OrganicHAR: Towards Activity Discovery in Organic Settings for Privacy Preserving Sensors Using Efficient Video Analysis

PRASOON PATIDAR, Carnegie Mellon University, United States
RIKU ARAKAWA, Carnegie Mellon University, United States
RICARDO GRAÇA, Fraunhofer Portugal AICOS, Portugal
RÚBEN MOUTINHO, Fraunhofer Portugal AICOS, Portugal
ADRIANO SOARES, Fraunhofer Portugal AICOS, Portugal
ANA VASCONCELOS, Fraunhofer Portugal AICOS, Portugal
FILIPPO TALAMI, Fraunhofer Portugal AICOS, Portugal
JOANA COUTO DA SILVA, Fraunhofer Portugal AICOS, Portugal
INÊS SILVA, Fraunhofer Portugal AICOS, Portugal
CRISTINA MENDES SANTOS, Fraunhofer Portugal AICOS, Portugal
MAYANK GOEL, Carnegie Mellon University, United States
YUVRAJ AGARWAL, Carnegie Mellon University, USA

Deploying human activity recognition (HAR) at home is still rare because sensor signals vary wildly across houses, people, and time, essentially requiring in-situ data collection and training. Prior approaches use cameras to generate training labels for privacy-preserving sensors (LiDAR, RADAR, Thermal), but this forces sensors to detect predefined activities that cameras can see yet the sensors themselves cannot reliably distinguish. In this work, we introduce OrganicHAR, an activity discovery framework that inverts this relationship by placing sensor capabilities at the center of activity discovery. Our approach identifies naturally occurring signal patterns using privacy-preserving sensors, leverages Vision Language Models (VLMs) only during these key moments for scene understanding, and discovers discrete activity labels at granularities that these sensors can reliably detect. Our evaluation with 12 participants demonstrates OrganicHAR's effectiveness: it achieves 79% accuracy for coarse (4-5) activities using only basic ambient sensors (radar, lidar, thermal arrays), and 73% accuracy for fine-grained (8-9) activities when a wearable IMU, depth, and pose sensor are added. OrganicHAR maintains 77% accuracy on average across configurations while discovering 4-8 categories per user (15 across all users) tailored to each environment and sensor capabilities. By triggering video processing only at key moments identified by local sensors, we reduce queries to VLM by 90%, enabling practical and privacy-preserving activity recognition in natural settings.

Authors' Contact Information: Prasoon Patidar, Carnegie Mellon University, Pittsburgh, United States, prasoonpatidar@cmu.edu; Riku Arakawa, Carnegie Mellon University, Pittsburgh, United States, rarakawa@cs.cmu.edu; Ricardo Graça, Fraunhofer Portugal AICOS, Porto, Portugal, ricardo.graca@aicos.fraunhofer.pt; Rúben Moutinho, Fraunhofer Portugal AICOS, Porto, Portugal, ruben.moutinho@fraunhofer.pt; Adriano Soares, Fraunhofer Portugal AICOS, Porto, Portugal, adriano.soares@aicos.fraunhofer.pt; Ana Vasconcelos, Fraunhofer Portugal AICOS, Porto, Portugal, ana.vasconcelos@fraunhofer.pt; Filippo Talami, Fraunhofer Portugal AICOS, Porto, Portugal, filippo.talami@fraunhofer.pt; Joana Couto da Silva, Fraunhofer Portugal AICOS, Porto, Portugal, joana.couto@aicos.fraunhofer.pt; Inês Silva, Fraunhofer Portugal AICOS, Porto, Portugal, ines.silva@aicos.fraunhofer.pt; Cristina Mendes Santos, Fraunhofer Portugal AICOS, Porto, Portugal, cristina.santos@fraunhofer.pt; Mayank Goel, Carnegie Mellon University, Pittsburgh, United States, mayankgoel@cmu.edu; Yuvraj Agarwal, Carnegie Mellon University, Pittsburgh, PA, USA, yuvraj@cs.cmu.edu.

ACM 2474-9567/2025/12-ART203
https://doi.org/10.1145/3770674

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Collaborative interaction**; **Interactive systems and tools**; **Ambient intelligence**; • **Computing methodologies** → *Planning for deterministic actions.*

Additional Key Words and Phrases: Activity Recognition, Ambient Intelligence, Interactive Agents

ACM Reference Format:
Prasoon Patidar, Riku Arakawa, Ricardo Graça, Rúben Moutinho, Adriano Soares, Ana Vasconcelos, Filippo Talami, Joana Couto da Silva, Inês Silva, Cristina Mendes Santos, Mayank Goel, and Yuvraj Agarwal. 2025. OrganicHAR: Towards Activity Discovery in Organic Settings for Privacy Preserving Sensors Using Efficient Video Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 203 (December 2025), 32 pages. https://doi.org/10.1145/3770674

1 INTRODUCTION

Smart homes that can infer the activities of their occupants and offer them insights, automation, and assistance without privacy-invasive cameras or microphones have been a longstanding vision of the research community and companies [16, 17, 61]. Yet despite significant advances in Human Activity Recognition (HAR) techniques, their deployment remains largely unrealized in most households [65, 69]. While these HAR systems achieve impressive performance using benchmarks and in controlled settings, they falter in real home settings. This persistent disconnect stems not from technological limitations alone, but from a fundamental misalignment in how we have conceptualized the problem: we face the dual challenge of (i) deploying models in organic settings where users perform diverse, evolving activities in countless variations and (ii) simultaneously obtaining high-quality training labels to identify what activities are occurring in these complex, unstructured environments.

Most existing HAR approaches start with a set of activities to detect and classify. For instance, Patidar *et al.* [56] trained privacy-preserving sensors (*e.g.*, radar, lidar) to detect 17 common activities of daily living at home. While these approaches prove effective for short durations, the assumption that the same set of activities happens with similar consistencies across all environments breaks down in the long term, as users engage in everyday activities in their very own idiosyncratic ways [30, 65]. This reliance on pre-defined activity sets creates challenges in aligning with activities that matter to users and what activities these sensors can reliably detect in diverse environments. Rather than imposing a fixed set of activities across all environments, we need HAR systems that can autonomously discover what activities are reliably detectable given the specific sensors deployed in each unique setting. This discovery capability is crucial for practical adoption—users must understand their sensors' actual detection capabilities before making informed decisions about how they want to use these systems.

In this paper, we introduce OrganicHAR[1] (Figure 1), a novel framework that takes the first step towards this vision by discovering activities that naturally emerge from available sensing capabilities, rather than imposing predefined categories. We start by identifying potentially meaningful patterns (*i.e.*, recurring spatial patterns and temporal fluctuations) within sensor data. Next, we selectively leverage Vision Language Models [15] (VLMs) to understand what activities these patterns represent, and convert descriptions (in natural language) into discrete labels to train HAR models. Our approach offers two advantages critical for practical deployment: (1) it allows us to control the granularity of recognized activities for different environments and available sensing capabilities; and (2) it reduces computational overhead by processing only essential video segments during training.

