HabitSense: A Privacy-Aware, AI-Enhanced Multimodal Wearable Platform for mHealth Applications

GLENN J. FERNANDES, Northwestern University, USA
JIAYI ZHENG, Northwestern University, USA
MAHDI PEDRAM, University of North Texas, USA
CHRISTOPHER ROMANO, Northwestern University, USA
FARZAD SHAHABI, Northwestern University, USA
BLAINE ROTHROCK, Northwestern University, USA
THOMAS COHEN, Northwestern University, USA
HELEN ZHU, Northwestern University, USA
TANMEET S. BUTANI, Northwestern University, USA
JOSIAH HESTER, Georgia Institute of Technology, USA
AGGELOS K. KATSAGGELOS, Northwestern University, USA
NABIL ALSHURAFA, Northwestern University, USA

Wearable cameras provide an objective method to visually confirm and automate the detection of health-risk behaviors such as smoking and overeating, which is critical for developing and testing adaptive treatment interventions. Despite the potential of wearable camera systems, adoption is hindered by inadequate clinician input in the design, user privacy concerns, and user burden. To address these barriers, we introduced HabitSense, an open-source[1], multi-modal neck-worn platform developed with input from focus groups with clinicians (N=36) and user feedback from in-wild studies involving 105 participants over 35 days. Optimized for monitoring health-risk behaviors, the platform utilizes RGB, thermal, and inertial measurement unit sensors to detect eating and smoking events in real time. In a 7-day study involving 15 participants, HabitSense recorded 768 hours of footage, capturing 420.91 minutes of hand-to-mouth gestures associated with eating and smoking data crucial for training machine learning models, achieving a 92% F1-score in gesture recognition. To address privacy concerns, the platform records only during likely health-risk behavior events using SECURE, a smart activation algorithm. Additionally, HabitSense employs on-device obfuscation algorithms that selectively obfuscate the background during recording, maintaining individual privacy while leaving gestures related to health-risk behaviors unobfuscated. Our implementation of SECURE has resulted in a 48% reduction in storage needs and a 30% increase in battery life. This paper highlights the critical roles of clinician feedback, extensive field testing, and privacy-enhancing algorithms in developing an unobtrusive, lightweight, and reproducible wearable system that is both feasible and acceptable for monitoring health-risk behaviors in real-world settings.

[1] https://github.com/HAbitsLab/HabitSense

Authors' addresses: Glenn J. Fernandes, Northwestern University, Chicago, Illinois, USA; Jiayi Zheng, Northwestern University, Chicago, Illinois, USA; Mahdi Pedram, University of North Texas, Chicago, Illinois, USA; Christopher Romano, Northwestern University, Chicago, Illinois, USA; Farzad Shahabi, Northwestern University, Chicago, Illinois, USA; Blaine Rothrock, Northwestern University, Chicago, Illinois, USA; Thomas Cohen, Northwestern University, Chicago, Illinois, USA; Helen Zhu, Northwestern University, Chicago, Illinois, USA; Tanmeet S. Butani, Northwestern University, Chicago, Illinois, USA; Josiah Hester, Georgia Institute of Technology, Atlanta, Georgia, USA; Aggelos K. Katsaggelos, Northwestern University, Chicago, Illinois, USA; Nabil Alshurafa, Northwestern University, Chicago, Illinois, USA.

2474-9567/2024/9-ART101
https://doi.org/10.1145/3678591

CCS Concepts: • **Computing methodologies → Activity recognition and understanding**; *Video segmentation*; *Scene understanding*; • **Human-centered computing → Mobile devices**; • **Hardware → Sensor devices and platforms**; • **Applied computing → Consumer health**.

Additional Key Words and Phrases: wearable, multimodal, eating, smoking, thermal, camera, vision transformers, privacy, machine learning

ACM Reference Format:
Glenn J. Fernandes, Jiayi Zheng, Mahdi Pedram, Christopher Romano, Farzad Shahabi, Blaine Rothrock, Thomas Cohen, Helen Zhu, Tanmeet S. Butani, Josiah Hester, Aggelos K. Katsaggelos, and Nabil Alshurafa. 2024. HabitSense: A Privacy-Aware, AI-Enhanced Multimodal Wearable Platform for mHealth Applications. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 101 (September 2024), 48 pages. https://doi.org/10.1145/3678591

1 INTRODUCTION AND BACKGROUND

In the United States, obesity represents a significant public health concern, affecting approximately 40% of adults [1, 2]. Obesity is closely associated with an elevated risk of various chronic diseases, including heart disease, certain types of cancer, and other long-term health issues [1, 2]. Similarly, cigarette smoking remains the leading cause of preventable death [3–5]. Health-risk behaviors such as overeating, excessive food consumption beyond physiological needs, and tobacco smoking are major contributors to premature morbidity and mortality worldwide. [6–8]. Eliminating these behaviors can help prevent several chronic diseases. However, this requires understanding and modifying what individuals put into their mouths. Accurate and efficient detection of consumption patterns will improve our knowledge of these behaviors and facilitate the development of effective interventions to prevent them.

In current approaches to understanding the factors influencing the daily habits of individuals, clinicians often rely on self-report methods to document food consumption, timing, and psychological factors related to the behavior of interest. Although self-report methods effectively capture psychological factors, these methods may introduce bias, impose a significant burden on participants, and are frequently susceptible to errors from forgetfulness [9–12]. Integrating automatic, objective measures would significantly enhance the accuracy in documenting the timing, patterns, and context surrounding consumption behaviors [13, 14]. The increasing popularity of wearable technology, such as smartwatches, offers the potential for automated behavior monitoring, but these devices lack essential visual confirmation capabilities for real-life applications. *Wearable cameras offer a promising solution by providing visual evidence that enhances contextual understanding and offers more reliable approaches to automated confirmation of health-risk behaviors.* Therefore, there is a need for objective, unobtrusive wearable camera systems that can enhance the diagnosis of health-risk behaviors, allowing for personalized treatments and real-time triggers to test interventions that improve treatment outcomes [15, 16].

Wearable cameras developed by the ubiquitous computing community have shown promise in monitoring behaviors such as overeating and smoking. However, their adoption, particularly in clinical settings, faces significant challenges, primarily due to the lack of clinical involvement in the design process [17]. This oversight often results in systems that do not align with the specific needs and workflows of healthcare providers, preventing their integration into routine clinical practice [17]. Engaging clinicians early in the development process is essential to ensure that the physical design and capabilities of the system are practical and valuable in real-world healthcare settings. [18]. However, the role of clinician involvement in designing wearable cameras and qualitatively analyzing potential obstacles that might impede their acceptance and feasibility in healthcare settings has not been extensively investigated.

Based on feedback from our focus groups with clinicians (N=36) and experience from user studies involving wearable camera systems (N=105, across studies), privacy concerns continue to be a significant barrier to adopting wearable cameras. To address user privacy concerns, recent studies have explored the concept of activity-oriented cameras (AOCs) [19]. In contrast to traditional surveillance and egocentric cameras, AOCs are designed to focus

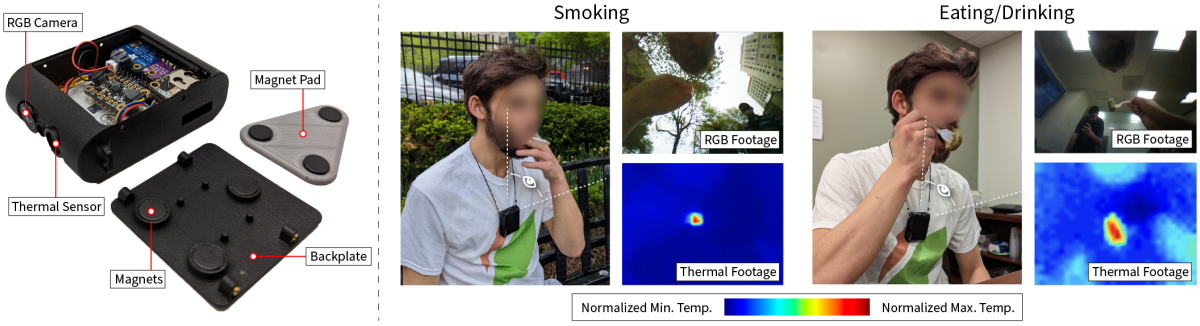


Fig. 1. The HabitSense device (left) and its applications (right). HabitSense is a wearable multi-modal system (thermal, RGB, accelerometer) optimized for privacy, enabling all-day recording and on-device processing to detect health-risk behaviors associated with smoking and eating.

only on capturing the intended activity of interest, inspired by the principle of least privilege (PoLP)[20, 21]. For example, one approach to capturing hand-to-mouth gestures involves orienting the lens of a neck-worn camera toward the mouth of the ***wearer*** (the person wearing the camera) and obfuscating background pixels to prevent capturing ***bystanders*** (individuals other than the wearer) or unintended behaviors in the video.[22]. Research has demonstrated a significant increase in system acceptability among both AOC wearers and bystanders while simultaneously maintaining the utility of visually confirming hand-to-mouth behavior [23]. However, there currently exists no AOC-based wearable system capable of running such obfuscation algorithms on-device in real time, while also detecting health-risk behaviors.

Developing an AOC-based wearable camera system that can capture eating and smoking gestures while running obfuscation algorithms in real time presents two challenges. First, the device must operate continuously for a full, 16-hour day [24]. Second, the camera must be computationally capable of running real-time obfuscation algorithms with the least possible impact on battery life. Smart activation mechanisms can improve battery life by using a low-energy sensor to detect preliminary signs of specific behaviors, activating more power-intensive tasks only as needed [25]. Given that eating and smoking are intermittent activities, this approach will allow selective activation of obfuscated recordings only *when a sensor detects preliminary gestures of these activities*. This will not only conserve battery life but also enhance privacy by ensuring that only detected health-risk behaviors are recorded.

Given the need for a sensor that can detect preliminary gestures of interest to activate high-energy processes and considering the necessity to obfuscate the user's background for privacy, we decided to build a wearable camera that incorporates an infrared (IR) thermal sensor alongside RGB. The choice of an IR thermal sensor is strategic because it allows the detection of the wearer and any objects they might handle by exploiting temperature differentials. Because the wearer is closest to the camera lens, their body temperature will create a distinct thermal signature, effectively distinguishing them from the background, enabling methods that distinguish the wearer and foreground from the background [22].

In this paper, we aim to synthesize diverse research strands surrounding wearable cameras into a comprehensive whole, providing clinicians and users with a privacy-preserving, unobtrusive wearable camera platform capable of providing visual evidence and detecting health-risk behaviors that users are willing to wear throughout the day. In doing so, HabitSense can assist clinicians in improving treatment efficacy by identifying the contextual factors that drive these behaviors when they occur to aid in personalizing treatment strategies. To achieve this

goal of addressing challenges related to acceptance, privacy, and user burden, our paper presents the following contributions:

(1) *To increase clinical adoption and inform design*, we report on insights from a qualitative analysis of focus group discussions involving 36 weight management and smoking treatment specialists to inform the need, design, and functionality of wearable cameras.
(2) *To increase user adoption*, we leveraged our extensive iterative experience (N=105, in all studies) in wearable camera design to develop HabitSense (illustrated in Figure 1), an RGB thermal (RGB-T) wearable camera that is designed to be easy to use, unobtrusive, reproducible, lightweight, and capable of operating all day (16 hours). HabitSense is also designed to detect eating and smoking behaviors while incorporating on-device obfuscation to enhance privacy.
(3) *To enhance user privacy and reduce burden*, we developed the Sensor-Enabled Control of Ubiquitous Recording and Evaluation (SECURE) algorithm. This algorithm efficiently activates obfuscated RGB video recording, successfully capturing 90% of hand-to-mouth gestures associated with eating and smoking while turning on the camera only 52.4% of the time. As a result, it achieved a 48% reduction in data storage and a 30% reduction in battery usage, significantly optimizing device efficiency.
(4) *To evaluate our device*, we tested HabitSense in a natural setting with 15 participants (8 with obesity, 7 smokers), yielding 768 hours of footage, and reported on the high user acceptability of the HabitSense device.
(5) *To validate the utility of our device*, we trained machine learning models on RGB, thermal, and obfuscated video data to run offline and in real time. We then rigorously evaluated their ability to recognize eating, smoking, and hand-to-mouth gestures and reported on the performance of our models. Our comprehensive assessments demonstrated a 92% F1-score in hand-to-mouth gesture recognition, highlighting the effectiveness and versatility of our models. Additionally, our eating and smoking detection models proved to be robust across various conditions, including night time, low-light settings, and during intense head movements.

Although this paper focuses on smoking and eating, HabitSense enables a paradigm shift in real-time behavior monitoring to test the effectiveness of timely interventions for multiple health-risk behaviors, including substance abuse, alcohol consumption, and medication adherence.

2 RELATED WORK

2.1 Current Wearables for Detection of Eating and Smoking Activities

This section highlights a few of the existing wearables for eating and smoking detection and their limitations. A more comprehensive discussion is provided in Supplementary Section 9.1.

Wrist-worn inertial measurement unit (IMU) sensing modalities are recognized for their ability to detect eating and smoking activities; however, they face challenges in practical application. These include false positives from unrelated movements, issues with sensor positioning (i.e., dependency on the dominant hand), and difficulties in generalizing across different individuals [16]. However, IMU-based sensors continue to provide utility by supplementing multi-modal devices [26].

Acoustic-based wearables [27–29] have shown promise in capturing chews, swallows, and jaw motion for eating and respiratory monitoring [30] to detect smoking inhalation. However, microphones are susceptible to noise interference, particularly in free-living environments. Moreover, using microphones raises significant privacy concerns in everyday settings[19].

Alternate-based sensing modalities attempt to capture smoking behavior by capturing multiple consecutive behaviors, such as the use of a smart lighter [31], or by combining detection of respiratory patterns with hand-to-mouth gestures [32]. Although current methods for automated detection of eating and smoking are promising,

	Work	Popular Device Name	IMU	Thermal	RGB Camera	Form Factor	Privacy	Participants	Meals	Smoking Events	In-lab vs Free-living	Free-living Days	Clinician Study	Hours of Data	Detection Targets
Eating	Bedri et al. 2017	EarBit	●					10	16		free-living	NR		45	chewing sounds, chewing motion
	Bi et al. 2018	Auracle						14	26		both	NR		2+32	chewing sounds
	Bedri et al. 2020	FitByte	●			●		23	69		both	NR		91	chewing and swallow motion, h2m gestures
	Morshed et al. 2020		●				●	28	1,305		free-living	21		NR	wrist motion
	Nyamukuru et al. 2020	Tiny Eats						20	20		in-lab			11	chewing sounds
	Zhang et al. 2020	NeckSense	●				●	20	40		free-living	14,2		470	chewing motion, forward-lean angle
	Bedri et al. 2022	FitNibble	●				●	12	NR		free-living	18		NR	h2m gestures, chewing motion
	Bi et al. 2022					●		10	NR		free-living	NR		55	h2m gestures
	Morshed et al. 2022		●					18	54		in-lab			11.5	chewing motion
	Shin et al. 2022	MyDJ	●				●	24	94		free-living	7		237	chewing motion
Smoking	Sazonov et al. 2013	PACT					●	20		40	in-lab			20	breathing motion, h2m gestures
	Lopez-Meyer et al. 2015	PACT					●	20		40	in-lab			20	breathing motion, h2m gestures
	Echebarría et al. 2017							2		6	in-lab			NR	breathing sounds
	Añazco et al. 2018		●					4		4	in-lab			NR	h2m gestures
	Senyurek et al. 2019		●					35		443	both	1		821	h2m gestures
	Senyurek et al. 2019	PACT2.0	●					35		443	both	1		871	h2m gestures
	Tai et al. 2020						●	NR		NR	in-lab			NR	nicotine biomarkers
	Imtiaz et al. 2020					●		10		50	free-living	1		108	h2m gestures
	Alharbi et al. 2022	SmokeMon			●		●	19		115	free-living	NR		110	heat signature
Privacy	Chattopadhyay & Boult 2013					●		●							
	Wrinkler & Rinner 2010					●		●							
	Wrinkler et al. 2014					●		●							
	Fernandes et al. 2024	**HabitSense**	●	●		●	●	●	15	522	283	free-living	7	✓	768	h2m gestures, heat signature

Key: NR Not Reported; Smartwatch; Earbuds; Necklace; Microphone; Eyeglasses; Hat; Vest; Nasal Cannula; Sweatband

Table 1. Comparison of HabitSense and existing systems. This table highlights the distinctions between HabitSense and other systems in terms of sensor modalities, form factor, privacy protection, and the nature of studies conducted

deploying these methods in free-living settings often relies on self-reporting or external tools for confirmation, which can introduce inaccuracies and biases into the ground truth data given there is no visual evidence to confirm the occurrence of the events.

2.2 The Need for Wearable Camera Systems for Detecting Eating and Smoking

To circumvent problems of bias in self-reporting, researchers have turned to wearable camera devices that continuously record activities, providing rich contextual information and reliable visual confirmation for ground truth data. The introduction of SenseCam marked a notable stride in this direction–a wearable digital camera capable of capturing a comprehensive record of the wearer's daily activities through still images and sensor data [33]. Building on this concept, Bedri et al. [34] integrated a camera into an eyeglasses-based system that used IMUs and proximity sensors to detect eating behaviors. Similarly, head-mounted video cameras [35] have been developed specifically for video-based eating detection. In both cases, the addition of cameras provides visual ground truth from the wearer's perspective, enhancing data reliability. Wearable camera systems in the form of eyeglasses have also been used for smoking detection in free-living settings [36]. Such eyeglasses-based systems, outfitted with multi-modal sensing capabilities, excel in gesture recognition for eating and smoking detection tasks. However, the eyeglass form factor impacts its generalizability. Privacy concerns, practical issues such as hair obstructing the camera, using a single eyeglass frame type, and discomfort or reluctance to wear glasses among some individuals further impede adoption [36]. Addressing some of these issues, recent research introduced a thermal camera in a necklace form factor for smoking detection [24]. The authors conducted a free-living study with the device, collecting thermal data as participants self-reported their smoking activities. However, the reliance on potentially inconsistent and inaccurate self-reported logs posed a challenge due to the lack of visual evidence for confirmation. Building on earlier discussions of multi-modal systems with IMUs

and camera-based systems using RGB and thermal technologies, there is potential for developing a multi-modal platform that integrates RGB, thermal imaging, and IMU sensors.

2.3 The Need for Clinical Feedback in Informing the Design of Wearable Systems

Wearable technology is increasingly recognized as a vital component in healthcare, offering the potential to passively monitor health, assist in diagnosis, and enhance patient care [37–39]. However, several challenges impede the adoption of wearable technology in clinical practice, such as device accuracy, privacy concerns, and cost, as well as limited clinical involvement in design and technical and interoperability issues [37, 40–45]. Integrating wearable devices into medical practice necessitates clinician input to ensure device effectiveness and relevance in clinical settings. Qualitative analysis with clinicians has shown potential in integrating clinician feedback to improve mobile health (mHealth) technology and increase adoption [17, 46, 47].

2.4 Sensing Multiple Health Behaviors with Privacy-Preserving Wearable Cameras

Using AOCs along with obfuscation algorithms to obfuscate backgrounds—including bystanders and sensitive activities—has proven effective in mitigating privacy concerns [19, 23]. Surveys [23] have examined the balance between privacy concerns and data utility using various obfuscation techniques, including blurring, applying a canny edge algorithm, entirely blacking out the background, or using Generative Adversarial Network (GAN) -based cartoon obfuscation [48] in activity-oriented videos captured using neck-worn cameras. User studies have also demonstrated that these techniques effectively reduce privacy concerns [23, 48]. However, implementing these algorithms in real time on resource-constrained devices is yet to be tested and presents a challenge.

Several camera systems employ on-device obfuscation algorithms to improve privacy [49–51]. However, these systems are not designed as wearable cameras that can be worn all day on a single charge, and they lack comprehensive user studies. Recent research has highlighted the potential of thermal cameras for obfuscation [22]. Thermal data can isolate the wearer's foreground activities from the background. Although this approach is promising, it was tested only in offline scenarios. Additionally, it lacks a comprehensive assessment of how well obfuscation and the use of thermal-only versus RGB-only data impact detecting activities such as hand-to-mouth, smoking, and eating gestures. Thus, further extensive investigations are necessary to refine thermal camera efficiency for real-time, on-device applications in real-world settings. This collective evidence suggests significant potential for developing an AOC RGB-T wearable camera system that uses the thermal sensor for obfuscation over the need for complex computer vision or deep learning algorithms on the device.

Based on these motivations, the novelty of HabitSense stems from integrating four key aspects: (1) incorporation of feedback from clinician focus groups and user studies into its development, (2) on-device obfuscation capabilities, (3) intelligent activation of obfuscated recording to enhance privacy and conserve battery life, and (4) a comprehensive evaluation framework that assesses the system's performance across a variety of real-world scenarios. Table 1 compares our innovative features to existing methods for detecting health-risk behaviors associated with eating and smoking.

