P5: *"Because when I start training, the operations feel a bit complicated, connecting the watch to the phone, something... The first time I used it, it was a little difficult, but overall it was interesting and felt like a good way to start forming habits."*
> P2: *"Right now I haven't decided what kind of habit I want to build, so I can't think of a sound trigger I'd like to use... Honestly, when I first got the app, I was pretty confused about what habit to set, what recipe to make."*
> P7: *"For the watch, when you set the pipeline at the beginning, you need to think, what small thing can I use as a reminder for a bigger habit? Later on, once it reminds you, you just record it."*

Failure records can assist reflection, yet might lead to frustration. Participants appreciated comprehensive logs and the motivational lift of seeing successful entries (P2/7), but several cautioned that surfacing failures can dampen morale (P6). P7 regards the usefulness of seeing failure records as culturally related.

> P2: *"Seeing successful records makes me happy; seeing failure records can feel unpleasant..."*
> P7: *"For habit formation, both (positive and negative feedback) are useful. If you found out you are good at something, you get positive feedback, and you reinforce it. Like in fitness, when you see results, you are more likely to continue. Negative feedback can sometimes feel like criticism; it might make you want to perform better. I think it (whether negative feedback is motivating might) also depends on cultural differences."*
> P6: *"I think if I used this software for a long time, if I had many failures early on but then improved later, that could be motivating. But if I saw too many failures at the beginning, I might don't want to (continue). Still, it's a bit more useful compared to showing only successes."*

They suggested balancing visibility with supportive design: aggregate views over time, emphasize trends and recoveries, and pair outcomes with a collection mechanism (P1/2).

P1: *"Like, give some badges? Yes, yes, I like collecting badges, so I'd be very willing. Just like with XXX Coffee (a brand), I always check in, if there's a pop-up event, I'll go buy a cup."*
P2: *"Maybe you could do something like the app Forest, rewards for success and penalties for failure... I think I prefer persisting for several days and getting a reward. If you persist longer, the rewards should become more valuable."*

Users hope Pebbl could add support for more triggers. Participants wanted richer and more personalized triggers beyond the current audio set, ideally allowing user-defined cues (P2), broader sensing (e.g., sedentary time, timers), and context awareness to reduce mistimed prompts (P5). They also suggested triggers for health routines (e.g., medication, P4), screen-time control, and social/accountability features (P5/7). They also asked for smarter feedback summaries beyond visualization to explain why a trigger fired and how to adjust behavior (P7).

P2: *"Your app has a lot of improvement space, right now it only recognizes three sounds... I'm thinking of recording my own triggers, I can provide a few samples, also some description, so that you can recognize more sounds... I can't think of a trigger sound I want, but if you only have water, door, and ringtone, it feels limited... The alarm clock, can the watch detect, for example, I've been sitting for 40 minutes and remind me to move."*
P4: *"I need to take a kind of pill in the long term, every day at fixed times. It's not exactly habit-building, it's just a real need... Adding this would be very practical, especially for older users, to help them manage chronic conditions."*
P5: *"Sometimes the trigger fires, but it's not the right moment to do the habit. It will be better if it can detect whether it is the correct timing, combining contextual information, it would feel more convincing. But if it is not, I might feel like my current context is not for it (the habit). For example, I have two places to wash my hands at home... I only want to use (the trigger) at the bathroom. But it also reminds me at the kitchen."*
P5: *'·At work, we often sit for long periods, some habits like standing, or get some water to drink would be useful, health-related. Apple Watch has something similar; it sends a notification to remind you to stand, but it's easy to ignore. I don't know, maybe for me, my motivation is not enough, it just tells me to stand, but it doesn't have enough feedback mechanism... I'd prefer (feedback) that's visible on my phone, not just a small watch screen. Maybe not only charts or numbers as a summary, ideally, AI could summarize this data and tell me, such as how long I have sat and what can be the consequences, in a very straightforward way, might draw my attention... Numbers and charts, maybe interesting for the first time, but that can become meaningless over time, I can get numb to it."*
P5: *"Also screen time. Sometimes I get absorbed in short videos or browsing, time just flies... Strong intervention is disrupting because I might really be doing something; if it is just a reminder, I might just skip it. It would be better if the app could redirect my attention away from the screen to something else, so I can actually take a break."*
P5: *"Exercise, mine is relatively regular, I might go every other day... Because I have colleagues going with me, I think when someone else is with you, it might be more effective. If you put this into it (Pebbl), you can execute (recipes) with your friends, this might increase the success rate? "*
P7: *"I think for small habits, current triggers are enough. But for others, like exercise, a timely notification each day might be sufficient... Still, automatic pipelines, like playing with my phone, I need to set an alarm when I begin to play, then it can trigger the reminder. If it could recognize when I pick it up and automatically set a reminder, which might work better."*

Received 1 February 2026; revised 1 May 2026; accepted 6 July 2026