Developing the OrganicHAR framework required solving two major technical challenges. First, we developed novel techniques to identify meaningful interaction moments from multimodal time-series sensor data, without requiring prior activity models. Second, to address the inconsistent activity descriptions generated by VLMs, we implemented a clustering method that transforms these variable VLM descriptions into consistent, sensor-appropriate activity labels—effectively connecting rich semantic understanding of VLMs with the limited capabilities of privacy-preserving sensors.

[1] https://github.com/synergylabs/OrganicHAR

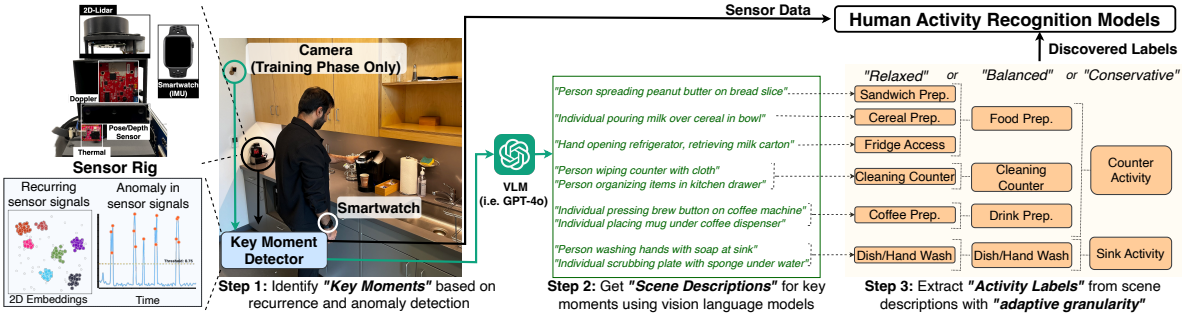


Fig. 1. The OrganicHAR framework discovers activity labels through a three-step sensor-first approach. **Step 1:** Privacy-preserving sensors identify "key moments" with recurring patterns (clusters) or anomalies (peaks). **Step 2:** A vision language model (VLM) generates natural language descriptions only for these key moments. **Step 3:** Descriptions are converted into activity labels with adaptive granularity tuned for downstream applications, following human-intuitive hierarchies where specific activities (*e.g.*, "Sandwich Prep") naturally group into broader categories ("Food Prep" → "Counter Activity") while maintaining semantic meaning. Unlike prior approaches that use predefined activities and continuous video monitoring (top right), OrganicHAR discovers activities organically from sensor capabilities while reducing video processing by 90%.

To build and evaluate the performance of OrganicHAR, we collected a comprehensive multimodal dataset (comprising a Doppler RADAR, a 2D LiDAR, a thermal array, a wearable IMU, pose and depth sensors) from 12 participants, who performed four different breakfast preparation tasks (preparing sandwiches, tea, coffee, and cereal). Each participant performed these tasks naturally without instructions, allowing us to capture organic activity patterns as they would occur in real homes. Our results show that OrganicHAR can reliably discover up to 4-8 labels per user (15 unique labels across 12 users) with 87% accuracy. Using these automatically discovered labels, we trained downstream activity recognition models for privacy-preserving sensors that achieved 70%-88% accuracy (F1-Score: 68%), with performance varying based on activity label granularity and sensor configuration. This performance is notable because it emerges from activity patterns discovered in an organic setting and is not limited to detecting manually defined activities. We further validated OrganicHAR's effectiveness in the wild, through deployment in 5 homes for 7 days each, where the system adapted to natural daily routines and diverse kitchen environments, achieving 74.2% accuracy while demonstrating effective learning over time. Moreover, our approach is computation and cost-efficient since it only sends approximately 10% of video data to VLMs in the cloud, which also reduces over time. These results highlight the potential for practical at-home HAR systems that require no human labor and minimize reliance on privacy-invasive sensors. In summary, we make the following contributions:

- OrganicHAR, a framework that reorients HAR by discovering activities from sensor patterns rather than predefined categories, minimizing privacy concerns while enabling practical deployment in real homes without continuous video monitoring. The source code is publicly available.
- Evaluation of OrganicHAR, which demonstrated that it achieves 70%-88% accuracy for different hardware configurations while automatically discovering up to 4-8 activities per user (15 unique across 12 users) without manual annotation. Our real-world deployment in 5 homes further validates the practical effectiveness of OrganicHAR, showing 74.2% accuracy across natural daily routines.
- Multimodal dataset comprising controlled evaluation data from 12 participants performing breakfast preparation tasks across three distinct kitchen environments and real-world deployment data from 5 homes over 7 days each (11 hours), captured through six sensor modalities [59]. We hope this dataset enables further research in the space of HAR in unconstrained settings.

2 BACKGROUND AND MOTIVATION

Ultimately, our goal is to develop context-aware systems that assist people in their homes—enabling aging in place, supporting complex daily activities, and promoting safety. In deploying such systems, cameras and vision-based models are powerful, but continuous monitoring violates privacy. On the other hand, sensors, such as RADAR, Thermal sensor, and IMU, protect privacy but require extensive manual labeling of training data that users cannot realistically provide.

We have worked toward this vision through a series of projects, each solving certain challenges while revealing others. VAX [56] demonstrated that privacy-preserving sensors (RADAR, LIDAR, Thermal sensor) could achieve 85% accuracy after bootstrapping from camera data during the training phase, eliminating continuous camera-based monitoring. However, like several prior studies, VAX uses a rigid taxonomy of predefined activities and fails to capture real-world diversity, with unrecognized behaviors relegated to an "Other" category. Then, as one of the applications meaningful to users, PrISM [4–7] was developed as a framework to support procedural tasks, such as cooking and self-care, with a context-aware, mixed-initiative assistant that combines HAR and dialogue interaction. While demonstrating effectiveness in certain scenarios, the PrISM assistant often faces challenges in HAR; people improvise, combine tasks, and create their own routines—variations that predefined activity models cannot anticipate. These prior projects have motivated this work: we need to discover what activities (1) occur in an environment ; (2) can be sensed by the system; and (3) are beneficial to be tracked for the user.

Hence, we built OrganicHAR, which inverts the traditional paradigm by letting sensor capabilities drive activity discovery. OrganicHAR's "sensor-first" approach enables each deployment to develop its own activity taxonomy matched to both the available sensing resources and how individuals actually behave. The convergence of these three technologies—VAX's privacy-preserving sensing, PrISM's procedural assistance, and OrganicHAR's organic discovery—now enables systems that learn individual household patterns, monitor for safety concerns, and provide assistance tailored to each person's unique routines. We are currently working with people living with Dementia [10] and post-operative skin cancer patients [26, 67] to provide daily task assistance and safety monitoring. In our deployments, the system learns each person's unique patterns, monitors for safety concerns like forgotten appliances (*e.g.*, stove left on), and provides task assistance tailored to their specific routines and cognitive abilities. By integrating bootstrapped sensing, adaptive interaction, and organic activity discovery, we move closer to systems that adapt to people's actual needs and routines instead of requiring people to adapt to predefined technological constraints.

3 RELATED WORK

We begin with previous work using privacy-preserving sensors for recognizing human activities. Then, we review machine-learning methods for addressing the challenges of training HAR models in new environments. Finally, we focus on prior work for in-situ training to clarify the novelty of our approach.