3 CLINICIAN-INFORMED DESIGN: INTEGRATING FOCUS GROUP FEEDBACK INTO WEARABLE CAMERA DEVELOPMENT

To learn about the importance and feasibility of using privacy-conscious RGB-T wearables for monitoring various health-risk behaviors, we engaged in focused discussions with healthcare professionals experienced in managing eating and smoking habits. We recruited clinical dietitians (N=18) and tobacco treatment specialists (N=18) to ascertain problems faced in their practices, to inform design priorities from their experiences, to assess their willingness to use and adopt our proposed wearable camera system, and to identify practical challenges in implementing the system into treatment programs. Each session included either dietitians or tobacco treatment

specialists; none included participants from both groups. All participants worked in the Chicago area and had at least two years of professional experience. The focus groups took place during the early stages of the development of the HabitSense system.

In the focus group sessions, the moderator used a topic guide to ask questions targeting four user-centered design elements: **Motivation:** What measurement tools do the clinicians have, and what are their limitations?; **Desirables:** What additional measures do clinicians want to improve their understanding of their patients' behaviors?; **Advantages:** How could a passive-sensing visual/thermal wearable with real-time activity recognition benefit clinical practice?; **Barriers:** What challenges do practitioners foresee in utilizing a system like ours? To avoid biasing responses to questions about the clinicians' experiences with contemporary measures, no information about the proposed system was given until after addressing the Desirables portion of the focus group. The proposed system shown to participants was an early prototype that incorporated design lessons learned across Gen 1 and Gen 2 deployments. In addition to planned questions, the moderator asked follow-up questions and encouraged participants to elaborate on their answers and comment on other participants' answers.

We structured our analysis to reflect the role of our participants as administrators, rather than end-users, of the proposed system and to account for the value of their domain expertise in our design process. First, we separated focus group transcripts by topic and labeled the participants. Two authors then independently analyzed and thematically coded the transcripts. Once independent code lists were generated, the authors met to compare lists, resolve discrepancies, and produce and iterate upon a common set of themes. All unique responses that were "seconded" (i.e., repeated or affirmed) by another participant in the same session were included as themes. Additional themes were derived inductively and proposed by the authors in keeping with standard thematic analysis practices [52]. Proposed themes that achieved consensus between authors were added to the final list of themes, detailed in Table 2.

Unsurprisingly, there was a high similarity in how dietitians and tobacco treatment specialists answered the moderator's questions. Both types of practice hinge on providers' understanding of habitual health-risk behaviors for any given patient. Both providers rely on similar measurement tools and, though measuring different behaviors, experience similar measurement limitations.

3.1 Motivation: Existing Measures and Their Limitations

Treatment providers in both groups reported that eating and smoking are traditionally measured by self-report, in which patients either estimate their dietary/smoking activities over a given time range or record them in real time via a paper journal or software application. Providers further confirmed that real-time self-reporting is burdensome and thus achieves limited adherence. Several participants from both groups indicated difficulty obtaining patient self-reports that are timely, accurate, and consistent over time. The participants reported that comprehensive calorie or tobacco intake accounting requires too much work for the typical patient. Additionally, dietitians reported that the quality of self-reported eating data correlates to the dietary knowledge of the patient. For instance, patients with low dietary knowledge frequently neglect to report sauces and their ingredients because they are unaware of their importance as food items. In both groups, some providers reported encouraging patients to use mobile applications for self-reporting. These applications can modestly improve reporting accuracy for some patients but ultimately suffer from the same adherence problems stemming from the burden involved in any form of continuous self-reporting.

Providers reported that memory recall approaches (where a trained specialist guides memory exercises to prompt maximum recollection of events) to smoking and dietary monitoring address the burden problem by combining the reporting of multiple smoking or eating events into a single recall event. However, both groups of providers reported that memory recall suffers its own limitations, namely forgetting and dishonesty, both of which variably bias patient data.

Motivation (what do they have?)		**Themes**	**Design Considerations**
Existing measures		Self-report	-
		Memory recall	-
Existing measures limitations	*self-report:*	Inadherence	Low burden
	self-report:	Inaccuracy/dishonesty	Real-time detection
	self-report:	Dietary knowledge *(dietitians only)*	RGB, All-day wear, RGB buffer*
	memory recall:	Forgetting/dishonesty	Real-time detection, All-day wear
Desirables (what do they want?)			
		Behavior Context	RGB, FOV, Orientation, Resolution
		Other substance-use *(tobacco TS only)*	RGB, Thermal, FOV, Orientation
		Food budget/Income *(dietitians only)*	Cost
Advantages (how could wearables help?)			
		Real-time treatment monitoring	Real-time detection
		Behavior trigger identification	RGB, RGB buffer*
		Improved reporting accuracy	Real-time detection, All-day wear
		Fine-grain features of behavior	RGB, Thermal, Model robustness
		Food temperature *(dietitians only)*	Thermal
Barriers (what challenges do they predict?)			
		Detection failure cases	Model robustness
		Dishonesty	-
		Operational conditions	Waterproofing*, Drop-safety
		Social stigma	Form-factor, Acceptability
		Privacy	FOV, Obfuscation
		Cost *(dietitians only)*	Cost

Table 2. Themes and associated design considerations. Themes are categorized based on those identified in focus group sessions and are organized by participant group and linked to relevant design considerations. *Retained for future work.

3.2 Desirables: Desired Information About Patients

Because we are designing new modalities of behavior measurement, we asked participants not just what problems they experience but also what information they would want about their patients in a hypothetical "perfect world" in which providers can magically obtain any data they want about their patients. Across both participant groups, responses converged around two key points: (1) context: environmental or social factors surrounding a given smoking or eating event (e.g., location, presence/absence of others, presence/absence of screens); and (2) secondary activities: behaviors that co-occur with smoking or eating (e.g., reading, talking on the phone, drinking coffee). With reliable detection of eating or smoking, our device could trigger timely smartwatch or smartphone notifications to obtain context or secondary activities.

3.3 Advantages: Benefits of Passive Sensing

After explaining the premise of the HabitSense system to participant groups, we sought their input on the benefits of an RGB-T passive-sensing system in contrast to the tools currently used in their practice. In linking the system's design to the constraints of existing measurement methods, both groups agreed that the proposed system could enhance the precision of dietary and smoking data accessible to healthcare professionals.

In both participant groups, the automatic delivery of real-time behavioral data to clinicians was viewed as a strong benefit. Specifically, one clinician stated that *"In therapies ... we could check in during the day-to-day through*

the device to actually know about the improvement or if they are lagging behind." Another clinician expanded on this perspective, emphasizing that *"... the disparity of information we are trying to close can be achieved if we get live statements and summaries from this sensing device."* Beyond monitoring treatment efficacy, clinicians reported great interest in using data reported by the system to infer behavioral triggers from the context and/or secondary activities associated with eating or smoking episodes. An example provided by a tobacco treatment specialist illustrates this point: *"If the day after tomorrow this person took 15 cigarettes, and maybe today this person took 9, we can say what happened in that particular day.... Maybe this person was engaging in very tedious activity, and that craving was a little bit higher."* Another tobacco specialist observed that *"Using the video, you'd be able to see if the individual is in a social environment, or among friends... is it peer pressure? Or is it such kinds of things where the individual is being influenced in a social environment?"*

This indicates that these providers are not solely concerned with the observable behaviors but also with the underlying causes, which are more readily determined when rich contextual information is present before the event takes place. One dietitian stated, *"... the device would be really helpful because we would be able to get accurate data on what this individual usually does in their normal life, in their normal environment, so that we'll be able to understand specifically the activities which usually surround their behavior."* In the interest of identifying these patterns, both groups saw value in fine-grained behavioral measurements. Dietitians expressed interest in episode-level metrics, including meal duration, number of bites, and number of chews. Likewise, tobacco treatment specialists expressed interest in smoking topography metrics such as smoking episode duration, number of puffs, puff duration, inter-puff interval, and puff volume. Dietitians also expressed interest in food temperature as an indicator of food type and preparation methods.

3.4 Barriers: Challenges in System Implementation

We prompted participants to predict the challenges of implementing the HabitSense system with real patients. Both groups shared concerns regarding potential instances in which the system fails to detect a smoking/eating episode (false negative) or erroneously detects an episode when none occurred (false positive). Similarly, participants raised concerns over operational conditions the device may encounter in practical use, potentially impacting its effectiveness, such as water exposure, physical damage, or unusual environments. Participants speculated that issues related to privacy and social stigma might affect the willingness of patients to use the device, citing concerns over the collection of sensitive data and how the device's appearance might be perceived by individuals with whom the wearer interacts. Participants highlighted that dishonesty remains relevant because patients who aim to conceal their eating or smoking behaviors from healthcare providers could still do so if they know how the system operates. Lastly, dietitians reiterated that cost is a major factor patients and providers will consider when determining whether to adopt the device.

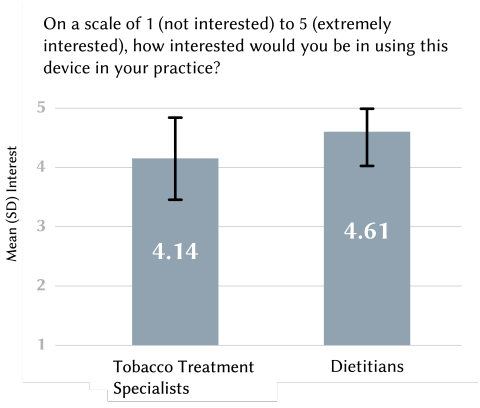


Fig. 2. Mean reported interest in using the HabitSense system by participant group.

3.5 Treatment Provider Interest in Using the HabitSense Wearable Camera

At the end of each session, all participants were asked to rate their interest in using the proposed HabitSense system in their practice on a scale of 1 to 5, with 1 indicating "not interested" and 5 indicating "extremely interested." A two-sample t-test was performed to compare interest in using HabitSense in clinical dietitians and treatment providers. There was no significant difference in interest, $t(34)=1.67$, $p=0.11$, despite 18 clinical dietitians reporting higher levels of interest ($M=4.61$, $SD=0.60$), compared with the 18 tobacco treatment specialists ($M=4.14$, $SD=0.69$). Certain participants specified that their rating was contingent on the HabitSense system "working as advertised" (i.e., performing at the levels of accuracy, generalizability, reliability, and efficiency stipulated in the design). These findings suggest a significant level of interest among both groups in wearable intake-monitoring systems and reflect their high expectations regarding the design and performance of these systems. The results of this focus group question are shown in Figure 2.

Although the difference was not significant, several factors may account for the slightly lower interest in the HabitSense device among tobacco treatment specialists. More dietitians reported using mobile applications and related technologies to support diet self-tracking than tobacco treatment specialists, which may contribute to dietitians having a slightly higher interest in adopting new technologies. Additionally, tobacco treatment specialists cited slightly greater concern over conditions interfering with the device's ability to capture contextual information since cigarettes are smoked in less favorable conditions (outdoors, dark, in motion) compared with eating (well lit, indoors, stationary plate). Lastly, a picture of a meal may provide slightly greater informational value than an image of a lit cigarette.

4 DESIGN IMPLICATIONS FROM EXPERIENTIAL INSIGHTS IN WEARABLE CAMERAS

Drawing on insights from dietitians and tobacco treatment specialists, we recognized the key design considerations outlined in the previous section as including privacy, RGB-T capabilities, sufficient resolution and field of view, affordability, minimal burden for users, and real-time detection capabilities during all-day wear. Refining our design in response to feedback and identified considerations, in addition to clinician input, we also uncovered hardware and firmware challenges through free-living studies with our evolving devices (see Supplementary Section 9.2 for Gen 1 and Gen2 study protocols). This experiential knowledge also informed the design and development of HabitSense. The subsequent section details our progressive evolution from Gen 1 to the culmination of HabitSense (see Figure 3 and Table 3).

4.1 Iterative Design

4.1.1 Gen 1. To overcome challenges encountered in neck-worn [53–55] and wrist-worn devices [56] designed to detect hand-to-mouth gestures related to eating and smoking activities, we set out to develop wearable cameras as a viable solution. We first developed Gen 1, which integrated a thermal sensor (GridEye 8x8 IR array) and RGB camera (OV2640) onto a printed circuit board (PCB). The addition of thermal imaging as a secondary sensing modality complemented the RGB data, particularly in detecting the wearer through the heat signature of pixels in the foreground. Privacy concerns were addressed by orienting the sensors toward the user's face and upper torso. Through Gen 1, we improved data collection while reducing user discomfort and privacy concerns.

Challenges and lessons learned: Although the device was successfully deployed in free-living settings (see Supplementary Section 9.2.1 for study protocol), the research team encountered three challenges during the study: (1) the device being warm to touch while charging, (2) time synchronization-related inaccuracies, and (3) limited field of view of the GridEye sensor. We addressed these challenges in the second iteration of our device.

4.1.2 Gen 2. To mitigate the problem of elevated device temperature during charging, we reduced the charging current of the onboard Li-Po charger in Gen 2, trading increased charging time for enhanced user comfort. To remedy time synchronization inaccuracies, we embedded a real-time clock (RTC) within the device powered by

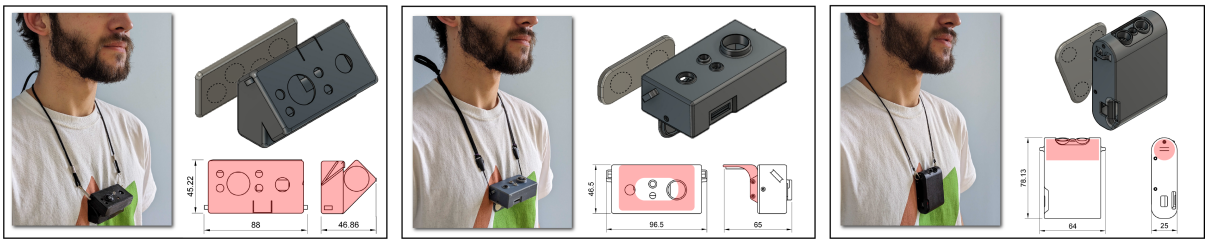


Fig. 3. Iterations through different generations (devices and magnetic back plates). Showcasing the transition from Gen1 to Gen2 to HabitSense, this figure reflects the design process of HabitSense, guided by expert insights and user feedback, emphasizing key considerations such as privacy, affordability, and real-time detection. Pink regions indicate parts used for orienting the sensor lenses with the activity of interest. All dimensions are in mm.

Table 3. Specifications of different generations of devices.

Device	Sensors	Weight (g)	Volume (cm^3)	Battery life (h)	Connectivity	Applications
Gen 1	Grid-EYE, OV2640	106	102.18	12	BLE	Eating Detection
Gen 2	MLX9640, OV2640	132	114.75	14	BLE	Eating Detection
Gen 3	LIS3DH, MLX9640, OV2640	87	93.75	17.6	BLE, WiFi	Eating/Smoking Detection and EE Estimation

BLE: Bluetooth Low Energy, EE: Energy Expenditure.

the onboard Li-Po battery. Concurrently, to prevent time loss due to battery depletion, we refined the device's firmware to include a sleep mode feature; this mode deactivates power-intensive functions when the battery falls below a set threshold, thus prioritizing RTC power supply and prolonging the device's operational longevity up to 20 days in sleep mode. To address the constraints posed by the limited 90° field of view and 8x8 pixel resolution of the GridEye sensor, we transitioned to the MLX90640 thermal camera, which offered a wider 110° field of view and a higher 32x24 pixel infrared array resolution.

Challenges and lessons learned: In the post-COVID landscape, the research and design team faced significant assembly challenges due to a marked shortage of STM chips and other key components. Despite an improved time-keeping mechanism extending battery life, time-drift issues led to unforeseen synchronization challenges, prompting repeated manual time corrections and reprogramming efforts. The device's non-modular architecture, based on a fixed PCB layout, posed challenges to incorporating new sensors or computational modules, such as machine learning accelerators, which would require a PCB redesign specific to the chosen accelerator. Feedback from participants again highlighted the need for a device that was smaller, lighter, and more ergonomically designed for comfort; that is, despite improvements from the first to the second generation, the device's form factor still did not achieve the level of unobtrusiveness necessary. Finally, an unfavorable center of gravity (see Fig. 3) and an inadequately designed magnetic pad failed to maintain the device's position consistently, leading to frequent displacements and consequently shifts in lens alignment that affected the integrity of the data collected.

4.2 Design Implications

Drawing on insights from focus group discussions and our iterative design experience in Gen 1 and 2, we delineate the guiding principles and objectives that informed the design process of the third-generation HabitSense device.

4.2.1 Prioritizing user privacy through on-device obfuscation algorithms. Due to burden and biased data, our thematic analysis and participant feedback from wearable camera studies indicate that health professionals and users favor camera usage over traditional self-report methods; however, they caution against the potential invasiveness and extensive field of view of cameras that may capture more than intended. Researchers have addressed privacy concerns in wearable camera technology by proposing two key privacy-enhancing approaches: first, orienting cameras toward the activity of interest to ensure focused capture [19] and second, applying obfuscation algorithms to protect the confidentiality and integrity of the data collected [23]. Both the Gen 1 and Gen 2 cameras were designed to orient toward the activity of interest; however, they lacked the capability to execute obfuscation algorithms on-device due to the PCB's configuration. Executing obfuscation algorithms offline exposes sensitive information to risks that could jeopardize user privacy and the integrity of the data. Section 5.5 details the design and execution of our proposed on-device obfuscation algorithm, which serves as an effective measure to protect user privacy by curtailing the exposure of sensitive data. Developing on-device obfuscation algorithms necessitates preserving the data's context to ensure that health professionals can accurately interpret user activities, thereby balancing user privacy with the practical value of the data for behavioral analysis.

4.2.2 Acquiring minimal data for maximal utility in activity detection. The principle of least privilege (PoLP) represents a foundational concept in secure design, advocating for minimal privilege allocation to programs and users to fulfill their roles [21, 57]. This principle guides our approach to data acquisition in wearable camera research. Continuous video recording in free-living studies generates vast, oftentimes irrelevant data, burdens storage media, and depletes device battery. Moreover, acquiring unneeded data, even if obfuscated, raises significant security concerns. We introduce the SECURE algorithm in Section 5.3 to navigate these issues. This algorithm strategically activates sensors only upon detecting a high likelihood of the targeted behaviors, thus limiting the data recorded to that which are pivotal for researchers and health professionals.

4.2.3 Modularity, reliability, reproducibility: pillars of system integrity and sustainability. The progression from our Gen 1 to Gen 2 devices highlighted a pivotal lesson: addressing device malfunctions due to field damage or component failures on the PCB required specialized technical proficiency for intricate repair work. Such procedures are not only time-intensive but also impose significant costs if entire PCBs are to be replaced. Modular designs not only facilitate expedited repairs and reduce costs but also bolster the device's dependability and consistency in production. The capacity for swift component testing and validation is crucial for the authenticity of data in free-living studies. Our system's modularity, supported by I2C communication and Qwiic connectors, also promotes expandability and seamless sensor augmentation, thereby extending its utility to monitor various health-risk behaviors. For instance, the platform could be reconfigured to assess UV exposure or outdoor time for melanoma survivors, creating a versatile tool for various health applications [58–61]. Furthermore, in light of the recent supply chain constraints on electronic components post-COVID, the capacity to interchange parts highlights the critical importance of a modular system design. This adaptability not only ensures uninterrupted research continuity but also contributes to the sustainability and recyclability of the device, as components can be replaced or upgraded without the need for discarding the entire unit, reducing electronic waste and fostering a more sustainable life cycle for research tools.

5 METHODS: DESIGN AND DEVELOPMENT OF THE HABITSENSE SYSTEM

This section explores the concrete application of our design implications in developing the novel HabitSense system. It includes the rationale behind component selection and the iterative enclosure design process, resulting in a comprehensive end-to-end system. Further, we detail our methodology for the free-living study and describe our proposed SECURE algorithm to address PoLP, as well as our proposed real-time on-device obfuscation algorithms to protect privacy. The section concludes with details about our proposed gesture-detection framework.

Table 4. HabitSense hardware design space. Components used in HabitSense are marked by *.