3.1 Human Activity Recognition (HAR) with Privacy-preserving Sensors

Human Activity Recognition (HAR) has been extensively studied for its promising applications. Video-based HAR has traditionally dominated [55], benefiting from large-scale datasets [19, 28] and advanced understanding toolboxes [48, 49]. VLMs have recently revolutionized this field, enabling zero-shot activity recognition [45, 62]. However, deploying video-based systems in real homes faces significant challenges: substantial computational resources (often gigabytes of GPU memory) and serious privacy concerns from continuous monitoring [1, 80]. Researchers have investigated alternatives such as privacy-preserving ambient sensors—Doppler radars [3, 13], lidars [36], low-resolution thermal arrays, subsampled audio sensing [50], and environmental sensors [37]—and wearable devices with IMUs [14, 75, 82]. Wang *et al.* provide a comprehensive review of multi-modal sensor fusion

techniques [71]. Researchers have also explored multi-modal sensing [47], sensor fusion [2, 52], and generative training approaches [43, 74] to enhance accuracy.

Most existing HAR approaches rely on supervised learning with predefined activity labels, requiring designers and developers to anticipate activities before data collection. This creates challenges in selecting the appropriate granularity: too fine-grained (*e.g.*, distinguishing between chopping different vegetables) reduces accuracy due to subtle signal differences and increased complexity; too coarse-grained (*e.g.*, simply detecting "cooking") diminishes utility by lacking meaningful and actionable task insights. These challenges highlight the need for approaches that *discover* and recognize activities autonomously at appropriate granularity. Notably, while clustering-based methods can identify patterns without predefined labels [8, 21], they require manual interpretation to assign meaningful labels [29, 73], limiting practical applications like monitoring specific home events. OrganicHAR addresses this gap by streamlining the process through advances in vision and language foundational models.

3.2 Transfer Learning and Domain Adaptation

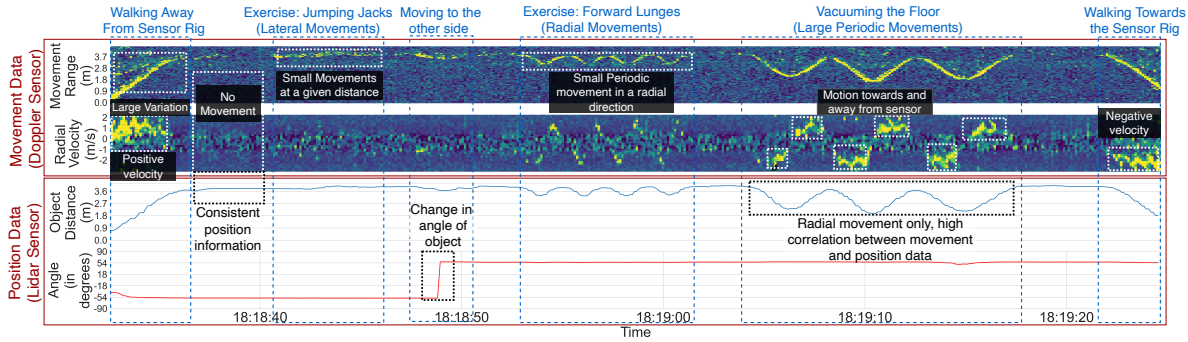
Transfer learning and domain adaptation techniques address HAR training challenges with privacy-preserving sensors by leveraging knowledge from data-rich but privacy-invasive modalities (video and raw audio) to generate synthetic training data for privacy-preserving sensors. Researchers have developed methods converting audio/video/images to synthetic IMU signals [35, 42, 78], video to doppler sensor readings [3, 24], or IMU signals to doppler sensor readings [13]. This paradigm has strengthened with advances in foundational models [39, 40], with researchers using VLMs as zero-shot or few-shot learners for HAR [9, 41, 62]. These approaches require continuous video data access and substantial computational resources, making them impractical where privacy, cost, and latency matter. Self-supervised learning offers another direction to minimize reliance on labeled data [23, 31, 60], but still requires predefined activity sets and struggles to discover new activities autonomously. Unlike approaches that transfer knowledge into predefined activity sets, our work explores a novel bootstrapping framework focusing on the signal representation of each sensor. We use targeted video analysis of key moments to build robust models that subsequently operate using only privacy-preserving sensors, leveraging VLMs' understanding capabilities while maintaining privacy and practical deployability in real-world settings.

3.3 Approaches for In-situ Training

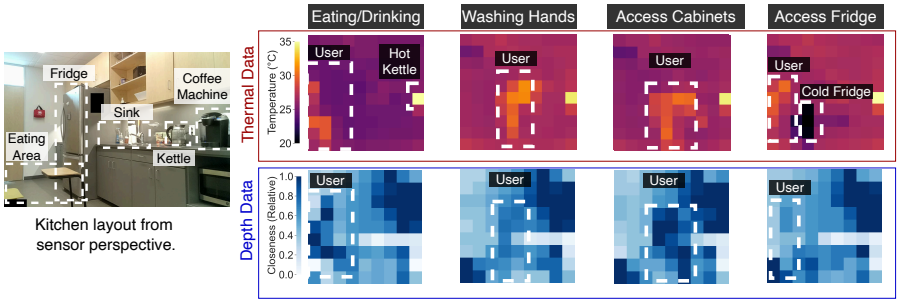
In-situ training approaches aim to bootstrap HAR systems to new environments while minimizing data collection and annotation burden. Clustering-based methods group similar sensor patterns, and then request user labels for representative samples from each cluster [29, 73]. While reducing annotation burden, users must still manually provide labels, and systems remain limited to predefined activities. For instance, some systems temporarily deploy cameras to capture ground truth labels for training models that operate on privacy-preserving sensors alone, but these typically struggle with complex or unknown activities [48, 56]. Other approaches explore interactive learning strategies that gradually refine activity models based on user feedback [34]. These often require significant user involvement or struggle to capture activity diversity. Unlike prior approaches, OrganicHAR uses privacy-preserving sensors to identify key moments for targeted video analysis rather than requiring continuous video recording or extensive user annotation. By leveraging VLMs' scene understanding capabilities during these brief moments, we bootstrap location awareness and rich activity recognition without predefined labels. This enables autonomous activity discovery while minimizing privacy concerns and user burden.

4 HARDWARE DESIGN

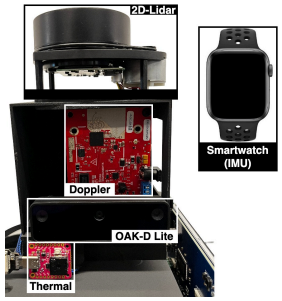
Our system implements a modular sensing infrastructure that can be configured with different combinations of sensors based on application requirements and user privacy preferences. We categorize our sensors into three tiers: (1) a basic ambient sensing setup (*Ambient (Basic) Only*) ambient sensors including position sensing with 2D



(a) Movement and position information using Doppler and Lidar sensors over time.



(b) Thermal data (10x8) and Depth Array information at a single timestamp (snapshot).



(c) Hardware setup used for data collection.