Modules	Alternatives	Features
Camera (RGB)	OV7670	VGA (640 x 480), 60mW/15fps(YUV)/maxRes
	OV2640*	2 megapixel (1600 x 1200), 140mW/15fps(JPEG)/maxRes
	OV5640	5 megapixel (2592 x 1944), 294mW/15fps(JEPG)/maxRes
Thermal Camera	Adafruit AMG8833	8x8 IR Array with 60° FOV, 0℃ to 80℃ range, 4.5 mA
	Grove MLX90641	16x12 IR Array with 110° FOV, -40℃ to 300℃ range, 12 mA
	Adafruit MLX90640*	24x32 IR Array with 110° FOV, -40℃ to 300℃ range, 23 mA
IMU	**Adafruit LIS3DH***	3-axis IMU, inexpensive, 2 uA
	Adafruit LSM6DSOX	6-axis IMU, built-in gesture recognition, 550 uA
	Adafruit BNO055	9-axis IMU, ARM Cortex-M0, 13.7 mA
RTC	**Adafruit PCF8523***	Dedicated battery, inexpensive, Moderately precise
	Adafruit DS3231	Dedicated battery, High-precision
	SparkFun RV-1805	Rechargeable supercapacitor, High-precision
Voltage Regulator	Adafruit LM3671	600 mA, 90-95% efficiency, 2 MHz frequency
	Adafruit TLV62569*	Inexpensive, 1.2 A, 90-95% efficiency, 1.5 MHz frequency
	Adafruit TPS62827	2 A, 90-95% efficiency, 2.2 MHz frequency

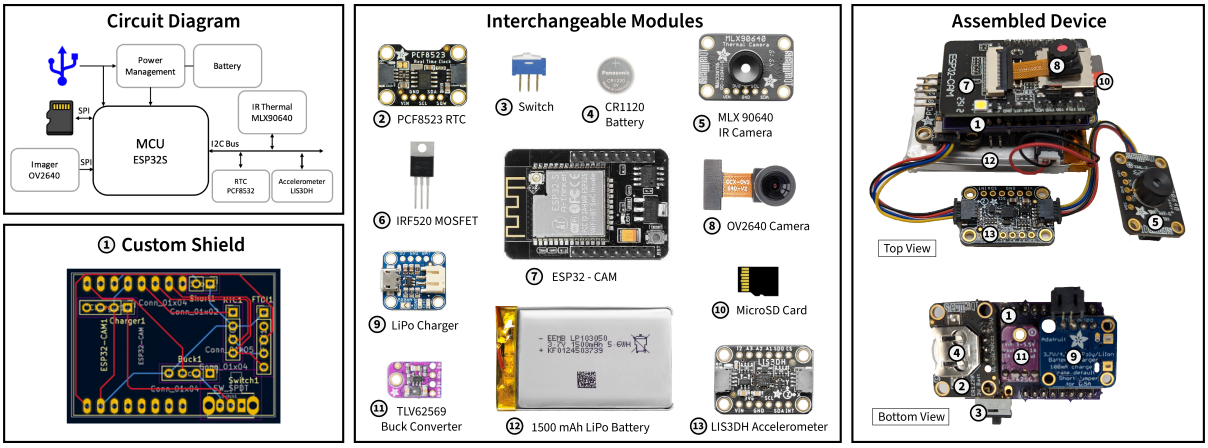


Fig. 4. Internal design schematic and electronic components used in HabitSense device.

5.1 Hardware Design Details

In our endeavor to develop HabitSense, a device capable of running obfuscation algorithms and real-time detection with a focus on reliability and reproducibility, we embraced a modular approach in its design. The development of Habitsense involved the following critical components: a microcontroller unit (MCU), custom shield, sensors, and power management modules and support components such as MOSFETs and external memory (see Figure 4). Our meticulous review of the hardware design landscape, as outlined in Table 4, guided our choice of the ESP32-cam development board to serve as the central platform for the system. This choice was motivated by its affordability ($5), compatibility with a 24-pin camera, built-in SD card port, WiFi and BLE capabilities (to trigger timely smartwatch or smartphone notifications to obtain further user context), a dual-core processor (240MHz)

with substantial RAM (520 KB SRAM and an external 4MB PSRAM) for on-device machine learning through TensorFlow Lite, and nine input/output ports that support various communication protocols (e.g., UART, I2C, SPI, ADC, and DAC). To seamlessly integrate additional components such as the IMU, thermal camera, and RTC, we engineered a custom shield facilitating easy connection to the ESP32-CAM board. This shield utilized the board's pins, channeling them to meet our data transfer, sensor integration, and power management needs. Addressing the focus groups' demand for high resolution and all-day battery life, we selected the OV2640 camera module with a 160-degree fisheye lens. This module not only offers a 1600x1200 resolution with a power draw of 140 mW, as detailed in Table 4, but is also equipped with an on-chip JPEG encoder to reduce the processing load on the MCU and RAM usage. The camera's standby mode draws less than 1 mA current. Should a future need arise for higher resolution without power constraints, we can effortlessly switch to the OV5640 camera module, which offers a 2592x1944 resolution but at a higher power consumption rate. To overcome the limitations of the 8x8 IR array in Gen 2, we selected the MLX90640 thermal camera, comprising a 24x32 IR sensor array and a wide field of view. For motion detection, we selected the LIS3DH triple-axis accelerometer for its low power consumption and proficiency in discerning device orientation and user engagement. This empowers the system to actively discern optimal periods for data collection or to execute complex real-time analyses. The accelerometer module can be upgraded to more advanced IMUs like the BNO055 for sophisticated motion detection tasks. Furthermore, in response to the drawbacks encountered with RTC integration in Gen 1 and 2, we opted for the PCF8523 RTC due to its standalone power source, guaranteeing independent time keeping from the system's main power supply while also allowing for potential upgrades to high-precision RTCs like the DS3231 for users necessitating superior time-keeping accuracy. To optimize the size, weight, and unobtrusiveness of HabitSense, we equipped the system with a 3.7 V, 150 0mAh LiPo battery. We ensured the provision of stable voltage and current to all components by integrating a high-efficiency (90-95%) voltage regulator (Adafruit TLV62569), which converts the input voltage to a consistent 3.3 V in conjunction with a micro-USB LiPo charger. While this configuration met our initial requirements, we designed the system with the flexibility to accommodate alternative voltage regulators, such as those with a broader input range (4.5-21 V), for adaptation to higher-voltage power sources as necessary. The charging circuit of our micro-USB LiPo charger, capable of delivering up to 500 mA, enabled a charge cycle of approximately three hours and included LED indicators to signify full charge status. In contrast to directly soldering the battery to the PCB in Gen 1 and 2, we adopted a JST connector for the battery in HabitSense, simplifying the battery replacement process and allowing for adjustments in battery capacity according to user needs.

The 3D-printed enclosure for the device underwent numerous iterations to refine its size, shape, mechanism, materials, and printing technology (FDM to SLA, SLS), as seen in Figure 16. These modifications ensured secure hardware containment, unobstructed camera views, easy micro-USB access, and compatibility with a neck lanyard attachment. To address issues of detachment and instability observed in earlier designs, a new triangular magnetic back plate was introduced (Figure 3). The final design also featured a hinge mechanism that orients the camera lenses toward the wearer's face, allowing for an activity-oriented, privacy-preserving design (see Section 9.3 and Figure 16 for details regarding the iterative case design).

5.2 Study Design and Data Collection

Utilizing the final iteration of our embedded RGB-T camera system, contained within the final iteration of its corresponding 3D-designed enclosure, we assembled multiple HabitSense devices and recruited 15 participants (7 smokers, 5 with obesity, 3 without obesity) from the Chicago area to participate in a free-living observational study. Recruitment was conducted through online advertisements, seeking adults that were over 18 years of age, fluent in English, who own a smartphone, with a body mass index (BMI) over 30 and under 30 kg/m^2 for people with obesity and without obesity, respectively; to obtain enough data, smokers needed to average at least ten

cigarettes daily. We screened and subsequently enrolled eligible participants, equipped them with a HabitSense wearable device, and provided instructions for its use. Participants installed SNaPsHOT (Smoking, Nutrition, and Personal Habit Observation Tool), an app developed at HABits Lab, and were instructed to log eating or smoking events by opening the app and tapping the appropriate "I'm Eating" or "I'm Smoking" button to record a timestamp for each activity. Over a week, participants wore the device during waking hours and used the app to track their behaviors before returning the device to the lab and filling out surveys about their experience. Notably, participants were not restricted to using the device exclusively during daylight hours or at night, with the only directive being not to wear the device while showering. This precaution was necessary for their safety, as the current version of the device is not waterproof.

5.3 SECURE: Sensor-Enabled Control for Ubiquitous Recording and Evaluation

We propose SECURE, a sensor-driven algorithm designed to optimize data collection for devices such as HabitSense, ensuring the acquisition of only essential data useful for researchers, dietitians, or smoking cessation counselors and the users themselves. The core idea of SECURE is to capture obfuscated RGB data that preserves privacy, initiating recording solely during instances with a high probability of a person engaging in eating or smoking activities. Unnecessary data acquisition results in (1) excess power consumption, (2) suboptimal utilization of computational resources, and (3) potential privacy concerns due to over-collection of data. To tackle these challenges, our algorithm employs a three-tiered hierarchical strategy tailored for real-time operations. We leveraged the microcontroller's deep sleep functionality, intermittently awakening it to assess the user's device wearing status using accelerometer data (see Figure 5). Subsequently, we analyzed a batch of thermal frames for hand-to-mouth gestures, activating the recording of obfuscated RGB frames upon gesture detection before setting the microcontroller back to its sleep state.

5.3.1 Pipeline.

Level 0. Our approach involves managing the clock frequency of the ESP32 MCU for power efficiency (see Figure 5). By default, the MCU operates at 240 MHz, but we set it to enter deep sleep mode when not in active use by using the ESP32's ultra-low power (ULP) co-processor, reducing the clock frequency to 150 KHz. After a predetermined time, the primary processor periodically awakens and operates at its minimum clock frequency of 80 MHz. To preserve temporary variables during sleep mode, we store them in the system's RTC memory, effectively conserving energy during device inactivity.

Level 1: Training a wearing-detection model. Upon reactivation, the primary processor performs accelerometer polling to determine the device's wear status, as shown in Figure 5. We detect wear over a 150-sample window (30 seconds) of three-axis accelerometer data by computing a set of ten running metrics: mean and variance of the x, y, and z-axis; mean, standard deviation, and variance of the L2-normalization of the three axes (removing gravity [energy]); and max change in energy differential. At the end of each 30-second segment, we compute a wear/not wear prediction using a decision tree. The decision tree was trained with a maximum depth of 10 to run on the device efficiently. The metrics were chosen based on individual thresholding optimization using f-beta scores. To stabilize and reduce false predictions, we only commit a change in wear-state when we observe five sequential stable predictions (2.5 minutes; e.g., five repeats of "wearing" or "not wearing"). The state of "not wearing" is rare since the device is usually switched off when doffed. Acceleration alone cannot detect all edge cases of "not wearing" precisely when the device is in motion but not in the correct positioning; therefore, so as to not miss data, the model was optimized to reduce false-negative predictions of wear time (0.60% false-negative rate), allowing other levels to filter out false positives (20.26% false-positive rate).

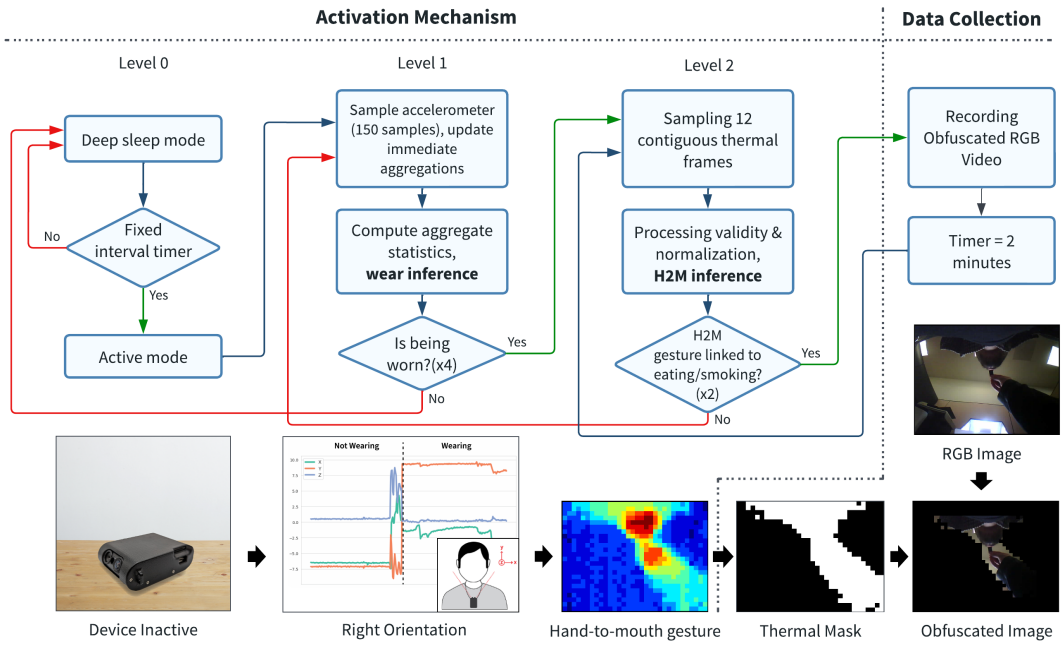


Fig. 5. Sensor-Enabled Activation Algorithm. Outlines the process from power-on to data recording, starting with deep sleep mode, followed by accelerometer checks for the device being worn (level 1), thermal frame sampling for hand-to-mouth gesture detection (level 2), and concluding with obfuscated RGB data recording, designed to record only essential data to optimize battery life and memory use.

Level 2: Should a person be detected wearing the device at level 1, we activate our thermal camera through a MOSFET and initiate sampling thermal frames. We then use a window of 12 contiguous thermal frames to estimate the presence of potential hand-to-mouth gestures associated with eating or smoking gestures. To identify these hand-to-mouth gestures, we developed a lightweight machine learning model capable of real-time detection on-device. We trained a two-layer neural network with dense layers on a sequence of contiguous thermal frames to incorporate the temporal aspect of thermal patterns associated with hand-to-mouth gestures. This is supported by the distinct temperature signatures observed during eating or smoking gestures (see Figure 6). A potential challenge when working with thermal frames lies in the fluctuations of thermal signatures caused by changes in the surrounding environment, which we addressed by applying min-max normalization to the entire window of thermal frames before their utilization in model training.

Data collection level: When a hand-to-mouth gesture is detected, the system starts capturing obfuscated RGB video (obfuscation algorithm described in the following section) for a user-defined period of time. The on-device obfuscation mechanism provides privacy-preserving data collection while providing visual confirmation (ground truth) of the activities of interest.

5.3.2 Energy profiling. To determine the achievable hours of daily usage with our SECURE algorithm, we calculated the device's average current usage, considering its dynamic operation across level 0, level 1, and level 2 and the data-collection level. We estimated the average daily wear time (t^w) and non-wear time (t^{nw}) based on our wearing detection algorithm, as well as the average obfuscated recording time (t^{obr}). We simulated 'Z'

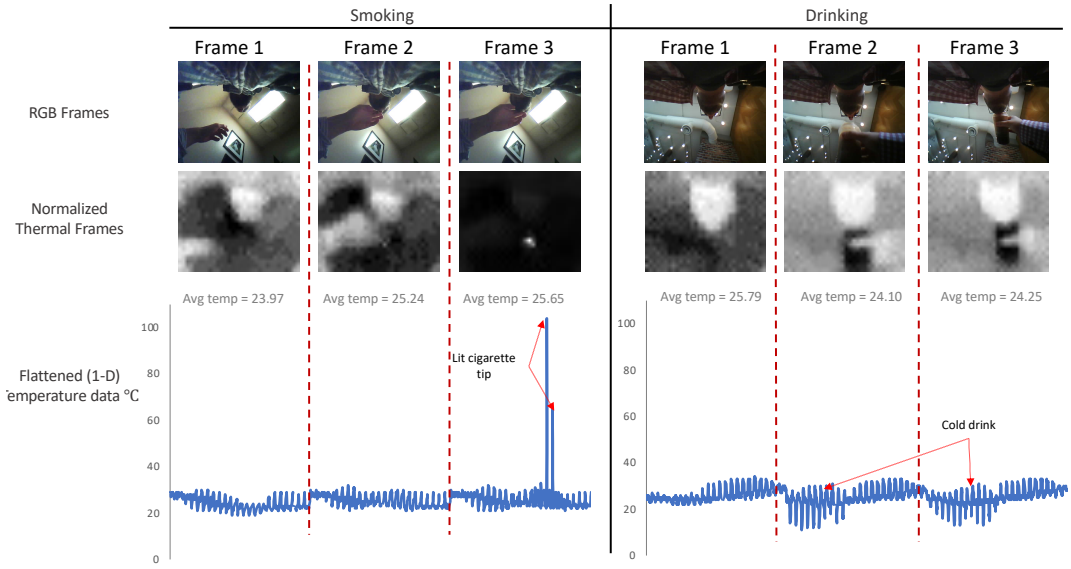


Fig. 6. HabitSense gesture characterization. Illustrates smoking and drinking gestures in RGB and thermal data, with RGB frames confirming gestures, normalized thermal frames showing temperature distribution, and flattened thermal histograms highlighting the heat signatures during eating and smoking.

minutes of data recording following a detected hand-to-mouth gesture. Denoting the current consumption of each level as i_{l1} for level 1, i_{l2} for level 2, and i_{dc} for data collection, we calculated the average daily current consumption:

$$Average\ current = \frac{(i_{l1} \cdot \sum t_j^{nw}) + i_{l2} \cdot (\sum t_j^{w} - \sum t_j^{obr}) + (i_{dc} \cdot \sum t_j^{obr})}{\sum t_j^{nw} + \sum t_j^{w}}$$

Careful consideration was given to the distinction between the system's operational states. Specifically, when the system operated in level 2, the device was being worn but not actively recording; this contrasts with periods where obfuscated data recording was in progress, necessitating a higher current draw as seen in Figure 7. Therefore, in the equation presented, we ensured the accuracy of our average current estimation by subtracting the obfuscated recording time (t^{obr}) from the total wear time (t^{w}) to reflect the differentiated current levels during device wear (level 2).

5.4 On-Device Obfuscation Algorithm

On-device obfuscation strategically masks background details to maintain focus on foreground activities, particularly hand-to-mouth gestures relevant to eating and smoking detection. We utilize the thermal sensor's capabilities, leveraging the proximity of the wearer's head and hands to the camera, which manifests as a distinct heat signature. This thermal information is overlayed onto the RGB data to create a human thermal mask that isolates and preserves the wearer's head and hand imagery while rendering the surrounding region opaque.

Our on-device obfuscation algorithm utilizes an adaptive masking operation to overlay thermal camera data onto RGB frames while accommodating disparities in resolution (640x480 for the RGB sensor and 32x24 for the thermal sensor), as well as differences in field of view. The superimposition process requires a one-time initial

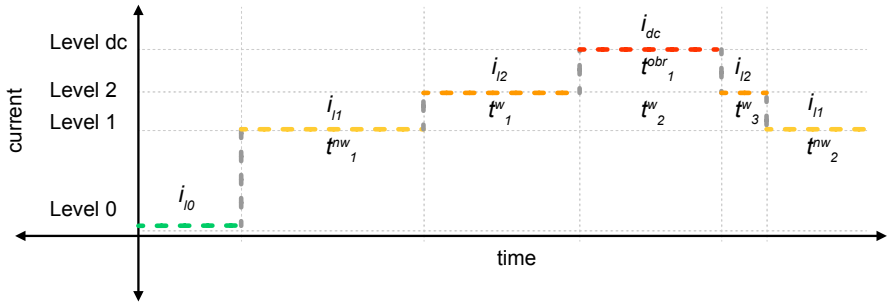


Fig. 7. SECURE pipeline energy profiling current consumption during different levels.