Fig. 2. Visualizing information across various activities from various privacy-preserving sensors inspired by our prior work on VAX [56]. (a) shows the movement range and radial velocity of moving objects using a Doppler sensor and object distance and angle using a Lidar sensor in the sensor plane across a sequence of activities happening in the living room. (b) shows snapshots of the Thermal and Depth Array at a given timestamp for various kitchen activities. For depth visualization, pixels grow darker when the user is closer to the sensor. (c) shows our hardware rig used to collect sensor data.

lidar, movement sensing with a doppler radar, and infrared sensing using low-resolution (10x10) thermal arrays; (2) an intermediate sensing setup (*Ambient (Basic) + Wearable (IMU)*) combining ambient sensors with wearable motion (IMU) data from smartwatches; and (3) an advanced setup (*Ambient (Advanced) + Wearable (IMU)*) that includes on-device human pose and depth estimation. Our current prototype integrates these sensing modalities (See Figure 2c) through a small form factor PC (Intel NUC, 8-core, 16GB RAM) that handles all processing locally. For the initial training phase only, we use an iPhone for video capture to enable VLM-based scene understanding.

Basic Ambient Sensing: Our basic configuration focuses on non-optical sensors that capture coarse movement and presence information. This includes: (i) *FMCW Doppler Radar* using a 77GHz mmWave radar (AWR1642BOOST-ODS) operating at 5Hz that captures movement through RF reflection, providing velocity and range information for detecting significant movements (Figure 2a), (ii) *2D Lidar* that measures distance using Time of Flight (ToF) or laser beam parallax. We utilize a Slamtec RPLIDAR A1M8, which delivers 360-degree horizontal measurements at 6-8Hz with 1-degree angular resolution. This data identifies spatial occupancy patterns and environmental changes (Figure 2a), and (iii) *Low-Resolution Thermal Camera* that provides a 10x10 pixel thermal map at 8Hz.

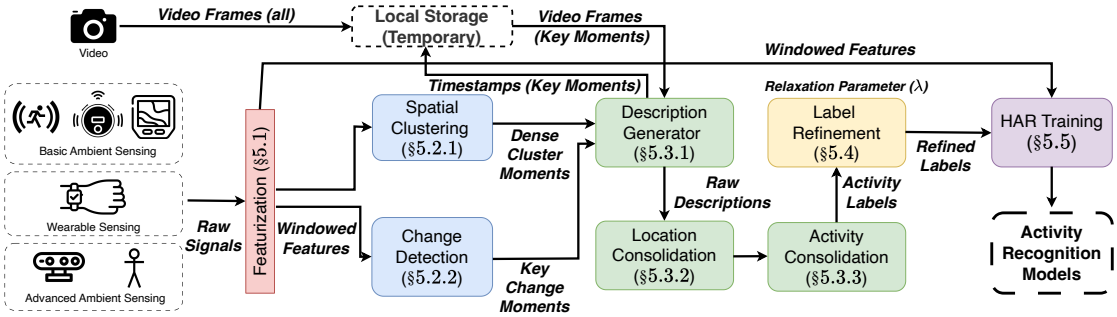


Fig. 3. Overall architecture of OrganicHAR. Raw sensor signals from different hardware configurations (§4) are featurized for each sensor separately (§5.1), and used to identify *key moments* (§5.2) for targeted video processing using VLMs. The raw descriptions are then converted to discrete activity labels using LLMs (§5.3), and further refined using a relaxation parameter λ (§5.4). The final activity labels are used to train activity recognition models (§5.5) for deployment.

While the original sensor (FLIR Lepton 3.5) offers higher resolution, we reduce it to preserve privacy [56] while still capturing thermal signatures associated with different activities, such as using appliances or engaging in tasks that have thermal signatures (*e.g.*, heating kettle, accessing a fridge; see Figure 2b).

Wearable Sensing: We also support configurations where a user-worn wearable device is utilized, enabling the capture of personalized motion data using *Smartwatch IMU*, and a custom iOS app to capture detailed motion data from Apple Watch. The IMU provides acceleration, gyroscope, and magnetometer data at 50Hz, enabling recognition of hand movements and gestures that complement the ambient sensing data. This sensor captures personal movement data and does not record any environmental information.

Advanced Ambient Sensing: Our advanced configuration adds on-device processing capabilities through the OAK-D Lite platform [58]. It includes: (i) *On-Device Pose Sensing* that outputs normalized pose coordinates (640x480 frame) from device, and (ii) *Low-Resolution Depth Maps* using stereo vision capabilities on OAK-D Lite (Figure 2b). This low-resolution depth information provides valuable context about spatial relationships while making it extremely hard to reconstruct any part of the original images [32].

Each sensing tier presents distinct privacy-capability tradeoffs. The basic ambient configuration maximizes privacy as its sensors cannot capture personally identifiable information, requiring no user interaction after installation. The wearable configuration enables personalized motion tracking but requires the device to be worn consistently. The advanced configuration provides detailed user information through processed pose and depth information, though optical sensors may raise privacy concerns despite no raw image output, making it better suited for common areas. Notably, future hardware-limited sensors that constrain functionality at the hardware level [25, 51, 72] could potentially address these concerns. As we demonstrate later, OrganicHAR's multiple configuration options enable deployments that balance privacy preferences with recognition capabilities based on individual needs and spatial contexts.

5 SYSTEM DESIGN

Figure 3 presents the overall architecture of our OrganicHAR framework, highlighting its major components. OrganicHAR operates through a sensor-first approach where privacy-preserving sensors drive activity discovery. Raw sensor signals from different hardware configurations (§4) are first featurized for each modality separately (§5.1). These features are then used to identify *key moments* (§5.2) that warrant targeted video processing

using VLMs. The resulting raw descriptions from VLMs are converted to discrete activity labels using LLM-based clustering (§5.3) and further refined using a relaxation parameter λ (§5.4) to control semantic granularity. Finally, these discovered activity labels are used to train activity recognition models (§5.5) that operate solely on privacy-preserving sensors during deployment, eliminating the need for continuous video monitoring.

5.1 Featurization

Each sensor modality produces unique signal patterns with inherent capabilities and limitations, requiring tailored featurization methods to maximize their strengths. OrganicHAR employs a sliding window approach with fixed-length 5-second segments and 0.5-second stride length for both training and inference. For feature extraction at timestamp t_s, we compute features using sensor data from interval $[t_s - 5, t_s]$ seconds. This window size is consistent with recent sensor fusion approaches [12, 66] to capture sufficient temporal context while maintaining computational efficiency. During final activity recognition, we aggregate consecutive predictions from overlapping windows to determine activity boundaries and durations, accommodating both brief interactions and extended tasks. This multimodal approach achieves complementary fusion, which enables our system to identify scene-relevant moments and activity patterns across different environments and sensor configurations more robustly than any single modality. We provide sensor-specific featurization details in Appendix A.1.

5.2 Key Moments Identification

Once we have featurized data, we start by identifying *key moments* for each modality. This critical module determines which segments of sensor data warrant further analysis by the VLM. The module ingests featurized sensor streams and processes them through two complementary approaches: (1) a spatial clustering module that identifies recurring patterns representing consistent activities in the feature space, and (2) a temporal change detection module that identifies segments with notable signal shifts indicating activity transitions.