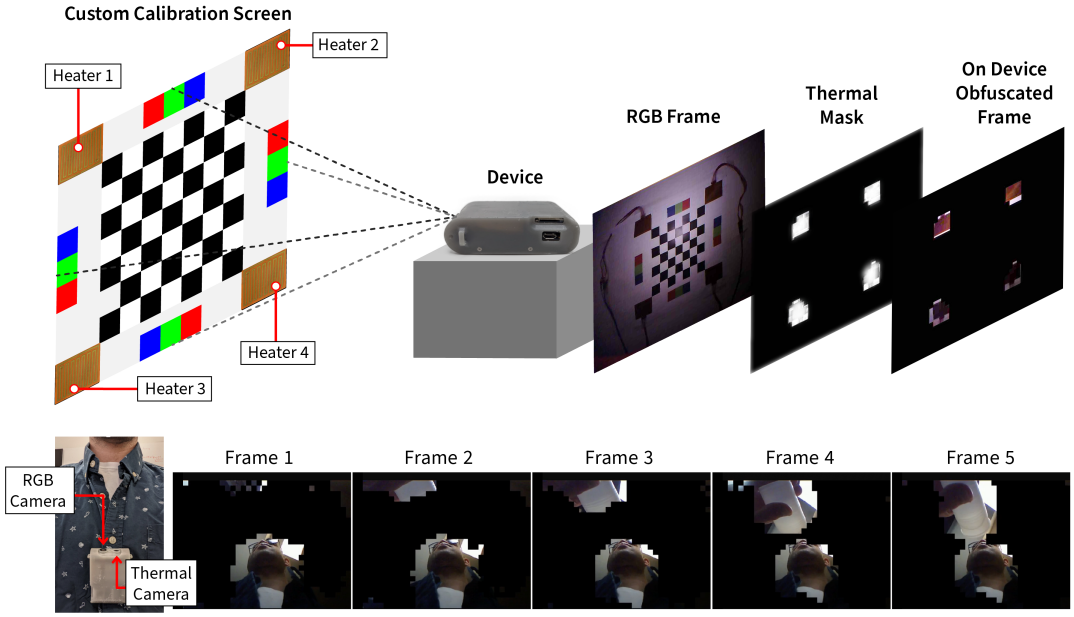


Fig. 8. Sensor Calibration and RGB Obfuscation Process. Depicts calibration via a screen with heated and colored sections for sensor alignment and shows how calibrated thermal data enables obfuscation in RGB sequences, keeping heads, hands, and objects in hands clear while obfuscating the background.

calibration to synchronize the two data modalities. Following this calibration, the derived parameters apply to all HabitSense devices, attributable to our 3D design that maintains a fixed positional relationship between the RGB and thermal sensors. For calibration, we built a custom-designed calibration screen with four heaters at the corners and a checkerboard pattern in the center, as shown in Figure 8. We positioned the HabitSense device in front of this screen and recorded frames while actively running a calibration algorithm on the device. This algorithm superimposed thermal masks onto the RGB frames while iterating through varying positional and scaling parameters in real time. Upon manual inspection of the recorded data, we identified and utilized the parameters that yielded the most accurate superimposition of the thermal mask on the RGB frames for

Algorithm 1 On device thermal mask-based obfuscation

function MaskCalibration($buffer$, $height$, $width$, $mlxFrame$, $threshold$)
  $offsetLeft, offsetRight, offsetTop, offsetBottom \leftarrow calibratedOffsetValues$
  $scaleX \leftarrow (width - (offsetLeft + offsetRight))/32.0$ ▷ Calculate X scaling factor
  $scaleY \leftarrow (height - (offsetTop + offsetBottom))/24.0$ ▷ Calculate Y scaling factor
  for $y \leftarrow 0$ **to** $height - 1$ **do** ▷ Iterate over height
    for $x \leftarrow 0$ **to** $width - 1$ **do** ▷ Iterate over width
      $bufferIndex \leftarrow (y \cdot width + x) \cdot 2$ ▷ Calculate buffer index
      $isEdge \leftarrow y \leq offsetTop$ **or** $y \geq height - offsetBottom$
      $isEdge \leftarrow isEdge$ **or** $x \leq offsetLeft$ **or** $x \geq width - offsetRight$
      if $isEdge$ **then** ▷ If it's the offset, blackout pixel
        $buffer[bufferIndex], buffer[bufferIndex + 1] \leftarrow 0$ ▷ Zero out offset edge
      else
        if $mlxFrame[mlxBufferIndex] < threshold$ **then** ▷ Check threshold
          $buffer[bufferIndex], buffer[bufferIndex + 1] \leftarrow 0$ ▷ Apply threshold masking
        end if
      end if
    end for
  end for
  return $buffer$ ▷ Return the masked buffer
end function

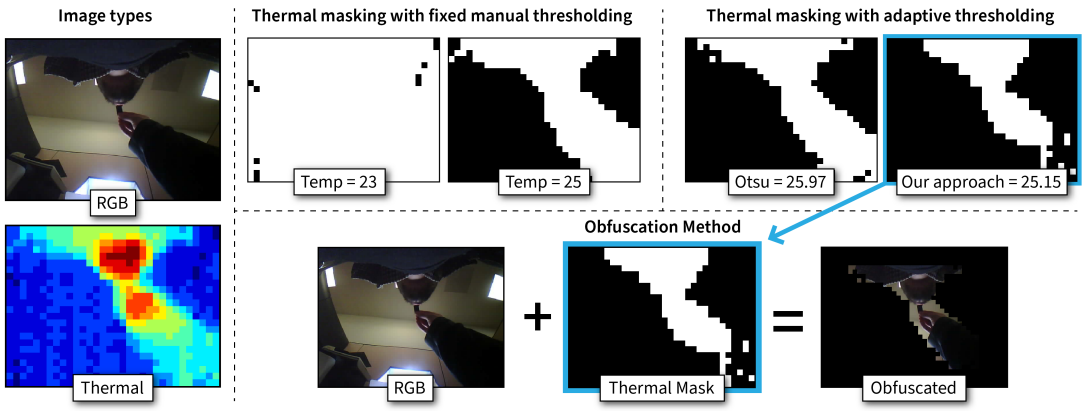


Fig. 9. Obfuscation threshold determination.

our on-device obfuscation algorithm. Our algorithm [ALG 1] initializes by constructing a margin ribbon for the RGB image, zeroing pixels within this margin, using the positioning and scaling parameters to account for the smaller field of view of the thermal sensor (compared with the RGB). We rescaled the thermal frame to match the dimensions of the active area of the RGB image, excluding the margin. Subsequently, the algorithm evaluates the thermal value of each pixel against a predetermined threshold, masking those in the visual buffer that fall below this criterion. Our experiments, illustrated in Figure 9, explored several strategies for determining this threshold. Initially, we considered fixed temperatures based on the human temperature range, but these

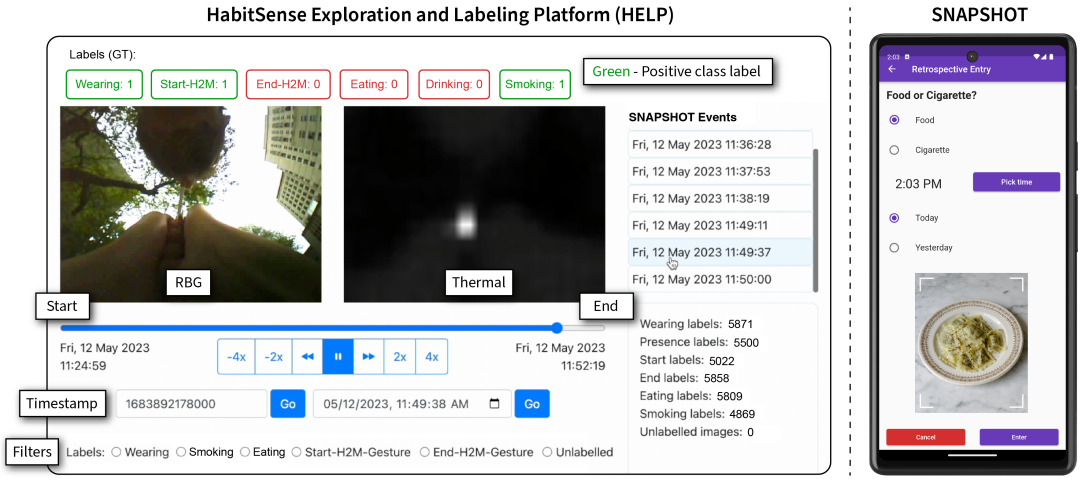


Fig. 10. HabitSense Exploration and Labeling Platform (HELP) web application (left) and the Smoking, Nutrition, and Personal Observation Tool (SNAPSHOT) mobile application (right).

proved suboptimal due to the influence of the device's distance from the subject's face and ambient temperature variations. Consequently, we adopted an adaptive approach. Although the Otsu method (used in prior work [22]) for image binarization was tested, we found that computing the average temperature of the pixels in each thermal frame yielded results comparable to those of Otsu, with the added benefit of reduced computational complexity for on-device execution. Finally, the algorithm generates a composite image that accentuates thermally notable regions, focusing on pertinent activities such as the clear depiction of the wearer's head, hands, and held objects—food, a glass of water, or a cigarette—while the surrounding area is obfuscated to safeguard the privacy of both the wearer and any non-consenting bystanders.

5.5 Gesture Recognition Framework

Having established an efficient data acquisition strategy, we aim to investigate the utility of HabitSense data in identifying instances of eating and smoking behaviors. To ascertain the effectiveness of our collected data in identifying eating and smoking behaviors, we annotated the raw data acquired during the free-living study, thereby enabling the training of machine learning models for behavior detection.

5.5.1 Data annotation: HELP and SNAPSHOT. Data annotation represents a significant commitment of time and effort in the data processing workflow. We explored existing annotation tools, such as labelImg and label studio [62], to annotate the start and end of gestures while simultaneously displaying multimodal data (RGB/thermal/IMU) during the annotation process. However, we found that these platforms are predominantly optimized for object-based annotation and do not provide, by default, the functionality for annotating gesture-specific segments or visualizing multimodal data in a manner that aligns with our project requirements. To address these challenges, we designed and developed the HabitSense Exploration and Labeling Platform (HELP) [Fig. 10], a dedicated system for streamlined data annotation of raw data acquired from video-based behavior monitoring tools.

During our free-living study, we deployed SNAPSHOT, a multi-platform mobile app developed in Flutter, allowing participants to document their eating and smoking activities while wearing the Habitsense device. We engineered the HELP platform such that annotators could reference events logged via the SNAPSHOT application to more efficiently find eating and smoking activities, given the prevalence of inactivity in the dataset.

We systematically annotated the start and end of hand-to-mouth gestures associated with eating and smoking activities, in addition to distinguishing periods of device wear from non-wear. The onset of a gesture was defined as the hand approaching the mouth for food intake or puffing from a cigarette, and its termination was marked by the hand retracting from the mouth.

5.5.2 Gesture recognition using state-of-the-art spatiotemporal architecture: Timesformer. In the context of consumption behaviors, the primary gesture preceding eating or smoking is a hand-to-mouth gesture. While 3D convolutional networks have successfully identified spatiotemporal gestures, the emergence of state-of-the-art models, notably transformers [63] built upon attention mechanisms, introduces more efficient approaches for training video data. TimeSformer [64] introduces a significant shift in video analysis using self-attention mechanisms instead of traditional convolutional methods. This change leads to better classification accuracy, quicker training times, and improved efficiency in testing, allowing the model to process long video sequences with spatial and temporal context effectively. Utilizing this architecture, we could also generate self-attention maps, which illuminate the model's focal areas, thereby rendering the model not only effective but also interpretable [65, 66].

We trained binary (eating vs. none, smoking vs. none, hand-to-mouth vs. none) and multi-class TimeSformer (eating vs. smoking vs. none) classifiers using annotated RGB and thermal data to detect eating and smoking activities. For each of the four classifiers, we train and test models on (1) RGB-only data; (2) Thermal-only sensor data; (3) RGB Obfuscated-blur (lower-privacy version that blurs the background, gaussian blur with radius 15); and (4) RGB Obfuscated-mask (a high-privacy version masking all background pixels). Through surveys, these two obfuscation methods effectively reduced privacy while maintaining contextual information surrounding hand-to-mouth activity [23]. To inform the sensors and data type used in the online model (Thermal-only, Obfuscated-blur, and Obfuscated-mask), we perform significance tests using an analysis of variance (ANOVA), followed by post hoc pairwise t-tests. Our data were segmented into video clips, each defined by the commencement and conclusion of the specific gesture pertinent to the model—eating, smoking, or hand-to-mouth. For the "none" class clips, a distribution was established mirroring the gesture lengths of the respective positive classes. Using this distribution as a benchmark, we randomly selected sequences of frames for the "none" class, ensuring they adhered to the established length distribution of the positive classes. These clips were then resized to a short edge of 256 pixels following the procedures specified in [67]. We maintained a balanced dataset with an equal representation of 50% for both positive and negative classes. For validation, we implemented a leave-one-participant-out (LOPO) cross-validation strategy, with an 80:20 split for training and validation.

5.5.3 Gesture recognition using quantized neural networks on-device. Training state-of-the-art offline models demonstrated the potential of the data collected by HabitSense, but for real-world applications, models that can detect hand-to-mouth gestures associated with eating and smoking in real time are essential. Based on the winning data type from the significance test, we train and test the data type that provides the greatest privacy preservation without impacting the accuracy of activity detection. For the real-time detection capability and to capture the temporal nature of the gestures, we selected a window size of 12 frames. This selection was grounded by the distribution of the length of eating and smoking gestures (Supplementary Figure 21). We also mirrored the approach used in the offline models to select the negative class, ensuring class balance during training. We used a training-to-validation ratio of 80:20 and performed leave-one-out cross-validation to evaluate the models. The on-device models, based on a two-layer neural network, were quantized to enhance the real-time detection performance. The inference time was estimated by calculating the number of multiply-accumulate operations (MACs) from the model weights/parameters. Then, by knowing our MCU's number of operations per second, dictated by its clock frequency, we estimate the inference time.

6 RESULTS

In this section, we present (1) the statistical data from our study; (2) an evaluation of our device conducted with 15 participants (comprising 8 people with and without obesity and 7 smokers), highlighting the advantages of capturing smoking and eating episodes through obfuscated RGB video, triggered by the SECURE algorithm; (3) the evaluation of machine learning algorithms for both the offline and real-time detection of eating and smoking behaviors; and (4) the positive feedback received for HabitSense following its third iteration.

6.1 Data Statistics

In total, we collected 46,060 minutes (768 hours) of data with a sampling rate of 5 Hz. The recording duration for each participant ranged from 1,644 to 4,901 minutes (27.4-81.7 hours). Table 5 presents the demographic characteristics of the participants. From these recordings, we captured 10,495 hand-to-mouth gestures across all participants (see Table 10). On average, each participant had 807 hand-to-mouth gestures (min=35, max=1726, std=531). Each hand-to-mouth gesture is represented by a start frame and an end frame. By analyzing the frame lengths of all the hand-to-mouth gesture clips captured across all participants, we found that, on average, each hand-to-mouth gesture clip consisted of 12.2 frames (min=9.39, max=20.10, std=2.98). Supplementary Figure 21 provides a visual representation of the distribution of hand-to-mouth gesture clip counts per frame length. The combined duration of all the hand-to-mouth intakes amounted to 420.91 minutes (7.02 hours) of smoking, eating, or drinking gestures. When examining individual participants, the time spent smoking, eating, and drinking varied from 1.7 to 97.2 minutes (min=1.7, max=97.2, std=25.32). Two participants were omitted from our data and analysis because of device malfunctions, including water damage, during the free-living study.

Table 5. Demographic information of participants.

Participant	Age	Sex	Race	Hispanic/Latino	Smoker	Hours recorded	Ave. hrs per Day
P1	36	male	white	no	no	27.4	5.5
P2	22	male	white	yes	no	60.1	7.5
P3	26	female	black	no	no	59.3	4.2
P4	28	female	white	yes	no	62.9	9.0
P5	47	female	black	no	no	61.9	6.9
P6	40	female	white	no	yes	62.8	7.0
P7	43	female	black	no	yes	69.1	9.9
P8	27	male	black	no	yes	81.7	7.4
P9	27	female	white	no	yes	47.0	5.9
P10	69	male	white	no	yes	74.3	9.3
P11	31	male	asian	no	yes	36.5	7.3
P12	30	female	white	no	yes	70.5	8.8
P13	25	female	white	no	no	54.4	5.4
Mean						59.1	7.2
Total						768.0	

6.2 SECURE Algorithm Evaluation

6.2.1 Assessment of volume of data captured through the SECURE algorithm. Using our pipeline, we evaluated the device wear-detection and hand-to-mouth gesture-detection models and, by assessing different parameter settings, determined the minimum amount of data required to capture the maximum number of hand-to-mouth gestures associated with eating and smoking (see Figure 11).

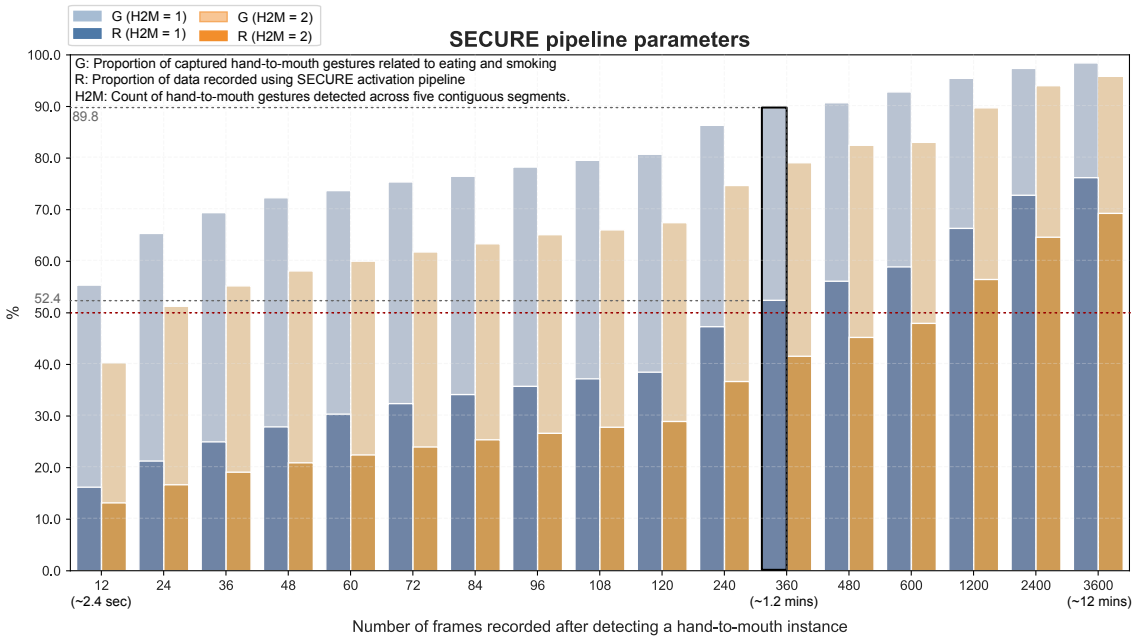


Fig. 11. SECURE pipeline parameter determination. This figure shows the effects of varying parameter settings on data collection. The graph displays the effectiveness of choosing a threshold of one hand-to-mouth gesture across five segments and a recording duration of 360 frames (1.2 minutes), achieving 89.8% capture of eating and smoking gestures with only 52.4% data recorded.

Wearing detection (Level 1). A decision tree was trained with an 80/20 train/test split on 150-second aggregated samples, resulting in an F1-score of 0.86 on "no wearing" predictions and 0.98 on "wearing" predictions. On the ESP32 microcontroller, the time to compute intermediate values for aggregate metrics at each pull of the accelerometer was negligible (<1 μs), whereas the time to compute metrics for each window was measured at 24 μs and the decision tree inference time at 5 μs. Overall, this time was insignificant compared to other steps in the pipeline.

Hand-to-mouth gesture detection (Level 2). Our two-layer network-based hand-to-mouth gesture detection model achieved an F1-score of 0.71. As detailed in Table 6, the model occupied 134 KB of memory and necessitated 48 ms for on-device inference time. Additionally, due to the model's training on a 12-frame window, frame normalization on the device required an extra 24 ms, resulting in a total inference time of 72 ms.

Table 6. Performance of each SECURE pipeline level of the system based on the leave-one-participant-out evaluation.

Level	Precision	Recall	Accuracy	F1-score	Model Size	Inference Time
Wearing detection (level 1)	0.97	0.99	0.97	0.98	15.4 KB	<50 μs
Hand-to-mouth detection (level 2)	0.79	0.67	0.74	0.71	134 KB	48 + 24 ms

Table 7. Percent of data recorded versus percent of captured gestures and resulting storage and energy savings using the SECURE pipeline. Utilizing the SECURE pipeline, we captured approximately 89.8% of hand-to-mouth gestures while recording only 52.4% of the data, thereby conserving about 47.6% of storage space and achieving 30.0% energy savings.

Participant	Recorded (%)	Hand-to-Mouth Captured (%)	Storage Saving (%)	Energy Saving (%)
P1	26.35	77.37	73.65	47.12
P2	72.25	96.43	27.75	16.54
P3	60.52	90.27	39.48	20.38
P4	60.82	93.96	39.18	22.25
P5	50.28	82.90	49.72	30.92
P6	55.15	86.66	44.85	23.81
P7	49.61	94.48	50.39	27.59
P8	41.89	93.64	58.11	43.57
P9	44.50	72.99	55.50	33.09
P10	71.94	96.63	28.06	16.63
P11	33.34	96.76	66.66	60.53
P12	64.38	96.85	35.62	18.51
P13	50.36	88.23	49.64	29.33
Mean	**52.41**	**89.78**	**47.59**	**30.02**

6.2.2 Energy profile. We evaluated the energy efficiency of our SECURE pipeline by measuring the HabitSense system's current consumption through the INA219 current sensor at each SECURE pipeline level, as depicted in Table 8. The SECURE pipeline's design inherently minimizes power usage. It maintains the system in its lowest power state—level 0—when idle, using only the ULP co-processor and drawing approximately 2.8 mA. Based on a 1500 mAh lithium polymer battery and assuming a constant current draw, this translates to a standby time of approximately 2.5 weeks per charge. During active periods, the system periodically polls the accelerometer to determine the device wear status, consuming 57.44 mA. If worn, the thermal camera is activated via a MOSFET to scan for hand-to-mouth gestures, increasing consumption to 79.44 mA. Should no gestures be detected, the system would operate for approximately 18 hours; however, upon detecting a gesture, it would begin recording obfuscated data using the RGB camera and storing data on the SD card, necessitating a current draw of 139.60 mA. Using the SECURE pipeline, we captured approximately 89.8% of hand-to-mouth gestures while recording only 52.4% of the data, thereby conserving about 47.6% of storage space and achieving 30.0% energy savings (Table 7).

Table 8. Energy requirements for each level of the SECURE pipeline.

Level	Active Components	Power (mW)	Current (mA)
Level 0	ULP co-processor	10.36	2.80
Level 1	ESP32 (80 MHz), accelerometer	212.52	57.44
Level 2	ESP32, accelerometer, thermal camera	295.03	79.74
Data collection	ESP32, accelerometer, microSD thermal and RGB cameras	471.42	139.60

6.3 Evaluation: Usability of Data Collected

In this section, we evaluate the overall performance of the method by assessing the F1-score for all eating and smoking activities of all the participants.

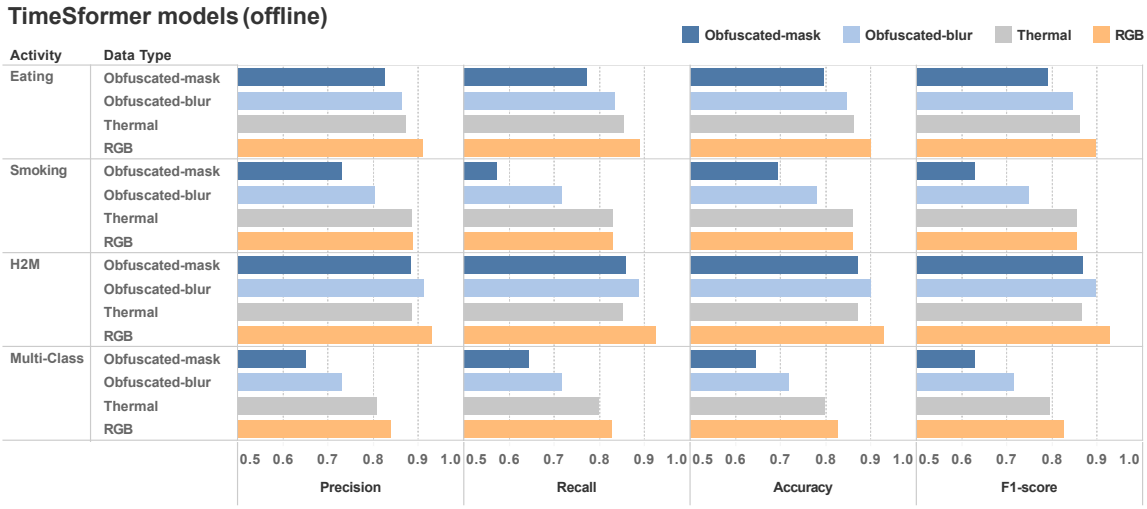


Fig. 12. Performance comparison of TimeSformer models (offline). H2M, hand-to-mouth.

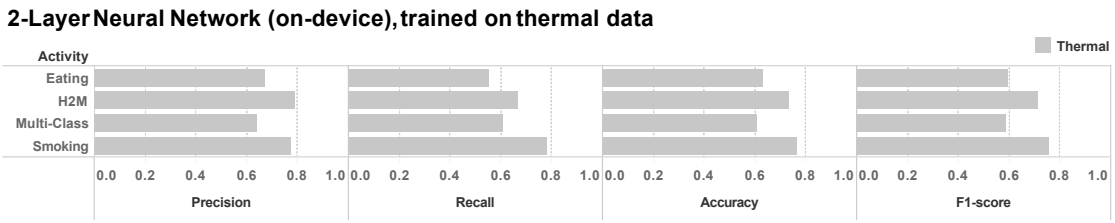


Fig. 13. Performance comparison of 2-Layer Neural Network, on-device models. H2M, hand-to-mouth.

6.3.1 Eating and smoking recognition offline analysis. We trained a series of offline models for each activity and input data, reporting the F1 score; Figure 12 highlights these results. The TimeSformer architecture-based classification models excelled in hand-to-mouth gesture recognition with an F1-score of 92.7%, closely followed by the obfuscated-blur model and eating detection using RGB data, each achieving 89.7%. However, for smoking detection, the results of RGB versus thermal were much closer, with thermal trailing closely behind RGB video by 85.4%. In addition to the models trained on individual activities, we explored multiclass models, which yielded less effective results, with an F1 score of 82.6% in RGB, thermal performance slightly below RGB at 80%, and obfuscated-blur around 71%. We observed a trend where the RGB-trained model generally outperformed models using only thermal data; however, unaltered RGB images pose privacy concerns and are computationally problematic on embedded systems. As a result, we aimed to process either the thermal, obfuscated-blur, or obfuscated-mask on-device.

6.3.2 Assesing the potential of on-device detection for real-time application. A one-way ANOVA revealed a statistically significant difference in F1-score between at least two groups ($p<.05$) for all eating, smoking, hand-to-mouth, and multi-class activities. Post hoc pairwise analyses indicate no significant difference between thermal and obfuscated-blur ($p<.05$ for all activity types; eating, smoking, hand-to-mouth, and multi-class). However, post hoc pairwise analyses indicated a significant difference between obfuscated-blur and obufscated-mask ($p>.05$ for all activity types; eating, smoking, hand-to-mouth, and multi-class). In light of these findings, we considered

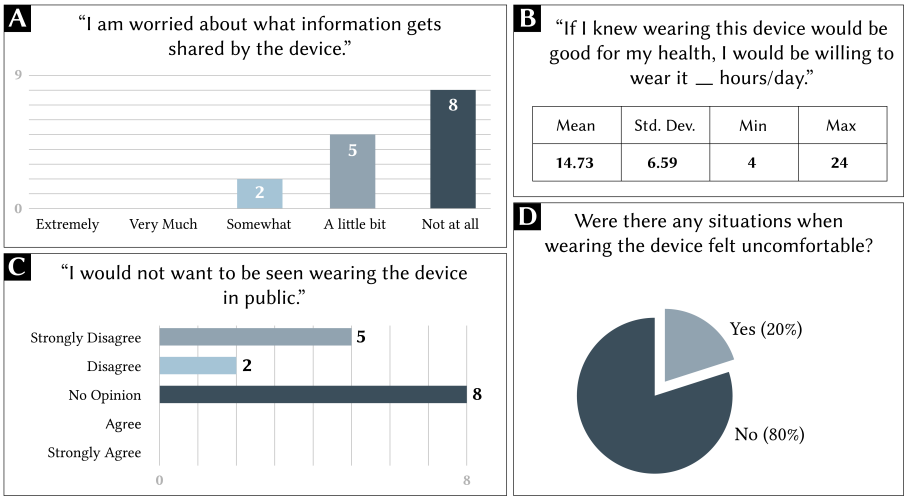


Fig. 14. Participant responses to burden metrics concerning (A) privacy, (B) obtrusiveness, (C) willingness-to-wear, and (D) physical comfort.

thermal to be the optimal target for the input of the on-device model due to the drastic reduction in input data (pixel count) and given no significance difference between thermal and obufscated-blur.

Using the internal input data, the on-device models were quantized two-layer fully connected neural networks (reduced size and inference time). These models were expected to have reduced performance due to these optimizations. The best-performing model, individual smoking detection, yielded an F1-score of 75.7%, whereas hand-to-mouth gesture recognition yielded an F1-score of 71.1%. Although the precision and recall of the smoking model were closely matched, with an approximately 1% difference, the hand-to-mouth gesture model displayed a much higher precision than recall.The model for eating detection produced an F1-score of only 59.5%, which warrants future investigation. As with the offline model, the multi-class model performed below the individual models with an F1-score of 58.8%.

Table 9. On-device smoking detection model (thermal data), participant-level analysis.

	Precision	Recall	F1-score	Accuracy
P6	0.61	0.65	0.63	0.62
P7	0.79	0.94	0.86	0.84
P8	0.75	1.00	0.86	0.83
P9	0.89	0.39	0.54	0.67
P10	0.84	0.79	0.81	0.82
P11	0.80	0.82	0.81	0.81
P12	0.70	0.88	0.78	0.76
mean	0.77	0.78	0.76	0.77

7 DISCUSSION AND LIMITATIONS

7.1 User Burden and Privacy Evaluation

In the study of participants' perceptions regarding the device, several key insights emerged (Figure 14). When inquiring about participant concerns regarding the information shared by the device (panel A), 87% (13/15) of participants had limited to negligible worries, implying minimal privacy concerns related to the data. For comparison, 77.7% (14/18) of participants who completed the Gen 2 study and 77.6% (38/49) participants who completed the Gen 1 study reported being "a little bit" or "not at all" concerned about what information gets shared by the device. When asked how long participants would be willing to wear the device for its health-related benefits, the average intended wear time was approximately 14.73 hours daily (panel B). It is worth noting, however, that a standard deviation of 6.59 hours was observed, and the wear time spanned a range from a minimum of 4 hours to a maximum of 24 hours daily. The actual mean daily wear time of HabitSense study participants was 7.2 hours, compared with 4.3 hours observed in the Gen 2 study. Together, these results indicate a 3 hours-per-day improvement in wear time between Gen 2 and HabitSense. Within the framework of social acceptability, approximately 53.33% of the participants assumed a neutral stance, indicating no distinct preference concerning their visibility while wearing the device in public (panel C). Meanwhile, about 47% of the participants were comfortable with being seen wearing the device in public spaces. One participant commented on the social comfort of the system, saying, "I did not really notice it until I saw people looking at me. They didn't ask, but they 'asked with their eyes.' It didn't bother me, but I was maybe a little conscious of it." Another participant said, "Nobody was even looking at it. They probably thought it was a radio or something." These responses and quotations suggest that the HabitSense wearable constitutes a meaningful step forward in the mitigation of social burden for wearable camera systems. Lastly, 80% of participants found the device comfortable in all situations they encountered, whereas only 20% reported instances of discomfort while wearing the device (panel D).

7.2 Privacy Algorithm: SECURE

At level 1 of the SECURE pipeline, acceleration data were used to determine whether the device was worn. However, this method faced challenges in detecting edge cases where the device, while active, was oriented similarly to when worn (while seated or stationary) but placed on the hook of a door hanger. However, the model was optimized to minimize false negatives (0.6%), i.e., instances where wearing was falsely detected as non-wearing, as this is more critical to the system's functionality. Future exploration could involve integrating gyroscopic data to enhance the detection of the orientation of the device. This addition, coupled with a revised case design that limits similar wear orientations when the device is not worn, may improve performance. Another potential avenue is to leverage the thermal camera in wear detection, provided that it is aligned with the system's energy budget constraints.

An important observation from Table 7 is the suboptimal performance of the SECURE system on participant P9, who was a smoker. Further investigation into the data, as seen in Table 9, showed that the smoking gesture of P9 often involved covering the cigarette tip with their hand, which likely contributed to these detection errors. Moreover, most of the participant's eating and smoking activities occurred inside a car, an environment characterized by higher indoor temperatures, and less contrast in heat signature. Future work should address the performance of eating and smoking in such scenarios. The final step of the SECURE pipeline involves recording obfuscated data using thermal information to mask the wearer and obfuscate background pixels. We can achieve a frame rate of about one frame per second with an image resolution of 640x480. In comparison, on-device obfuscation techniques have been implemented in various camera systems such as privacyCAM [49], which encrypts the facial region; trustCAM [50], which applies a Canny edge filter to the face; and trustEye [51], which renders the entire frame with a cartoon effect. These systems can operate at higher frame rates, with privacyCAM and trustEye testing at lower resolutions of 320x240. TrustCAM has achieved on-device obfuscation at 640x480

resolution and a higher frame rate of 5 frames per second. However, these systems are not designed as wearable technologies with all-day battery life, nor do they utilize thermal data for masking. Future work should focus on enhancing the frame rate at higher resolutions.

7.3 Comparative Analysis of Model Performances for Gesture Classification

In our study, we trained the on-device models using thermal data to achieve feasible inference times and to comply with the device's resource limitations while enhancing privacy for real-time applications. Our benchmarks using TimeSformer models with thermal data demonstrate performance comparable to obfuscated blur models across all scenarios. However, the on-device models demonstrated superior performance in detecting smoking gestures, with an F1-score of 75.7%, compared with eating gestures, which had an F1-score of 60%. This disparity likely stems from the distinct thermal signatures of smoking gestures, which are more discernible in thermal imagery.

Previous studies, including work by Bi et al. [35], have investigated eating detection using RGB data from head-mounted cameras, achieving F1-scores of 73.8% and 78.7% with 3D CNNs and slow-fast networks, respectively. In contrast, our use of TimeSformers on thermal data alone yielded F1-scores of 85.4% for smoking and 85.6% for eating, highlighting thermal data's potential to learn eating gestures. Despite these promising results, deploying transformer-based architectures on-device remains challenging. However, recent advancements in transformer technology, such as one-bit based mixtures of experts [68], and ongoing research into on-device integration [69, 70], suggest potential pathways for implementation. To overcome the limitations of resource-constrained devices, combining thermal and RGB data could enable smaller models to extract more comprehensive features, thereby improving on-device eating gesture detection.

7.4 Model Robustness

As depicted in Figure 19, our eating and smoking detection models demonstrated robustness not only during the night or dimly lit conditions but also during scenarios of intense head movement. The training dataset included participants with a diverse representation of age, sex, race, and ethnicity, as detailed in Table 5. To further evaluate robustness, we implemented an additional experiment to assess the impact of incorporating additional participant data into our model training. The results, illustrated in Figure 20, indicated that the F1-score of our hand-to-mouth detection model initially increased and then reached a plateau, showing no significant improvement as more participant data were added.

7.5 Cigarette vs. E-Cigarette: Gesture Detection Challenges

HabitSense aims to detect hand-to-mouth gestures associated with eating and cigarette smoking. Despite the increasing prevalence of electronic cigarettes (e-cigs), smoking remains the predominant preventable cause of death in the U.S., responsible for approximately 480,000 deaths annually [4], or approximately one in every five individuals. Therefore, in our free-living study, we recruited participants who primarily smoked cigarettes, averaging at least ten daily (recruitment details in section 5.2). However, we also found instances of e-cig use in our study data. Such instances were labeled as non-cigarette-smoking gestures, resulting in high precision observed in our LOPO analysis for P8. This is detailed in Table 9 of the on-device smoking detection model (thermal data), participant-level analysis. It is important to note that because the SECURE pipeline does not rely on specific eating or smoking detection models for data recording, relying instead on hand-to-mouth gestures, it will still record e-cig use instances.

The growing prevalence of e-cig use, which poses confounding factors for both conventional smoking and the recognition of eating gestures, warrants dedicated investigation [71]. To evaluate the effectiveness of HabitSense on different e-cigs, we performed preliminary tests on three popular devices on the market: Juul [72], Posh Max 5200 [73], and IQOS [74]. Figure 15 illustrates the heat signatures of a conventional cigarette versus the e-cigs

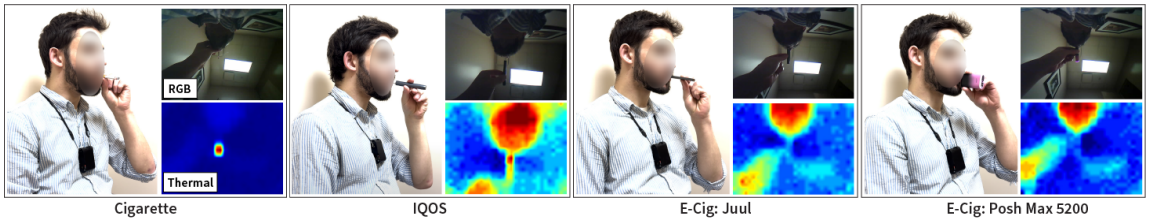


Fig. 15. Heat signature, cigarette vs e-cigarettes

tested. Although none of the e-cigs tested produced pronounced heat signatures, the IQOS signature is more promising than the Juul and Posh Max signatures, suggesting the potential for future iterations of HabitSense to detect e-cigs that operate based on thermal-passing heating mechanisms.

The detection of e-cigs that do not have thermal-passing mechanisms may require the integration of more sophisticated techniques and the fusion of data from both thermal and RGB cameras. This could be made feasible through HabitSense, which supports both modalities, unlike other wearable camera devices such as Smokemon [24], which only includes a thermal camera. Future work can also test the benefits of a gas sensor or enhanced data collection to discern subtle differences in hand-to-mouth gestures between conventional smoking and e-cig use.

7.6 Thermal Sensor Selection

In the transition from the first to the second generation of devices, we replaced the thermal camera with the MLX90640 thermal IR array, which comprises a higher resolution (32x24) and an expanded field of view (110 degrees). This enhancement improved the recognition of hand-to-mouth gestures, increasing the efficacy of the SECURE pipeline. However, the MLX90640 has a notable current draw of approximately 23 mA, as reported in the sensor specifications. During the operation of HabitSense, predominantly at level 3 of the SECURE pipeline, the thermal camera was responsible for 28% of total energy consumption. Another option exists with the MLX90641 sensor, which has a 110-degree field of view but a lower pixel count (16x12), better noise suppression, and lower current draw (12 mA). However, future work is needed to explore the trade-off between improved temperature noise and pixel resolution. If functionality is feasible, this substitution would decrease the current draw at level 2 of the SECURE pipeline from 79.74 mA to 68.74 mA, thus reducing the total energy consumption at level 3 by up to 14%. Another sensor, which would require significant hardware rework due to the higher supply voltage, is the Omron D6T-32L-01A MEMS thermal sensor, which has 1024 pixels (32x32) and a current draw of 12 mA. Future efforts to scale HabitSense should include efforts to optimize the selection of sensors based on the results of this work.

8 CONCLUSION

This research has successfully demonstrated the practicality and efficacy of the HabitSense platform, a novel, open-source RGB-T wearable camera system designed for the continuous monitoring and automated detection of health-risk behaviors such as eating and smoking. Informed by thematic analysis of focus groups with 36 experts, including dietitians and smoking treatment specialists, and bolstered by our iterative design, HabitSense balances user privacy considerations with the need for automated detection of hand-to-mouth activity. One of the system's most notable attributes is its SECURE algorithm. SECURE efficiently triggers obfuscated RGB video recording, significantly improving storage savings, reducing battery consumption, and minimizing overall data accumulation, while still maintaining robust event capture rates. Our comprehensive evaluations, in real-world

scenarios with 15 participants, have validated the high user acceptability and willingness to use HabitSense. Our study yielded extensive RGB-T data (768 hours), enabling the continued development and testing of advanced machine learning models for offline and real-time on-device analysis, which are robust to free-living environments with varying lighting conditions and head posture. HabitSense offers a low-cost, modular, and scalable solution for reliable all-day data collection while also addressing critical privacy concerns through on-device obfuscation. Our evaluation shows no significant difference between thermal-sensing and obfuscated RGB data in detecting eating and cigarette smoking, paving the way for two privacy-conscious approaches to the detection of health risk behaviors. Finally, the ability of HabitSense to objectively monitor health-risk behaviors in real-world scenarios, reducing reliance on self-reporting, elevates the practicality and appeal of wearable cameras in healthcare contexts, setting a new benchmark for the design, development, deployment, and translation of such technologies.

ACKNOWLEDGMENTS

We acknowledge the support from the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (NIH) under award numbers R03DK127128 and R01DK129843, as well as the National Institute of Biomedical Imaging and Bioengineering of the NIH under award number R21EB030305. The opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NIH. The authors also gratefully acknowledge the contributions of colleagues and collaborators Dr. Brian Hitsman, Dr. Angela Pfammatter, Dr. Annie W. Lin, Dr. Yang Gao, Bonnie Nolan, and Dr. Krystina Neuman, for their constructive feedback.