5.2.1 Spatial Clustering. We employ clustering in the feature space of each sensor modality to identify recurring patterns representing distinct activity signatures. While we refer to this as "spatial clustering," it operates not in the physical space but in the high-dimensional feature space specific to each sensor, where sensor signals from similar activities form natural clusters regardless of their physical location. We leverage HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [46], selected for its ability to discover clusters of varying densities without requiring a predefined cluster count, crucial for unsupervised activity discovery in unfamiliar environments. After applying standard preprocessing (robust scaling and dimensionality reduction using principal component analysis), we optimize the clustering process through careful hyperparameter tuning. The key hyperparameters governing our feature space clustering include:

- min_cluster_size $\in \{m_1, m_2, \ldots, m_a\}$: Determines the minimum points required to form a cluster, directly influencing the granularity of detected activity patterns.
- min_samples $\in \{s_1, s_2, \ldots, s_b\}$: Controls clustering conservativeness, with higher values producing more stringent cluster formation.
- n_components $\in \{c_1, c_2, \ldots, c_c\}$: Defines the dimensionality of the feature space after reduction.

Central to our approach is a carefully designed scoring function that evaluates clustering quality across all sensor modalities for a given combination of hyperparameters:

$$\text{score} = w_1 \cdot S_{count} + w_2 \cdot S_{noise} + w_3 \cdot S_{prob} \tag{1}$$

This function balances three essential aspects of effective activity clustering for our use case:

- S_{count}: Evaluates how well the number of discovered clusters aligns with activity diversity (between C_{min} and C_{max} clusters), penalizing overly granular and overly coarse solutions.

- S_{noise}: Assesses the proportion of data points assigned to meaningful clusters rather than noise, preferring solutions with noise ratios below N_{max}. It separates random sensor activations from meaningful patterns.
- S_{prob}: Captures the average cluster membership probability, reflecting the confidence in cluster assignments and favoring well-defined, distinct activity signatures.

This scoring framework consistently evaluates cluster quality across diverse sensing modalities while accommodating their varying signal characteristics. The weighted combination balances cluster count appropriateness (w_1) with noise handling and assignment confidence (w_2 and w_3). From each cluster, we select representative windows with high membership confidence as key moments for VLM analysis, significantly reducing the required video processing requests while capturing essential activity patterns. Our approach excels at activity discovery by identifying natural signal groupings that correspond to distinct behaviors — cooking activities might cluster in Doppler feature space despite different kitchen locations, while refrigerator interactions might form clusters in thermal feature space despite varying user movements. Through empirical evaluation across sensors, we selected hyperparameter ranges that provide optimal clustering: min_cluster_size $\in \{3-8\}$, min_samples $\in \{2-5\}$, and n_components $\in \{8-80\}$, with modality-specific adjustments (*e.g.*, thermal sensing uses cluster_selection_epsilon $\in \{0.01-0.03\}$ for tighter clusters). The scoring weights ($w_1 = 0.3-0.4, w_2 = 0.15-0.3, w_3 = 0.1-0.3$) and target parameters (desired clusters between 20-40, maximum noise ratio of 0.8) are consistent across modalities.

5.2.2 Change Detection. We also leverage the temporal changes in signals of each modality, expecting that these moments can capture useful actions. We used an algorithm based on multimodal anomaly detection, specifically the one proposed by Yamanishi *et al.* [76]. This algorithm is online, unsupervised outlier detection using the Gaussian Mixture Model (GMM), which can capture changes in time-series signals without prior knowledge. Whenever a new data sample is received, it computes an anomaly score based on its likelihood under the current GMM and simultaneously updates the model parameters. Specifically, let $\mathbf{y}_t \in \mathbb{R}^M$ denote the input sample at time t, and let $\pi_{i,t}$, $\mu_{i,t}$, and $\Sigma_{i,t}$ represent the weight, mean vector, and covariance matrix of the ith component (for $i = 1, \ldots, K$) of the GMM at time t. The anomaly score for $\mathbf{y}_t$ is then defined as

$$s_t = -\ln\left(\sum_{i=1}^{K} \pi_{i,t-1}\, \mathcal{N}\left(\mathbf{y}_t \mid \mu_{i,t-1}, \Sigma_{i,t-1}\right)\right). \tag{2}$$

Following this, the parameters of the GMM are updated according to

$$\begin{aligned}
\lambda_{i,t} &= \frac{\pi_{i,t-1}\, \mathcal{N}\left(\mathbf{y}_t \mid \mu_{i,t-1}, \Sigma_{i,t-1}\right)}{\sum_{j=1}^{K} \pi_{j,t-1}\, \mathcal{N}\left(\mathbf{y}_t \mid \mu_{j,t-1}, \Sigma_{j,t-1}\right)}, \\
\pi_{i,t} &= (1-\alpha)\, \pi_{i,t-1} + \alpha\, \lambda_{i,t}, \\
\tilde{\mu}_{i,t} &= (1-\alpha)\, \tilde{\mu}_{i,t-1} + \alpha\, \lambda_{i,t}\, \mathbf{y}_t, \\
\mu_{i,t} &= \frac{\tilde{\mu}_{i,t}}{\pi_{i,t}}, \\
\tilde{\Sigma}_{i,t} &= (1-\alpha)\, \tilde{\Sigma}_{i,t-1} + \alpha\, \lambda_{i,t}\, \mathbf{y}_t \mathbf{y}_t{}^T, \\
\Sigma_{i,t} &= \frac{\tilde{\Sigma}_{i,t}}{\pi_{i,t}} - \mu_{i,t}\mu_{i,t}^T,
\end{aligned} \tag{3}$$

where α is a forgetting factor that controls the influence of past observations. After computing the anomaly score for each t using Equation 2, we selected the top N points as detected moments. The parameters K, M and α are hyperparameters of this component. We empirically set $K = 2$, $M = 10$, and $\alpha = 0.1$, which works well on average across all modalities.

5.3 Label Discovery

After identifying *key moments* through spatial clustering and temporal change detection, our framework translates these sensor-identified moments into meaningful activity labels using a multi-stage approach described below:

5.3.1 Semantic Description Generation. For each *key moment*, we prompt a VLM to generate semantic descriptions from video frames—the only stage where video data is processed, occurring exclusively during initial training. We iteratively tested multiple configurations to determine optimal video input parameters. Higher frame rates and resolutions frequently led to hallucinations and inconsistent descriptions. Our empirical testing revealed that low-resolution frames (640×480) sampled at 1 FPS for 5-second clips provided the ideal balance between detail and consistency for current (as of July 2025) VLM capabilities, though this balance may shift as vision-language models improve.

A critical challenge in prompt design was achieving dual objectives: controlling VLM speculation while balancing specificity and consistency. VLMs tend to infer intentions beyond what is directly observable and struggle with appropriate granularity—either overgeneralizing activities (*e.g.*, labeling everything as "kitchen activity") or focusing on irrelevant details. We addressed these tensions by explicitly instructing the model to *"focus only on what you can clearly see"* while providing structured categories guiding appropriate detail levels. We directed the VLM to analyze four aspects: (1) *actions*, including movements between locations; (2) *objects* the person interacts with; (3) *location* where activity occurs; and (4) *activity structure* (initial conditions, main actions, results). For each category, we provided examples at appropriate granularity levels to establish consistent reference points. By requiring JSON-structured output with self-assessed confidence scores, we enabled systematic filtering of observations ($\theta_{conf} = 0.8$), rejecting both speculative and overly generic descriptions. While accurate, the resulting descriptions exhibited inconsistent terminology and granularity across similar activities—addressed in our label consolidation process. We provide our complete prompt in Appendix A.2.

5.3.2 Location Consolidation. We implemented a two-step label consolidation process: location consolidation followed by activity description consolidation. The location-based approach works well with privacy-preserving ambient sensors, which often provide clean separation in signals for spatial differentiation. Additionally, the VLM outputs were more consistent for location references than for action descriptions, providing a stable foundation for generating location-specific activity labels. We design an LLM prompt to consolidate various location references from VLM outputs (*e.g.*, "at sink," "by sink," "near sink basin") into consistent *functional zones* like "sink area," "counter area," and "coffee machine area." The prompt instructs the LLM to group locations based on supported activities, merge functionally similar spaces, and maintain separation between distinct activity zones. This spatial organization helps disambiguate semantically similar actions through their context (*e.g.*, distinguishing "washing hands" from "washing dishes" based on precise sink location) and creates a foundation for more accurate activity recognition that leverages the spatial detection capabilities of our sensors.

5.3.3 Activity Consolidation. With established functional zones, we transform unstructured action descriptions from the VLM into discrete activity labels. For each functional zone, we use an LLM prompt to create high-level clusters based on action descriptions from this zone. The prompt differentiates between actions (physical movements like "pouring" or "washing") and purpose/context (the goal or situation, such as "breakfast preparation" or "dishwashing"). This distinction helps maintain important separations – for example, keeping "pouring cereal" and "pouring coffee" as different activities despite sharing the same physical action. In the sink area, "washing dish with sponge" and "scrubbing plate with sponge" merge into "washing dishes with sponge," while remaining distinct from "washing hands with soap."

We then use LLM-assisted matching to assign each activity description to existing clusters based on interaction patterns. For each description, we extract action and object information, then use an LLM to calculate weighted similarity scores based on action alignment ($w_a = 60$), object consistency ($w_o = 25$), and location match ($w_l = 15$).

These empirically determined weights reflect the relative importance of each factor in determining activity similarity. The mechanism effectively consolidates variations like "filling cup using coffee machine" and "using coffee machine with mug" into "preparing coffee drink" while distinguishing them from functionally different activities occurring at the same location.

5.4 Label Refinement and Semantic Granularity Control

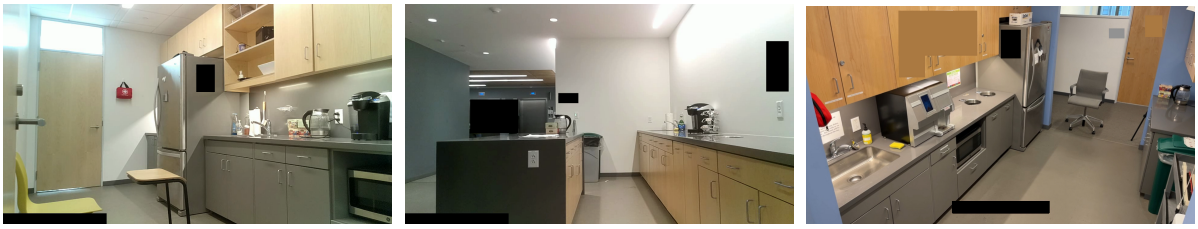
While our two-step consolidation produces well-defined activity labels, we also need to provide users control over label merging at different granularities. From a human perspective, certain distinctions are more meaningful than others—"rinsing dishes" and "washing dishes with sponge" should merge before "washing hands with soap" and "washing dishes" despite textual similarity. We analyze semantic relationships across six dimensions. We use an LLM to expand each activity label along: (1) action type (*e.g.*, "cleaning," "preparing"), (2) objects involved (*e.g.*, "dishes," "sponge," "water"), (3) sub-location (*e.g.*, "sink basin," "counter edge"), (4) purpose/goal (*e.g.*, "remove food residue," "prepare beverage"), (5) access patterns (*e.g.*, "reaching for scrubber," "turning faucet"), and (6) related activities (*e.g.*, "rinsing dishes" relates to "drying dishes"). This multi-dimensional characterization provides rich semantic representation. Next, we compute a pairwise similarity matrix S where the score S_{ij} between activity labels a_i and a_j is calculated by combining weighted cosine similarities across these dimensions:

$$\begin{aligned} S_{ij} = w_{\text{action}} \cdot \text{sim}_{\text{action}}(a_i, a_j) + w_{\text{object}} \cdot \text{sim}_{\text{object}}(a_i, a_j) + w_{\text{location}} \cdot \text{sim}_{\text{location}}(a_i, a_j) + \\ w_{\text{purpose}} \cdot \text{sim}_{\text{purpose}}(a_i, a_j) + w_{\text{access}} \cdot \text{sim}_{\text{access}}(a_i, a_j) + w_{\text{relation}} \cdot \text{sim}_{\text{relation}}(a_i, a_j) \end{aligned} \tag{4}$$

For each dimension, we generate textual embeddings using an embedding model and compute cosine similarities between them. The weights reflect each dimension's contribution to human-relevant distinction: action type ($w_{\text{action}} = 0.20$), object involvement ($w_{\text{object}} = 0.25$), sub-location ($w_{\text{location}} = 0.15$), purpose/goal ($w_{\text{purpose}} = 0.15$), interaction patterns ($w_{\text{access}} = 0.15$), and explicit relationships ($w_{\text{relation}} = 0.10$). Finally, we apply hierarchical clustering with a relaxation parameter λ to group activity labels into clusters C such that $\forall a_i, a_j \in c_k, S_{ij} \geq 1 - \lambda$. The values of λ are typically in the range from 0 (one cluster per activity) to 1 (merge all activities in a single cluster). This gives users control over recognition granularity through a single parameter—smaller λ values preserve fine distinctions between activity labels, while larger values merge semantically related activities.

5.5 Training Privacy-preserving Human Activity Recognition Models

The final phase trains HAR models on privacy-preserving sensors without video after initial training. For each sensor modality, we evaluate multiple classifiers robust to imbalanced datasets, implementing leave-one-session-out cross-validation to assess generalizability. This addresses real-world sensor challenges including missing values and class imbalance across different environments. After training individual classifiers, we employ grid search to identify suitable sensor-classifier combinations for each functional zone. We evaluate two ensemble methods: soft voting (combining probability distributions from multiple classifiers) and hard voting (weighted voting using each classifier's highest-confidence prediction) [11]. Our framework computes balanced accuracy (macro-recall) for each configuration to determine effective deployment combinations. Ultimately, after training, OrganicHAR implements a hierarchical approach towards inferring activities: first identifying the user's functional zone, then applying appropriate zone-specific activity recognition models using the selected sensor ensemble. This context-aware structure recognizes that activities present different signatures in different zones while managing computational resources. By activating only necessary sensors and models, the system balances privacy preservation with recognition accuracy while operating on local sensors after training.