REFERENCES

[1] CDC. Obesity is a Common, Serious, and Costly Disease — cdc.gov. https://www.cdc.gov/obesity/data/adult.html. [Accessed 28-04-2024].
[2] Cynthia L Ogden, Margaret D Carroll, Brian K Kit, and Katherine M Flegal. Prevalence of obesity among adults: United states. *NCHS data brief*, 2013(131):1–8, 2012.
[3] CDCTobaccoFree. Health Effects of Cigarette Smoking — cdc.gov. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm. [Accessed 28-04-2024].
[4] American Lung Association. Tobacco Facts | State of Tobacco Control — lung.org. https://www.lung.org/research/sotc/facts#:~:text=Smoking%20is%20the%20number%20one,in%20the%20U.S.%20each%20year. [Accessed 28-04-2024].
[5] Jonathan M Samet. Tobacco smoking: the leading cause of preventable disease worldwide. *Thoracic surgery clinics*, 23(2):103–112, 2013.
[6] Danielle E Ramo, Johannes Thrul, Erin A Vogel, Kevin Delucchi, and Judith J Prochaska. Multiple health risk behaviors in young adult smokers: stages of change and stability over time. *Annals of Behavioral Medicine*, 54(2):75–86, 2020.
[7] US Department of Health and Human Services. The health consequences of smoking—50 years of progress: A report of the surgeon general. *U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*, 2014.
[8] Bryan Stierman, Joseph Afful, Margaret D Carroll, Te-Ching Chen, Orlando Davy, Steven Fink, Cheryl D. Fryar, Qiuping Gu, Craig M Hales, Jeffery P Hughes, Yechiam Ostchega, Renee J Storandt, and Lara J Akinbami. National health and nutrition examination survey 2017–march 2020 prepandemic data files development of files and prevalence estimates for selected health outcomes. *National Center for Health Statistics (U.S.)*, 2021.
[9] EF Kremer, A Block, and MS Gaylor. Behavioral approaches to treatment of chronic pain: the inaccuracy of patient self-report measures. *Archives of Physical Medicine and Rehabilitation*, 62(4):188–191, 1981.
[10] Xitao Fan, Brent C Miller, Kyung-Eun Park, Bryan W Winward, Mathew Christensen, Harold D Grotevant, and Robert H Tai. An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field methods*, 18(3):223–244, 2006.
[11] Thea F Van de Mortel. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The*, 25(4):40–48, 2008.
[12] Hugh J Arnold and Daniel C Feldman. Social desirability response bias in self-report choice situations. *Academy of Management Journal*, 24(2):377–385, 1981.
[13] Benjamin Singh, Eva M Zopf, and Erin J Howden. Effect and feasibility of wearable physical activity trackers and pedometers for increasing physical activity and improving health outcomes in cancer survivors: A systematic review and meta-analysis. *Journal of sport and health science*, 2021.

[14] Yonatan Vaizman, Nadir Weibel, and Gert Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4), jan 2018.

[15] Tonmoy Ghosh, Yue Han, Viprav Raju, Delwar Hossain, Megan A McCrory, Janine Higgins, Carol Boushey, Edward J Delp, and Edward Sazonov. Integrated image and sensor-based food intake detection in free-living. *Scientific Reports*, 14(1):1665, 2024.

[16] Brooke M Bell, Ridwan Alam, Nabil Alshurafa, Edison Thomaz, Abu S Mondol, Kayla de la Haye, John A Stankovic, John Lach, and Donna Spruijt-Metz. Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review. *NPJ digital medicine*, 3(1):38, 2020.

[17] Mara Ulloa, Blaine Rothrock, Faraz S Ahmad, and Maia Jacobs. Invisible clinical labor driving the successful integration of ai in healthcare. *Frontiers in Computer Science*, 4:1045704, 2022.

[18] Dennis R Louie, Marie-Louise Bird, Carlo Menon, and Janice J Eng. Perspectives on the prospective development of stroke-specific lower extremity wearable monitoring technology: a qualitative focus group study with physical therapists and individuals with stroke. *Journal of neuroengineering and rehabilitation*, 17(1):1–11, 2020.

[19] Rawan Alharbi, Tammy Stump, Nilofar Vafaie, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. I can't be myself: effects of wearable cameras on the capture of authentic behavior in the wild. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 2(3):1–40, 2018.

[20] Jerome H Saltzer. Protection and the control of information sharing in multics. *Communications of the ACM*, 17(7):388–402, 1974.

[21] Jerome H Saltzer and Michael D Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[22] Rawan Alharbi, Sougata Sen, Ada Ng, Nabil Alshurafa, and Josiah Hester. Actisight: Wearer foreground extraction using a practical rgb-thermal wearable. In *2022 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 237–246. IEEE, 2022.

[23] Rawan Alharbi, Mariam Tolba, Lucia C Petito, Josiah Hester, and Nabil Alshurafa. To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 3(3):1–29, 2019.

[24] Rawan Alharbi, Soroush Shahi, Stefany Cruz, Lingfeng Li, Sougata Sen, Mahdi Pedram, Christopher Romano, Josiah Hester, Aggelos K. Katsaggelos, and Nabil Alshurafa. Smokemon: Unobtrusive extraction of smoking topography using wearable energy-efficient thermal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4), jan 2023.

[25] Soroush Shahi, Mahdi Pedram, Glenn Fernandes, and Nabil Alshurafa. Smartact: Energy efficient and real-time hand-to-mouth gesture detection using wearable rgb-t. In *2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4, 2022.

[26] Ziqi Gao, Yuntao wang, Jianguo Chen, Junliang Xing, Shwetak Patel, Xin Liu, and Yuanchun Shi. Mmtsa: Multi-modal temporal segment attention network for efficient human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(3), sep 2023.

[27] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, et al. Auracle: Detecting eating episodes with an ear-mounted sensor. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–27, 2018.

[28] Maria T. Nyamukuru and Kofi M. Odame. Tiny eats: Eating detection on a microcontroller. *CoRR*, abs/2003.06699, 2020.

[29] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. Earbit: Using wearable sensors to detect eating episodes in unconstrained environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), sep 2017.

[30] Iñigo Torres Uribe Echebarría, Syed Anas Imtiaz, Mingxu Peng, and Esther Rodriguez-Villegas. Monitoring smoking behaviour using a wearable acoustic sensor. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4459–4462, 2017.

[31] Volkan Senyurek, Masudul Imtiaz, Prajakta Belsare, Stephen Tiffany, and Edward Sazonov. Cigarette smoking detection with an inertial sensor and a smart lighter. *Sensors*, 19(3):570, 2019.

[32] Paulo Lopez-Meyer, Stephen Tiffany, Yogendra Patil, and Edward Sazonov. Monitoring of cigarette smoking using wearable sensors and support vector machines. *IEEE Transactions on Biomedical Engineering*, 60(7):1867–1872, 2013.

[33] Steve Hodges, Emma Berry, and Ken Wood. Sensecam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory*, 19(7):685–696, 2011.

[34] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[35] Shengjie Bi and David Kotz. Eating detection with a head-mounted video camera. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 60–66, 2022.

[36] Masudul H Imtiaz, Delwar Hossain, Volkan Y Senyurek, Prajakta Belsare, Stephen Tiffany, and Edward Sazonov. Wearable egocentric camera as a monitoring tool of free-living cigarette smoking: a feasibility study. *Nicotine and Tobacco Research*, 22(10):1883–1890, 2020.

[37] Lin Lu, Jiayao Zhang, Yi Xie, F. Gao, Song Xu, Xinghuo Wu, and Z. Ye. Wearable health devices in health care: Narrative systematic review. *JMIR mHealth and uHealth*, 8, 2020.
[38] A. Yetisen, J. L. Martinez-Hurtado, Barış Ünal, A. Khademhosseini, and H. Butt. Wearables in medicine. *Advanced Materials (Deerfield Beach, Fla.)*, 30, 2018.
[39] S. M. A. Iqbal, Imadeldin Mahgoub, E. Du, Mary Ann Leavitt, and W. Asghar. Advances in healthcare wearable devices. *npj Flexible Electronics*, 5:1–14, 2021.
[40] Karim Bayoumy, Mohammed A. Gaber, A. Elshafeey, Omar Mhaimeed, Elizabeth H. Dineen, F. Marvel, S. Martin, E. Muse, M. Turakhia, K. Tarakji, and M. Elshazly. Smart wearable devices in cardiovascular care: where we are and how to move forward. *Nature Reviews. Cardiology*, 18:581 – 599, 2021.
[41] A. Godfrey, Victoria Hetherington, Hubert P. H. Shum, P. Bonato, N. Lovell, and S. Stuart. From a to z: Wearable technology explained. *Maturitas*, 113:40–47, 2018.
[42] P. Bonato. Advances in wearable technology and applications in physical medicine and rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 2:2 – 2, 2005.
[43] Min Chen, Yin Zhang, Yong Li, Mohammad Mehedi Hassan, and Atif Alamri. Aiwac: affective interaction through wearable computing and cloud technology. *IEEE Wireless Communications*, 22:20–27, 2015.
[44] E. McAdams, C. Géhin, B. Massot, and J. McLaughlin. The challenges facing wearable sensor systems. *Studies in health technology and informatics*, 177:196–202, 2012.
[45] H. Lewy. Wearable technologies - future challenges for implementation in healthcare services. *Healthcare technology letters*, 2 1:2–5, 2015.
[46] Lisa A. Simpson, C. Menon, A. Hodgson, W. Ben Mortenson, and J. Eng. Clinicians' perceptions of a potential wearable device for capturing upper limb activity post-stroke: a qualitative focus group study. *Journal of NeuroEngineering and Rehabilitation*, 18, 2021.
[47] Glenn J Fernandes, Arthur Choi, Jacob Michael Schauer, Angela F Pfammatter, Bonnie J Spring, Adnan Darwiche, and Nabil I Alshurafa. An explainable artificial intelligence software tool for weight management experts (primo): Mixed methods study. *Journal of medical Internet research*, 25:e42047, 2023.
[48] Glenn Fernandes, Helen Zhu, Mahdi Pedram, Jacob Schauer, Soroush Shahi, Christopher Romano, Darren Gergle, and Nabil Alshurafa. Is cartoonized life-vlogging the key to increasing adoption of activity-oriented wearable camera systems? In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2023.
[49] Ankur Chattopadhyay and T.E. Boult. Privacycam: a privacy preserving camera using uclinux on the blackfin dsp. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
[50] Thomas Winkler and Bernhard Rinner. Trustcam: Security and privacy-protection for an embedded smart camera based on trusted computing. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 593–600, 2010.
[51] Thomas Winkler, Ádám Erdélyi, and Bernhard Rinner. Trusteye.m4: Protecting the sensor — not the camera. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 159–164, 2014.
[52] Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D. Abowd, and Lauren Wilcox. Students' experiences with ecological momentary assessment tools to report on emotional well-being. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), mar 2018.
[53] Nabil Alshurafa, Haik Kalantarian, Mohammad Pourhomayoun, Shruti Sarin, Jason J Liu, and Majid Sarrafzadeh. Non-invasive monitoring of eating behavior using spectrogram analysis in a wearable necklace. In *2014 IEEE healthcare innovation conference (HIC)*, pages 71–74. IEEE, 2014.
[54] Haik Kalantarian, Nabil Alshurafa, Tuan Le, and Majid Sarrafzadeh. Monitoring eating habits using a piezoelectric sensor-based necklace. *Computers in biology and medicine*, 58:46–55, 2015.
[55] Shibo Zhang, Yuqi Zhao, Dzung Tri Nguyen, Runsheng Xu, Sougata Sen, Josiah Hester, and Nabil Alshurafa. Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 4(2):1–26, 2020.
[56] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. 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*, pages 999–1010, 2015.
[57] Samuel Jero, Juliana Furgala, Runyu Pan, Phani Kishore Gadepalli, Alexandra Clifford, Bite Ye, Roger Khazan, Bryan C Ward, Gabriel Parmer, and Richard Skowyra. Practical principle of least privilege for secure embedded systems. In *2021 IEEE 27th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 1–13. IEEE, 2021.
[58] Nabil Alshurafa, Jayalakshmi Jain, Tammy K Stump, Bonnie Spring, and June K Robinson. Assessing recall of personal sun exposure by integrating uv dosimeter and self-reported data with a network flow framework. *Plos one*, 14(12):e0225371, 2019.
[59] Tammy K Stump, Lisa G Aspinwall, Elizabeth L Gray, Shuai Xu, Nenita Maganti, Sancy A Leachman, Nabil Alshurafa, and June K Robinson. Daily minutes of unprotected sun exposure (muse) inventory: measure description and comparisons to uvr sensor and sun protection survey data. *Preventive medicine reports*, 11:305–311, 2018.

[60] Tammy Stump, Siobhan M Phillips, Jayalakshmi Jain, June K Robinson, Payton Solk, Whitney A Welch, and Nabil Alshurafa. Unprotected sun exposure and physical activity among melanoma survivors and first-degree relatives. In *ANNALS OF BEHAVIORAL MEDICINE*, volume 55, pages S491–S491. OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, 2021.
[61] S Xu, TK Stump, J Jain, N Alshurafa, and JK Robinson. Variation in daily ultraviolet radiation exposure and sun protection behaviours of melanoma survivors: an observational single-arm pilot study with a wearable sensor. *British Journal of Dermatology*, 180(2):413–414, 2019.
[62] HumanSignal. labelimg: Image annotation tool. https://github.com/HumanSignal/labelImg, 2023. Accessed: 2023-11-11.
[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
[64] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
[65] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.
[66] Yiyi Xu et al. Timesformer-rolled-attention: Github repository. https://github.com/yiyixuxu/TimeSformer-rolled-attention, 2023. Accessed: 2023-11-09.
[67] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Timesformer: Github repository. https://github.com/facebookresearch/TimeSformer, 2023. Accessed: 2023-11-09.
[68] Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit llms: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024.
[69] Atila Orhon, Aseem Wadhwa, Youchang Kim, Francesco Rossi, and Vignesh Jagadeesh. Deploying transformers on the apple neural engine, 2023. Accessed: 2023-11-15.
[70] Apple. Apple introduces the advanced new apple watch series 9, Sep 2023. Accessed: 2023-11-15.
[71] Authors of the study. The prevalence of electronic cigarettes vaping globally: a systematic review and meta-analysis. *Archives of Public Health*, 80, 2022.
[72] https://www.juul.com/. [Accessed 27-04-2024].
[73] Posh Max 2.0 | Rechargeable Disposable Vape | Best Flavors — nowposh.com. https://nowposh.com/posh-max-2-0/. [Accessed 28-04-2024].
[74] iqos.com. https://www.iqos.com/global. [Accessed 28-04-2024].
[75] Mehrab Bin Morshed, Samruddhi Shreeram Kulkarni, Richard Li, Koustuv Saha, Leah Galante Roper, Lama Nachman, Hong Lu, Lucia Mirabella, Sanjeev Srivastava, Munmun De Choudhury, et al. A real-time eating detection system for capturing eating moments and triggering ecological momentary assessments to obtain further context: System development and validation study. *JMIR mHealth and uHealth*, 8(12):e20625, 2020.
[76] Mehrab Bin Morshed, Harish Kashyap Haresamudram, Dheeraj Bandaru, Gregory D Abowd, and Thomas Plötz. A personalized approach for developing a snacking detection system using earbuds in a semi-naturalistic setting. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, pages 11–16, 2022.
[77] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. Mydj: Sensing food intakes with an attachable on your eyeglass frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.
[78] Edwin Valarezo Añazco, Patricio Rivera Lopez, Sangmin Lee, Kyungmin Byun, and Tae-Seong Kim. Smoking activity recognition using a single wrist imu and deep learning light. In *Proceedings of the 2nd international conference on digital signal processing*, pages 48–51, 2018.
[79] Volkan Y Senyurek, Masudul H Imtiaz, Prajakta Belsare, Stephen Tiffany, and Edward Sazonov. Smoking detection based on regularity analysis of hand to mouth gestures. *Biomedical signal processing and control*, 51:106–112, 2019.
[80] Edward Sazonov, Paulo Lopez-Meyer, and Stephen Tiffany. A wearable sensor system for monitoring cigarette smoking. *Journal of studies on alcohol and drugs*, 74(6):956–964, 2013.
[81] Li-Chia Tai, Christine Heera Ahn, Hnin Yin Yin Nyein, Wenbo Ji, Mallika Bariya, Yuanjing Lin, Lu Li, and Ali Javey. Nicotine monitoring with a wearable sweat band. *ACS sensors*, 5(6):1831–1837, 2020.
[82] Samuel L. Battalio, David E. Conroy, Walter Dempsey, Peng Liao, Marianne Menictas, Susan Murphy, Inbal Nahum-Shani, Tianchen Qian, Santosh Kumar, and Bonnie Spring. Sense2stop: A micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.
[83] Find Sunrise and Sunset Times Anywhere - SunriseSunset.io — sunrisesunset.io. https://sunrisesunset.io/. [Accessed 29-04-2024].

9 SUPPLEMENTARY

9.1 Related-works: Wearables for Eating and Smoking Detection

9.1.1 IMU-based sensing. Wearable technologies like smartwatches, commonly equipped with inertial measurement units (IMUs), have gained popularity for automatically detecting eating and smoking activities. This functionality helps overcome certain limitations inherent in self-reporting methods. IMU-based approaches to eating detection largely focus on extracting hand-to-mouth gestures from motion data, typically collected at the wrist [16]. For instance, Morshed et al. [75] achieved real-time eating detection at the meal level by training a classifier to recognize eating gestures using smartwatch accelerometer data. Other on-body IMU sensor locations have also been explored. One example is the earbud accelerometer in the study by Morshed et al. [76], designed to detect chewing motions. These earbud sensors are less influenced by confounding hand-to-mouth gestures typical of wrist-worn IMUs, enabling more precise differentiation between eating activities such as meal consumption and snacking.

IMUs have also supported multimodal systems where non-IMU sensors primarily detect eating activities, such as in a study by Zhang et al. [55]. Their necklace device employs proximity and ambient light sensors aimed at the jaw to measure chews directly, while the IMU supplements by estimating the lean-forward angle, enhancing the prediction model. Bedri et al. [34] embedded an IMU and proximity sensor into eyeglasses. This design enables detecting eating gestures through hand-to-mouth proximity and captures chews from IMU-registered movements of the facial muscles. Similarly, recent developments feature the MyDJ device [77], which integrates both IMU and piezoelectric sensors onto eyeglasses for eating detection.

IMU-based systems have also been explored for smoking detection. Añazco et al. [78] demonstrated the efficacy of wrist-worn IMUs in detecting smoking-related hand gestures, successfully differentiating them from confounding activities such as eating and drinking. However, due to variations in gestures across individuals, their method lacks generalizability. To address this, Senyurek et al. [79] focused on analyzing gesture regularity from wrist-based IMUs, although performance significantly decreased in free-living settings. Further attempts to enhance performance led Senyurek et al. [31] to integrate a smart lighter with a wrist-worn three-axis accelerometer. Still, inconsistencies arose due to self-report errors and instances of participants using their lighters instead of the provided smart lighter.

To summarize, although wrist-worn IMU sensing modalities are recognized for their ability to detect eating and smoking activities, they face challenges in practical application. These include false positives from unrelated movements, issues with sensor positioning (i.e., dependency on the dominant hand), and difficulties in generalizing across different individuals [16]. However, IMU-based sensors continue to supplement multi-modal devices [26].

9.1.2 Acoustic-based sensing. Another area explored uses acoustics to capture sonic features of chews, swallows, and jaw movements. Bi et al. [27] found that chewing sounds could be retrieved from audio collected with a contact microphone on the mastoid bone. Nyamukuru et al. [28] improved on this approach by implementing a shallow gated recurrent unit neural network optimized for execution on a low-power microcontroller, detecting mastoid chewing sound with greater accuracy and efficiency. Acoustic sensors have been employed in multi-modal systems, such as in a study by Bedri et al. [29] where a neck-positioned microphone detects swallowing sounds. Additionally, an earbud equipped with an IMU and a proximity sensor was used to identify chewing. Acoustic sensing modalities have also been explored to detect smoking activities. Echevarria et al. [30] demonstrated the potential of wearable acoustic sensors, designed for respiratory monitoring, to identify smoking behavior by capturing the distinct acoustic properties of smoking inhalation. However, microphones are susceptible to noise interference, particularly in free-living environments. Moreover, using microphones raises significant privacy concerns in everyday settings [19].

9.1.3 Alternate sensing modalities. Several studies have explored alternate sensing modalities and methodologies to monitor cigarette smoking behavior precisely. One approach involved the development of the PACT system [80], which integrated a comprehensive sensor suite consisting of a respiratory inductance plethysmograph and hand gesture sensor capable of capturing multiple behavioral indicators associated with cigarette smoking. Using support vector machines trained on data from the PACT system, Lopez-Meyer et al. [32] highlighted the feasibility of autonomously identifying smoke inhalations by continuously monitoring respiratory patterns and hand-to-mouth gestures. Although this methodology holds promise for delivering precise and objective evaluations of smoking behaviors, the data collection and assessment of the PACT system were conducted within a controlled environment, limiting generalization to free-living contexts. Studies have also explored the feasibility of using wearable sweatbands to monitor nicotine levels in the sweat, thereby inferring smoking behavior [81]. Although the prospects appear promising, the frequent need for calibration, coupled with potential contamination from passive smokers' smoke deposition, could compromise the accuracy of readings. Battalio et al. conducted a micro-randomized trial using wrist-worn and ECG sensors, intending to refine a just-in-time-adaptive stress management intervention to mitigate smoking relapse [82]. This research emphasized the potential of wearable sensors in facilitating real-time data acquisition, thereby supporting personalized interventions.