(a) Kitchen 1 (b) Kitchen 2 (c) Kitchen 3

Fig. 4. Kitchen environments used in our study: (left) Kitchen 1 with compact galley layout, (middle) Kitchen 2 with island counter, and (right) Kitchen 3 with integrated appliances.

5.6 Detailed Implementation

OrganicHAR is implemented in Python with approximately 8,000 lines of code and consists of three main components: (i) key moment identification from privacy-preserving sensors, (ii) semantic label discovery using VLMs, and (iii) training privacy-preserving sensor models with discovered activity labels. For sensor data processing, we utilize NumPy [27] for array manipulation, Pandas [54] for time-series analysis, and OpenCV [22] for video processing. The key moment identification implements Gaussian Mixture Models from scikit-learn [57] for temporal change detection and HDBSCAN [46] for spatial clustering. During the training phase only, we leverage OpenAI's API [53] with the *GPT-4o* model to generate semantic descriptions for the key moments, which are then organized into hierarchical activity clusters using SciPy's [68] squareform and linkage functions. For model training, we employ multiple classifier implementations from scikit-learn and imbalanced-learn [38], including RUSBoost [64], Balanced Random Forest [77], EasyEnsemble [44], KNN [18], and SVM [18]. We open-sourced the OrganicHAR implementation [59], which can be used with different combinations of privacy-preserving sensors with minimal implementation changes.

6 DATA COLLECTION

We collected data from 12 participants in three distinct kitchen environments. Participants performed four breakfast preparation tasks: *preparing tea*, *making coffee from a coffee machine*, *preparing cereal*, and *making sandwiches* with various spreads. Critically, participants received no explicit instructions on how to execute these tasks, and researchers did not intervene or provide prompts during the activities. Each participant approached the tasks in their own way, using natural movements and sequences that reflected genuine everyday behavior rather than scripted demonstrations. The data collection for each user lasted approximately 1.5 hours, with participants performing each task 2-3 times with natural variations. To ensure clear segmentation while maintaining ecological validity, participants washed their hands before each task and used a hand clap to mark segment boundaries.

We collected data across three kitchenettes (Figure 4): **Kitchen 1 (P1-P4)** featured a compact galley-style layout with a single countertop, refrigerator at one end, sink in the middle, and coffee machine at the opposite end; **Kitchen 2 (P5-P8)** had an open layout with an island counter separate from the main preparation area, featuring a sink and coffee machine; and **Kitchen 3 (P9-P12)** was a medium-sized kitchenette with microwave, coffee machine, and sink aligned along a single wall, with a refrigerator perpendicular to the main counter.

Our dataset comprises ~4300 five-second snippets (approximately 6 hours) of breakfast preparation activities across six sensor modalities: doppler radar, 2D lidar, low-resolution thermal array, wearable IMU, on-device pose sensing, and depth processing. Unlike existing HAR datasets featuring scripted activities, our dataset captures natural variations in how people perform everyday tasks. Ground truth annotation was performed by

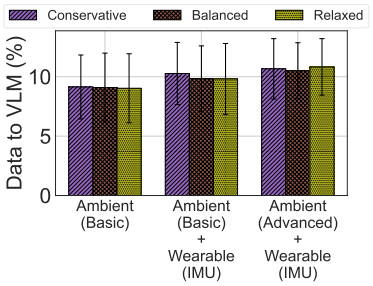


Fig. 5. Percentage of video data requiring VLM analysis across configurations. Our approach processes only 9-11% of total video data, demonstrating efficiency compared to continuous monitoring.

Granularity	Metrics	Sensor Config		
		Ambient (Basic) Only	Ambient (Basic)+ Wearable (IMU)	Ambient (Advanced)+ Wearable (IMU)
Conservative	Accuracy	90.4%±8.9%	91.1%±6.7%	91.7%±6.9%
	F1 Score	89.4%±9.8%	90.6%±6.4%	90.3%±7.4%
Balanced	Accuracy	89.9%±6.3%	85.5%±7.1%	87.1%±7.9%
	F1 Score	84.1%±8.7%	78.0%±9.9%	81.1%±11.8%
Relaxed	Accuracy	86.7%±4.9%	83.8%±6.5%	85.9%±6.7%
	F1 Score	73.2%±10.9%	72.4%±8.7%	75.6%±9.9%

Table 1. Average accuracy and F1 scores (mean±std) of discovered activity labels compared to ground truth across three semantic granularity settings. Conservative ($\lambda = 0.4$) represents coarse-grained activities, Balanced ($\lambda = 0.3$) shows medium granularity, and Relaxed ($\lambda = 0.2$) captures fine-grained activities. Higher performance is observed with more advanced sensor configurations, particularly for fine-grained recognition.

an unbiased observer who labeled each snippet without prior knowledge of system-generated labels. Our dataset is open-sourced [59], including all sensor data, VLM annotations, and ground truth labels.

7 EVALUATION

Our evaluation of OrganicHAR focuses on two primary research questions:

- **RQ1:** How effective and efficient is our activity discovery pipeline in identifying meaningful activity labels compared to traditional approaches?
- **RQ2:** How do the HAR models trained on these discovered labels perform across different sensor configurations and semantic granularity settings?
- **RQ3:** How does OrganicHAR perform in real-world deployment scenarios, and how does its performance evolve over time?

For each user and sensor configuration, we evaluated OrganicHAR across three semantic granularity settings, *i.e.*, (*Conservative*: $\lambda = 0.4$, *Balanced*: $\lambda = 0.3$, and *Relaxed*: $\lambda = 0.2$) to determine how permissively activities are merged based into single activity label based on their semantic similarity, as discussed in §5.4.

7.1 Performance of Label Discovery Pipeline

7.1.1 Efficiency of Key Moment Identification. One of the primary advantages of our sensor-first approach is minimizing the amount of video data that requires processing during the training phase. Figure 5 shows the percentage of video data selected for VLM analysis across different sensor configurations and granularity settings. The error bars show variability (standard deviation) across different users. Our key moment identification approach demonstrates remarkable efficiency, requiring analysis of only approximately 9-11% of the total collected video data across all configurations. The *Ambient(Basic) Only* configuration required the least video processing (around 9%), while the advanced configuration with *Ambient(Advanced)+Wearable(IMU)* utilized slightly more (around 10.5%). This modest increase is expected as richer sensing capabilities can identify more nuanced activity transitions that warrant VLM analysis.