Although current methods for automated detection of eating and smoking are promising, deploying these methods in free-living settings often relies on self-reporting or external tools, which can introduce inaccuracies and biases into the ground truth data. Additionally, there is no visual evidence to confirm the occurrence of the events.

9.2 Prior Deployment Protocols

9.2.1 Gen 1 Study. With the Gen 1 device, we collected 14 waking days of data with 60 participants in a free-living setting. Participants attended a laboratory visit to introduce them to the Gen 1 device and calibrate the neckband for optimal body placement. Participants were then instructed to take the device home, equip and turn on the device each morning, perform a synchronization event by aiming the device at a digital clock smartphone application, wear the device throughout the day, remove and turn off the device at the end of the day, and charge the device overnight each night. Participants were informed that recording status was indicated by a blue LED light inside the device's enclosure (blinking = on, solid/absent = off).

9.2.2 Gen 2 Study. We deployed the Gen 2 device in a free-living study to collect 14 waking days of data from 30 participants, systematically varying which obfuscation technique (blurring, edge obfuscation, cartoon/avatar obfuscation, or raw data/no obfuscation) each participant's device used to test the impact of obfuscation method on wear time. After a laboratory visit to introduce participants to the Gen 2 device and calibrate the neckband for optimal body placement, participants were instructed to take the device home, equip and turn on the device each morning, wear the device throughout the day, remove and turn off the device at the end of the day, and charge the device overnight each night. Participants were informed that recording status was indicated by a blue LED light inside the device's enclosure (blinking = on, solid/absent = off). Participants used a provided laptop to view brief clips of their data collected throughout the day, allowing them to view the effect of the obfuscation technique on their collected data, thus allowing the obfuscation method to affect their wear time.

9.3 Methods: Iterative Case Design

9.3.1 Iterations through different 3D printing technoglogies. The initial prototypes of the device's enclosure, depicted in Figure 16, were fabricated using fused deposition modeling (FDM) with polylactic acid due to its rapid and accessible printing capabilities despite producing lower-fidelity prints. Progressing to later stages, we adopted Stereolithography (SLA) with a FormLabs printer, capitalizing on its ability to handle resins with varied mechanical properties and create complex geometries beyond the scope of FDM. The higher resolution of the FormLabs

Table 10. Hand-to-mouth (H2M) statistics across participants.

Participant	Num. of H2M Gestures				H2M Gestures	
	Smoking	Eating	Drinking	Total	Total Frames	Total Duration (mins)
P1	non-smoker	128	58	186	2130	7.1
P2	non-smoker	538	481	1019	9574	31.9
P3	non-smoker	210	177	387	5547	18.5
P4	non-smoker	1233	299	1532	15750	52.5
P5	non-smoker	18	17	35	503	1.7
P6	957	486	283	1726	17892	59.6
P7	157	546	76	779	9610	32.0
P8	7	434	101	542	6257	20.9
P9	294	631	140	1065	10636	35.5
P10	1040	109	302	1451	29161	97.2
P11	531	66	43	640	6441	21.5
P12	79	199	54	332	4766	15.9
P13	non-smoker	582	219	801	8005	26.7
Mean	438	398	173	807	9713	32.4
Total	3065	5180	2250	10495	126272	420.9

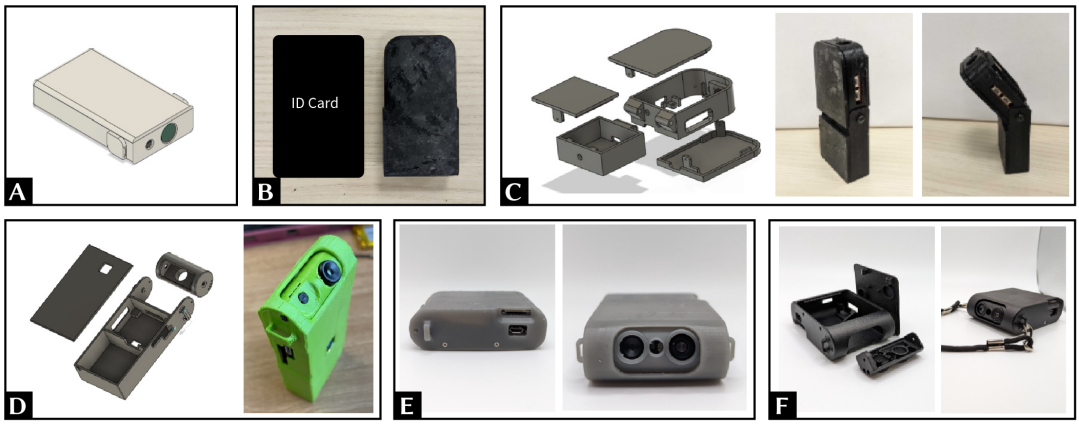


Fig. 16. Iterative Design of HabitSense Enclosure. (A) Slim initial 3D design, (B) size comparison with ID card, (C) two-piece hinge-connected enclosure, (D) rotatable internal mechanism for sensor adjustment, (E) revised enclosure with optimized dimensions, (F) final product-quality design.

printer was crucial, as it permitted the precise replication of small design features and components, achieving an optimal fit with minimal tolerances. Employing high-resolution SLA printing with a FormLabs printer allowed for the precise placement of RGB and thermal sensors, facilitating the standardization of calibration parameters across all devices, a process detailed in section 5.4 – on-device obfuscation. For the final design iteration, we utilized selective laser sintering (SLS) technology with nylon to surpass the resolution offered by SLA. We applied vapor smoothing to the SLS prints to alleviate their inherent rough texture.

9.3.2 Iterations in design: hinge mechanism. We systematically enhanced the device design through empirical prototype testing using various 3D printing techniques. Generations 1 and 2 utilized L-shaped components set at diverse angles, allowing the device to be positioned by leveraging the L-shape for variable inclinations. We made a notable improvement in Gen3 – HabitSense by adding a hinging mechanism into the enclosure itself to adjust the field of view for wearers, as documented in Figure 16. Initial designs for this mechanism included separate compartments for the battery and circuitry, resulting in a cumbersome design with exposed wiring [Fig 16C]. Later designs [Fig 16 D,E,F] compacted the structure by limiting the hinge's movement and relocating only the thermal and RGB cameras to a rotatable cylindrical section, which could be fixed at specific angles with screws for stability. These modifications not only reduced the bulkiness compared to Generations 1 and 2 but also improved the adaptability for users with different body types We also refined the case design from an initial compact, rectangular shape with sharp corners to a form with rounded edges, improving ergonomics and comfort. Adopting SLA and SLS printing technologies enabled the creation of these curves, which were not attainable with FDM. Despite a slight increase in size, this trade-off was considered beneficial for its ergonomic and aesthetic gains.

9.3.3 Magnetic backplate design. Ensuring the device remained stationary and did not sway when worn was critical for collecting high-quality data. To address the detachment issues and instability stemming from an inadequately strong magnetic plate in the Gen 1 and Gen 2 designs, which led to the device swaying or detaching during moderately active movements, we developed a new magnetic back plate. This redesigned plate featured a triangular shape with three points of contact, utilizing three magnets instead of the previous two, providing a more robust and secure device attachment through the fabric. The advancements in the system's volume and weight, as delineated in Table 3 (Gen 3: HabitSense), along with the adoption of stronger magnets and a triangular design for augmented contact points, effectively anchored the device in place, eliminating any sway. Consequently, users could wear the device and engage in their daily routines without the distraction or concern of its stability, assured by the magnetic backplate's reliable hold.

9.3.4 Privacy-preserving case design. We designed the device's case to position the RGB and thermal cameras towards the user's face, concentrating data collection on the target activity—hand-to-mouth gestures associated with eating and smoking. This orientation minimized the recording of bystanders, maintaining privacy for those who had not given consent.

9.4 Methods: Calculations for Day/Night, Variable Lighting Conditions, and Head Movement

The data collection for HabitSense included a free-living study where participants used the device continuously for 7 days, with unrestricted usage during both day and night. This resulted in data from varied lighting conditions—darker and brighter environments—and included variations in head movement. To ensure a rigorous evaluation, it is crucial to test the robustness of our models under these diverse conditions.

Using an API [83], we obtained the sunrise and sunset times for the participants' locations on the days they wore the device. Subsequently, we calculated the number of eating, smoking, and drinking gestures recorded during daytime versus nighttime.

Addressing potential concerns about the low-light performance of HabitSense, particularly as participants might wear the device at night in well-lit indoor environments, we analyzed the perceived luminosity of each of the 13 million RGB frames, totaling approximately 768 hours of data. Luminosity was calculated for each pixel; thus, for the entire frame, luminosity was determined by averaging the values for all the pixels in frames. We defined a luminosity threshold to distinguish between brighter and darker frames, guided by the perceptual evaluation of light intensity in RGB images and the statistical distribution of luminosity values across all frames (refer to Figure 17). The median of this distribution served as the threshold, satisfying our criteria for differentiating light

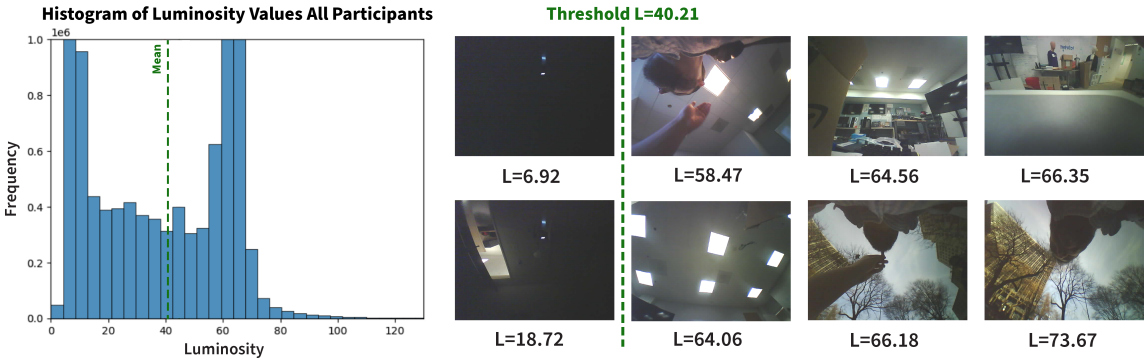


Fig. 17. Luminosity threshold determination

conditions. Accordingly, frames exceeding this threshold were classified as bright and those below as dark.

HabitSense was meticulously designed to ensure that the field of view (FOV) of our camera system lenses comprehensively captures the full rotational and translational motion of the wearer's neck and head (figure 18). To address the challenges associated with accurately detecting eating and smoking activities amid frequent head movements, it was essential first to quantify these movements precisely. We annotated the wearer's head with a bounding box and the nose with a keypoint across the 13 million frames collected from our free-living study. We analyzed head movements using three key metrics (Figure 18):

- The Euclidean distance between the centroids of the bounding boxes across adjacent frames to measure the movement of the head across frames.
- The amount of neck rotation was measured using the nose key points and the centroid of the bounding box across frames.
- The change in the area of the bounding box across adjacent frames to measure the movement of the head towards and away from the camera.

These metrics provided a comprehensive measure of head movements, encompassing both rotational and translational head motions captured throughout our study. We calculated these metrics for all eating and smoking segments in our dataset and identified instances of high versus low head motion.

We assessed the robustness of our models across various sub-categories of data and incorporated attention maps in the supplementary materials as an interpretability metric. This allows us to determine whether the models focus on relevant features for accurate gesture detection.

9.5 Results: Data Statistics for Day/Night, Variable Lighting Conditions and Head Movement

To assess the robustness of our models, we calculated statistics for different conditions, including day/night, brightness levels, and head motion intensity. Table 11 shows that 77.07% of the eating and smoking gestures were recorded during the daytime, with the remaining 22.93% occurring at night. Furthermore, as indicated in Table 12, 66.37% of the gestures were recorded under bright conditions, while 33.63% were in darker settings. According to Table 13, 80.85% of the gestures involved high-intensity head motion, and 19.15% featured low-intensity movements. These data points highlight the variability in our dataset and highlight the importance of testing model performance across these diverse conditions.

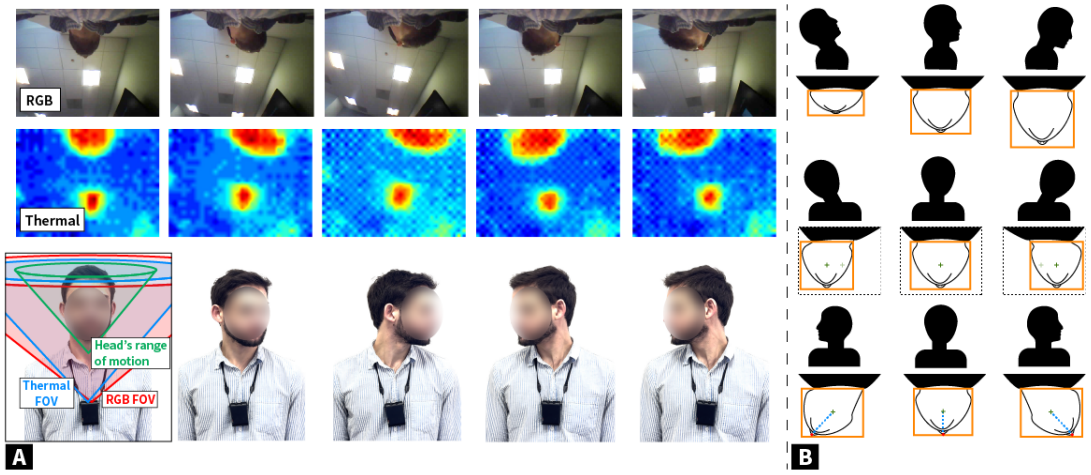


Fig. 18. (A) Head movement range and fields of view for RGB and thermal cameras, (B) Head movement characterization, including extension and flexion (top), lateral flexion (middle), and rotation (bottom)

Table 11. Daytime-Nightime table

	no. of gestures captured during day time				no. of gestures captured during night time			
Participant	Smoking	Eating	Drinking	Total	Smoking	Eating	Drinking	Total
P1		128	48	176		0	10	10
P2		386	350	736		152	131	283
P3		95	68	163		115	109	224
P4		1184	266	1450		49	33	82
P5		8	13	21		10	4	14
P6	644	455	194	1293	313	31	89	433
P7	136	539	70	745	21	7	6	34
P8	7	406	48	461	0	28	53	81
P9	164	297	55	516	130	334	85	549
P10	573	93	184	850	467	16	118	601
P11	507	66	43	616	24	0	0	24
P12	72	177	48	297	7	22	6	35
P13		566	199	765		16	20	36
Mean	300	338	122	622	137	60	51	185
Total	2103	4400	1586	8089	962	780	664	2406
% of gestures				77%				23%

9.6 Results: Model Robustness Performance Metrics

As seen in figure 19, our eating detection model demonstrates robust performance, achieving F1-scores exceeding 89% in both day and night conditions, as well as during high-intensity head motions. In darker settings, the model

Table 12. Luminosity table

	no. of gestures in bright environments				no. gestures in dark environments			
Participant	Smoking	Eating	Drinking	Total	Smoking	Eating	Drinking	Total
P1		128	52	180		0	6	6
P2		391	438	829		147	43	190
P3		78	17	95		132	160	292
P4		902	203	1105		331	96	427
P5		18	13	31		0	4	4
P6	317	331	167	815	640	155	116	911
P7	46	142	34	222	111	404	42	557
P8	7	336	88	431	0	98	13	111
P9	271	601	135	1007	23	30	5	58
P10	749	66	208	1023	291	43	94	428
P11	327	32	17	376	204	34	26	264
P12	47	117	30	194	32	82	24	138
P13		477	181	658		105	38	143
Mean	252	278	122	536	186	120	51	271
Total	1764	3619	1583	6966	1301	1561	667	3529
% of gestures				66%				34%

Table 13. Head Motion Table

	no. of gestures in high motion				no. gestures in low motion			
Participant	Smoking	Eating	Drinking	Total	Smoking	Eating	Drinking	Total
P1		105	43	148		22	15	37
P2		456	398	854		74	74	148
P3		156	113	269		49	44	93
P4		956	229	1185		188	46	234
P5		16	14	30		1	3	4
P6	643	424	221	1288	171	44	34	249
P7	81	405	36	522	24	120	23	167
P8	5	359	69	433	0	64	17	81
P9	243	551	102	896	45	64	38	147
P10	665	65	184	914	353	41	117	511
P11	426	58	32	516	40	0	3	43
P12	58	171	40	269	4	20	5	29
P13		451	131	582		58	71	129
Mean	303	321	124	608	91	57	38	144
Total	2121	4173	1612	7906	637	745	490	1872
% of gestures				81%				19%

achieves an F1-score of 85.58%. Similarly, our smoking detection model shows high efficacy, with F1-scores above

85% during daytime and high-intensity head motions. However, during nighttime and low-light conditions, the model's F1-scores are slightly lower, registering at 79.07% and 83.84%, respectively.

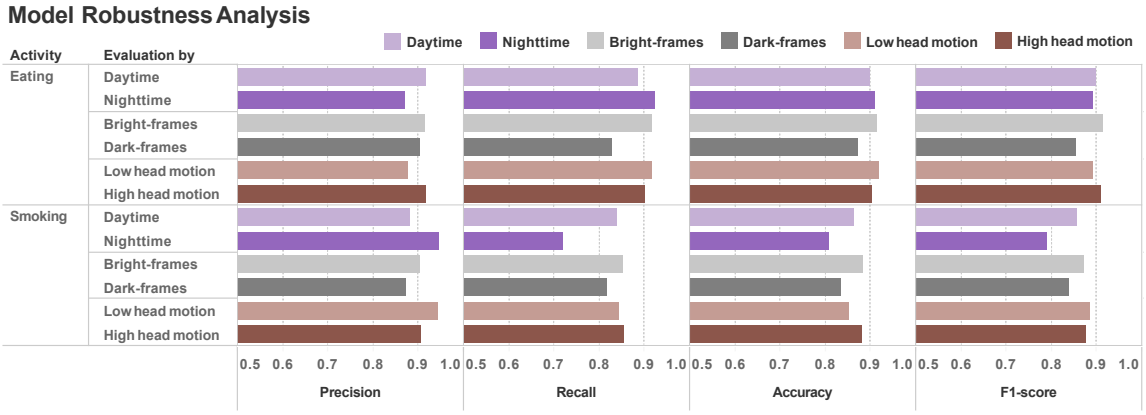


Fig. 19. Perfomance of eating and smoking detection TimeSformer models trained on RGB data in different circumstances to highlight robustness

9.7 Individual Performance metrics

The following list of tables provides individual performance metrics for each of the machine learning models we trained across different settings and data modalities:

- Eating Detection, Off-device RGB: Table 16
- Eating Detection, Off-device Thermal: Table 17
- Eating Detection, On-device Thermal: Table 18
- Eating Detection, Off-device RGB Obfuscated (Mask): Table 14
- Eating Detection, Off-device RGB Obfuscated (Blur): Table 15
- Smoking Detection, Off-device RGB: Table 25
- Smoking Detection, Off-device Thermal: Table 26
- Smoking Detection, On-device Thermal: Table 27
- Smoking Detection, Off-device RGB Obfuscated (Mask): Table 28
- Smoking Detection, Off-device RGB Obfuscated (Blur): Table 24
- H2M Detection, Off-device RGB: Table 21
- H2M Detection, Off-device Thermal: Table 22
- H2M Detection, On-device Thermal: Table 23
- H2M Detection, Off-device RGB Obfuscated (Mask): Table 19
- H2M Detection, Off-device RGB Obfuscated (Blur): Table 20
- Multiclass Detection, Off-device RGB (Weighted Scores): Table 31
- Multiclass Detection, Off-device Thermal (Weighted Scores): Table 32
- Multiclass Detection, On-device Thermal (Weighted Scores): Table 33
- Multiclass Detection, Off-device RGB Obfuscated (Mask) (Weighted Scores): Table 29
- Multiclass Detection, Off-device RGB Obfuscated (Blur) (Weighted Scores): Table 30

Received 15 November 2023; revised 1 May 2024

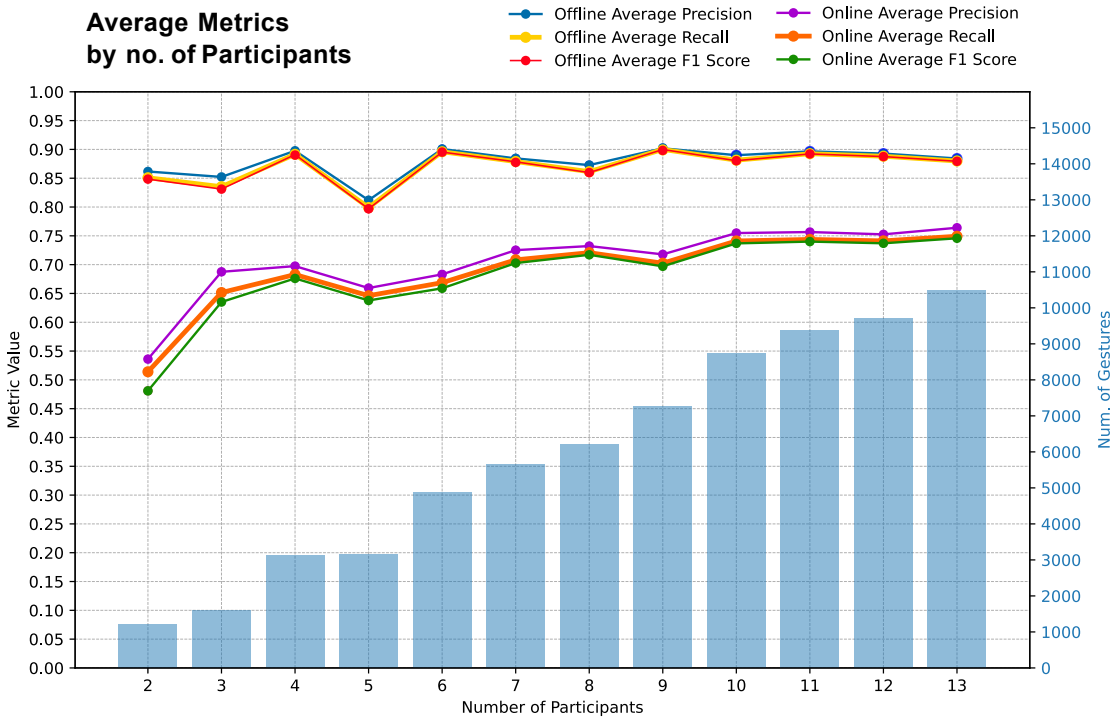


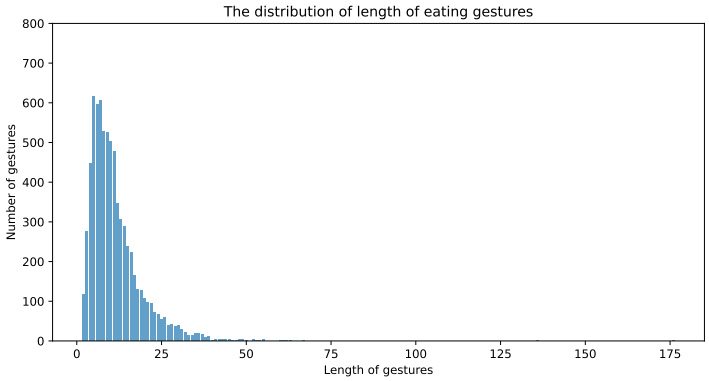
Fig. 20. This figure illustrates the results of our incremental data addition experiment designed to evaluate the impact of incorporating additional participant data into our model training.

Table 14. eating-offdevice-rgb-obfuscated-mask

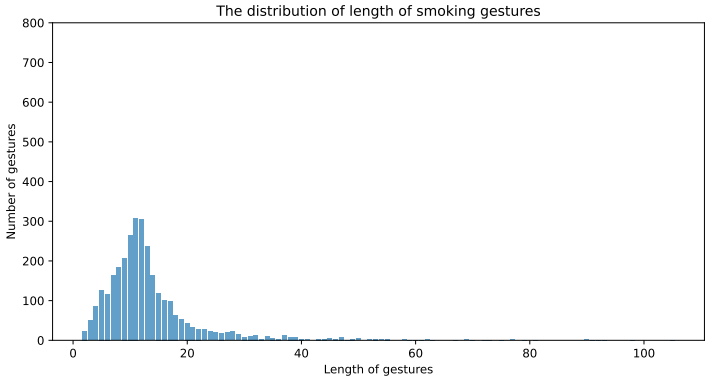
pid	precision	recall	f1_score	accuracy
P1	0.80	0.76	0.78	0.78
P2	0.84	0.89	0.87	0.86
P3	0.85	0.77	0.81	0.82
P4	0.91	0.74	0.82	0.83
P5	0.86	0.86	0.86	0.86
P6	0.79	0.82	0.80	0.80
P7	0.80	0.67	0.73	0.75
P8	0.95	0.53	0.68	0.75
P9	0.85	0.79	0.82	0.82
P10	0.63	0.82	0.71	0.67
P11	0.65	0.83	0.73	0.69
P12	0.88	0.79	0.83	0.84
P13	0.93	0.79	0.86	0.87
Mean	0.82	0.77	0.79	0.80

Table 15. eating-offdevice-rgb-obfuscated-blur

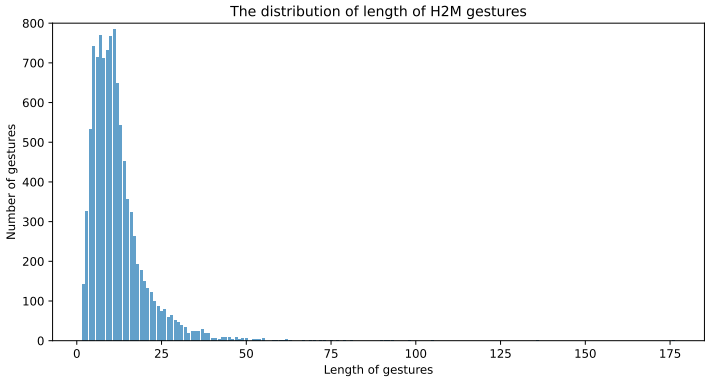
pid	precision	recall	f1_score	accuracy
P1	0.88	0.92	0.90	0.90
P2	0.86	0.93	0.89	0.89
P3	0.92	0.88	0.90	0.90
P4	0.95	0.79	0.86	0.87
P5	0.82	0.80	0.81	0.81
P6	0.87	0.91	0.89	0.89
P7	0.88	0.65	0.75	0.78
P8	0.96	0.65	0.78	0.81
P9	0.88	0.84	0.86	0.86
P10	0.72	0.85	0.78	0.76
P11	0.71	0.81	0.76	0.74
P12	0.86	0.94	0.90	0.89
P13	0.92	0.88	0.90	0.90
Mean	0.86	0.83	0.84	0.85



(a) Eating gesture (mean=10)



(b) Smoking Gesture (mean=12)



(c) Hand-to-mouth gestures (mean=10)

Fig. 21. Distribution of number of frames comprising one eating gesture (mean=10), smoking gesture (mean=12), and hand-to-mouth gestures (mean=10)

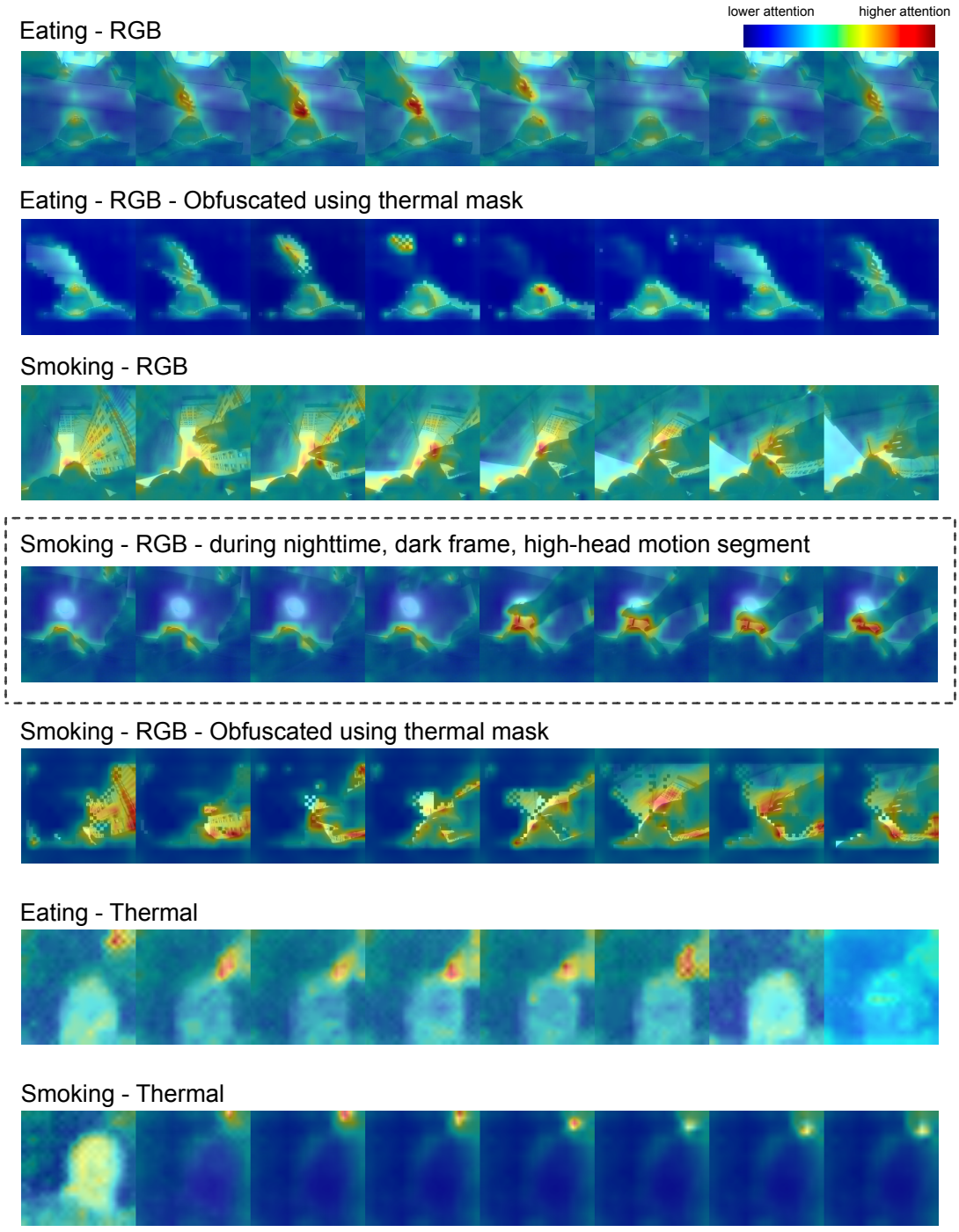


Fig. 22. Attention maps

Table 16. eating-offdevice-rgb

pid	precision	recall	f1_score	accuracy
P1	0.92	0.91	0.92	0.92
P2	0.92	0.93	0.92	0.92
P3	0.96	0.84	0.89	0.90
P4	0.96	0.80	0.87	0.88
P5	0.94	0.91	0.93	0.93
P6	0.86	0.98	0.92	0.91
P7	0.90	0.83	0.87	0.87
P8	0.95	0.83	0.89	0.89
P9	0.93	0.94	0.93	0.93
P10	0.77	0.84	0.80	0.79
P11	0.88	0.85	0.87	0.87
P12	0.90	0.95	0.93	0.92
P13	0.95	0.91	0.93	0.93
Mean	0.91	0.89	0.90	0.90

Table 17. eating-offdevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.84	0.97	0.90	0.89
P2	0.96	0.63	0.76	0.80
P3	0.89	0.94	0.91	0.91
P4	0.87	0.90	0.88	0.88
P5	0.76	0.80	0.78	0.77
P6	0.79	0.89	0.84	0.83
P7	0.92	0.90	0.91	0.91
P8	0.95	0.81	0.87	0.88
P9	0.90	0.87	0.88	0.88
P10	0.92	0.85	0.88	0.89
P11	0.81	0.89	0.85	0.84
P12	0.87	0.87	0.87	0.87
P13	0.86	0.82	0.84	0.84
Mean	0.87	0.86	0.86	0.86

Table 18. eating-ondevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.73	0.49	0.58	0.65
P2	0.81	0.60	0.69	0.73
P3	0.72	0.53	0.61	0.66
P4	0.71	0.60	0.65	0.68
P5	0.60	0.48	0.53	0.58
P6	0.58	0.47	0.52	0.57
P7	0.67	0.63	0.65	0.66
P8	0.78	0.46	0.58	0.66
P9	0.45	0.24	0.31	0.47
P10	0.57	0.71	0.63	0.59
P11	0.55	0.78	0.64	0.57
P12	0.65	0.63	0.64	0.65
P13	0.96	0.54	0.69	0.76
Mean	0.67	0.55	0.60	0.63

Table 19. h2m-offdevice-rgb-obfuscated-mask

pid	precision	recall	f1_score	accuracy
P1	0.83	0.82	0.82	0.82
P2	0.86	0.94	0.90	0.89
P3	0.91	0.90	0.91	0.91
P4	0.91	0.87	0.89	0.89
P5	0.85	0.80	0.82	0.83
P6	0.92	0.89	0.91	0.91
P7	0.91	0.77	0.83	0.85
P8	0.92	0.85	0.88	0.89
P9	0.93	0.89	0.91	0.91
P10	0.87	0.88	0.87	0.87
P11	0.78	0.87	0.82	0.81
P12	0.89	0.86	0.88	0.88
P13	0.92	0.79	0.85	0.86
Mean	0.88	0.86	0.87	0.87

Table 20. h2m-offdevice-rgb-obfuscated-blur

pid	precision	recall	f1_score	accuracy
P1	0.88	0.89	0.89	0.89
P2	0.86	0.98	0.92	0.91
P3	0.94	0.95	0.94	0.94
P4	0.95	0.88	0.91	0.92
P5	0.89	0.91	0.90	0.90
P6	0.96	0.90	0.93	0.93
P7	0.91	0.84	0.87	0.88
P8	0.96	0.89	0.92	0.92
P9	0.96	0.89	0.92	0.93
P10	0.91	0.84	0.87	0.88
P11	0.80	0.83	0.82	0.81
P12	0.89	0.84	0.86	0.87
P13	0.94	0.87	0.90	0.91
Mean	0.91	0.88	0.90	0.90

Table 21. h2m-offdevice-rgb

pid	precision	recall	f1_score	accuracy
P1	0.91	0.86	0.89	0.89
P2	0.91	0.99	0.95	0.94
P3	0.95	0.91	0.93	0.93
P4	0.96	0.93	0.94	0.94
P5	0.87	0.94	0.90	0.90
P6	0.95	0.96	0.96	0.96
P7	0.93	0.92	0.92	0.92
P8	0.95	0.95	0.95	0.95
P9	0.95	0.98	0.96	0.96
P10	0.94	0.90	0.92	0.92
P11	0.90	0.89	0.90	0.90
P12	0.92	0.89	0.90	0.90
P13	0.95	0.90	0.92	0.92
Mean	0.93	0.92	0.93	0.93

Table 22. h2m-offdevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.87	0.90	0.89	0.88
P2	0.96	0.66	0.78	0.82
P3	0.88	0.93	0.90	0.90
P4	0.91	0.91	0.91	0.91
P5	0.73	0.77	0.75	0.74
P6	0.91	0.89	0.90	0.90
P7	0.93	0.84	0.88	0.89
P8	0.94	0.87	0.90	0.91
P9	0.92	0.89	0.91	0.91
P10	0.87	0.83	0.85	0.85
P11	0.82	0.86	0.84	0.84
P12	0.89	0.93	0.91	0.90
P13	0.89	0.79	0.84	0.85
Mean	0.89	0.85	0.87	0.87

Table 23. h2m-ondevice-thermal

pid	precision	recall	f1_score	accuracy
P1	0.88	0.42	0.57	0.68
P2	0.75	0.72	0.74	0.74
P3	0.79	0.68	0.73	0.75
P4	0.77	0.69	0.73	0.74
P5	0.68	0.48	0.56	0.62
P6	0.80	0.71	0.76	0.77
P7	0.85	0.69	0.76	0.78
P8	0.68	0.73	0.71	0.69
P9	0.86	0.50	0.63	0.71
P10	0.72	0.79	0.75	0.74
P11	0.72	0.86	0.79	0.77
P12	0.78	0.82	0.80	0.80
P13	0.97	0.58	0.73	0.78
Mean	0.79	0.67	0.71	0.74

Table 24. smoking-offdevice-rgb-obfuscated-blur

pid	precision	recall	f1 score	accuracy
P6	0.88	0.85	0.87	0.87
P7	0.81	0.92	0.86	0.85
P8	0.50	0.29	0.36	0.50
P9	0.97	0.86	0.91	0.92
P10	0.88	0.56	0.69	0.74
P11	0.76	0.69	0.73	0.74
P12	0.82	0.85	0.83	0.83
Mean	0.80	0.72	0.75	0.78

Table 25. smoking-offdevice-rgb

pid	precision	recall	f1 score	accuracy
P6	0.94	0.90	0.92	0.92
P7	0.93	0.89	0.91	0.91
P8	0.75	0.86	0.80	0.79
P9	0.96	0.95	0.96	0.96
P10	0.91	0.67	0.77	0.80
P11	0.82	0.78	0.80	0.80
P12	0.92	0.76	0.83	0.85
Mean	0.89	0.83	0.86	0.86

Table 26. smoking-offdevice-thermal

pid	precision	recall	f1 score	accuracy
P6	0.90	0.72	0.80	0.82
P7	0.97	0.93	0.95	0.95
P8	1.00	0.86	0.92	0.93
P9	0.87	0.76	0.81	0.82
P10	0.83	0.83	0.83	0.83
P11	0.81	0.77	0.79	0.80
P12	0.82	0.92	0.87	0.86
Mean	0.89	0.83	0.85	0.86

Table 27. smoking-ondevice-thermal

pid	precision	recall	f1 score	accuracy
P6	0.61	0.65	0.63	0.62
P7	0.79	0.94	0.86	0.84
P8	0.75	1.00	0.86	0.83
P9	0.89	0.39	0.54	0.67
P10	0.84	0.79	0.81	0.82
P11	0.80	0.82	0.81	0.81
P12	0.70	0.88	0.78	0.76
Mean	0.77	0.78	0.76	0.76

Table 28. smoking-offdevice-rgb-obfuscated-mask

pid	precision	recall	f1_score	accuracy
P6	0.80	0.61	0.70	0.73
P7	0.81	0.83	0.82	0.82
P8	0.33	0.14	0.20	0.43
P9	0.94	0.74	0.83	0.84
P10	0.78	0.38	0.51	0.64
P11	0.72	0.54	0.61	0.66
P12	0.72	0.76	0.74	0.73
Mean	0.73	0.57	0.63	0.70

Table 29. multiclass-offdevice-rgb-obfuscated-mask (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.69	0.68	0.68	0.68
P7	0.66	0.65	0.62	0.65
P8	0.50	0.52	0.46	0.52
P9	0.75	0.73	0.73	0.73
P10	0.61	0.60	0.60	0.60
P11	0.65	0.63	0.63	0.63
P12	0.70	0.70	0.69	0.70
Mean	0.65	0.65	0.63	0.64

Table 30. multiclass-offdevice-rgb-obfuscated-blur (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.84	0.84	0.84	0.84
P7	0.72	0.69	0.67	0.69
P8	0.52	0.52	0.52	0.52
P9	0.82	0.80	0.80	0.80
P10	0.72	0.72	0.72	0.72
P11	0.70	0.68	0.68	0.68
P12	0.79	0.78	0.78	0.78
Mean	0.73	0.72	0.71	0.72

Table 31. multiclass-offdevice-rgb (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.90	0.90	0.90	0.90
P7	0.79	0.75	0.75	0.75
P8	0.82	0.76	0.76	0.76
P9	0.91	0.91	0.91	0.91
P10	0.82	0.82	0.82	0.82
P11	0.80	0.80	0.80	0.80
P12	0.84	0.84	0.84	0.84
Mean	0.84	0.83	0.83	0.83

Table 32. multiclass-offdevice-thermal (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.79	0.78	0.78	0.78
P7	0.85	0.84	0.84	0.84
P8	0.86	0.86	0.85	0.86
P9	0.80	0.79	0.79	0.79
P10	0.76	0.75	0.75	0.75
P11	0.74	0.72	0.72	0.72
P12	0.86	0.85	0.85	0.85
Mean	0.81	0.80	0.80	0.80

Table 33. multiclass-ondevice-thermal (weighted-scores)

pid	precision	recall	f1_score	accuracy
P6	0.46	0.46	0.46	0.46
P7	0.65	0.65	0.61	0.65
P8	0.87	0.78	0.75	0.78
P9	0.62	0.51	0.48	0.51
P10	0.69	0.69	0.69	0.69
P11	0.64	0.63	0.62	0.63
P12	0.56	0.56	0.50	0.56
Mean	0.64	0.61	0.59	0.61