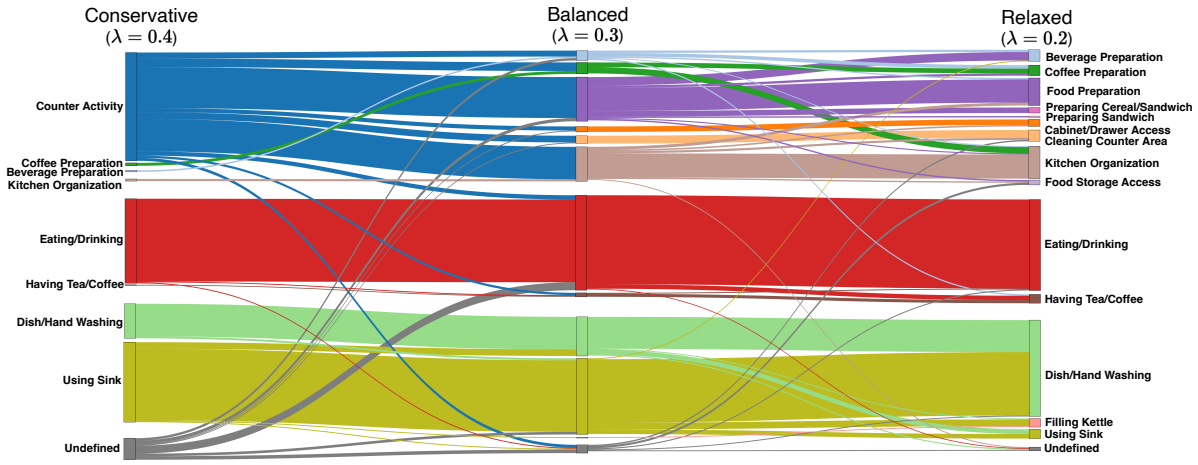


Fig. 6. Discovered activity labels across three semantic granularity settings: Conservative ($\lambda = 0.4$, left), Balanced ($\lambda = 0.3$, middle), and Relaxed ($\lambda = 0.2$, right). The flow width represents the proportion of data with each label, showing how broader categories branch into more fine-grained activities as granularity increases. For example, the "Counter Activity" broad category (left) differentiates into "Food Preparation," "Cleaning Counter Area," and other activities in the Balanced setting.

7.1.2 Discovered Activity with Different Granularity. Figure 6 illustrates the discovered activity labels across granularity settings, revealing how broader categories in the Conservative setting (left) branch into fine-grained distinctions in the Balanced (middle) and Relaxed (right) settings. The Conservative setting identified four primary activities: "Counter Activity," "Using Sink," "Dish/Hand Washing," and "Eating/Drinking," capturing kitchen zones and behaviors at a high level. Some users showed more nuanced counter activities, though with low discovery rates. In the Balanced setting, "Counter Activity" branches into specific categories like "Food Preparation," "Cleaning Counter Area," and "Cabinet/Drawer Access," discovering activity hierarchies matching the task groupings. The Relaxed setting continues this differentiation, breaking "Counter Activity" into fine-grained parts, with "Food Preparation" splitting into "Preparing Cereal/Sandwich" and other specialized tasks.

We observe several patterns through this granularity analysis. For instance, specific activities like "Eating/-Drinking" remain stable across all settings, having distinctive sensor signatures consistently recognized regardless of granularity. "Using Sink" appears broadly in Conservative mode but becomes less prominent in Relaxed settings as the system refines it to "Dish/Hand Washing" specifically, demonstrating how our approach clarifies ambiguous activities as similarity thresholds change. Flow patterns follow logical parent-child relationships rather than arbitrary recombination—"Cabinet/Drawer Access" emerges specifically from "Counter Activity," not unrelated categories. Overall, our system discovered 15 distinct activity labels, with users typically having 4-8 relevant activities depending on settings and sensor setup. This result suggests that OrganicHAR can adapt to what each sensor configuration reliably detects while maintaining meaningful relationships matching human understanding of kitchen activities, offering clear advantages over predefined activity sets.

7.1.3 Accuracy of the Discovered Activity. We evaluated our discovered activity labels against ground truth annotations as shown in Table 1. Since discovered labels often use different terminology than human-annotated ground truth, we performed manual mapping to convert annotations into the discovered activity label space for comparison. The Conservative setting (coarse-grained activities) demonstrates high accuracy above 90% and strong F1 scores across all sensor configurations. Performance moderately declines with increasing granularity, yet remains robust with accuracy above 83% even in the most challenging scenarios. The *Ambient(Basic)*

Table 2. Average accuracy and F1 scores (mean±std) of the HAR models across sensor configurations and granularity settings. These models perform effectively despite being trained on organically discovered labels rather than predetermined categories. The Balanced setting ($\lambda = 0.3$) offers an optimal compromise between recognition granularity and performance for most deployments, while Conservative settings excel in accuracy (84-88%) when coarser activity recognition suffices. Relaxed setting results (shown in gray) represent exploratory findings illustrating current technical boundaries rather than recommended configurations.

Granularity	Metrics	Sensor Config		
		Ambient(Basic) Only	Ambient(Basic)+ Wearable(IMU)	Ambient(Advanced)+ Wearable(IMU)
Conservative	**Accuracy**	83.9%±10.1%	85.1%±6.5%	87.8%±8.4%
	F1 Score	82.9%±10.6%	82.5%±9.3%	86.8%±9.3%
Balanced	**Accuracy**	75.2%±9.2%	72.5%±6.6%	78.8%±8.9%
	F1 Score	65.5%±13.3%	63.7%±9.7%	70.6%±11.1%
Relaxed	**Accuracy**	70.4%±11.5%	69.0%±9.0%	73.2%±10.4%
	F1 Score	51.4%±10.3%	47.8%±10.5%	55.6%±14.8%

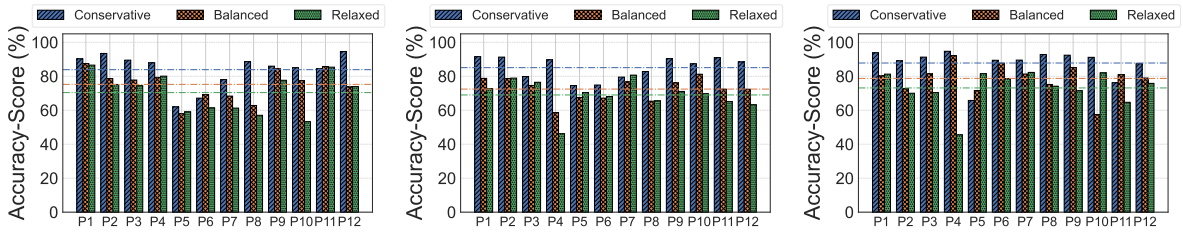
Only configuration sometimes outperforms *Ambient(Basic)+Wearable(IMU)* in finer-grained settings, suggesting wearable data occasionally introduces variability when distinguishing semantically similar activities. The *Ambient(Advanced)+Wearable(IMU)* configuration delivers the strongest Relaxed setting performance (85.9% accuracy), highlighting depth and pose information's value for fine-grained recognition. The more pronounced F1 score decline (72-75%) in the Relaxed setting reflects inherent challenges in precisely identifying fine-grained activities with VLMs processing low-frame-rate data (1 FPS), camera angle limitations, and partial occlusions. Given these results, we recommend the Balanced setting ($\lambda = 0.3$) as the optimal compromise for most deployments, offering meaningful activity differentiation while maintaining strong accuracy (85-89%) and F1 scores (78-84%). The high average performance demonstrates that OrganicHAR can automatically identify meaningful activity labels aligning well with human observations.

Moreover, to validate robustness across VLM architectures, we tested three state-of-the-art models (GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4). Our 1 FPS, 640×480 configuration remains optimal across all VLMs, though detection rates vary significantly (50.2-92%). Consistent fine-grained recognition limitations across all models (F1 scores of 35.6-55.4%) suggest these challenges reflect current VLM capabilities generally. Detailed analysis is presented in Appendix A.3.

7.2 Performance of Human Activity Recognition Models

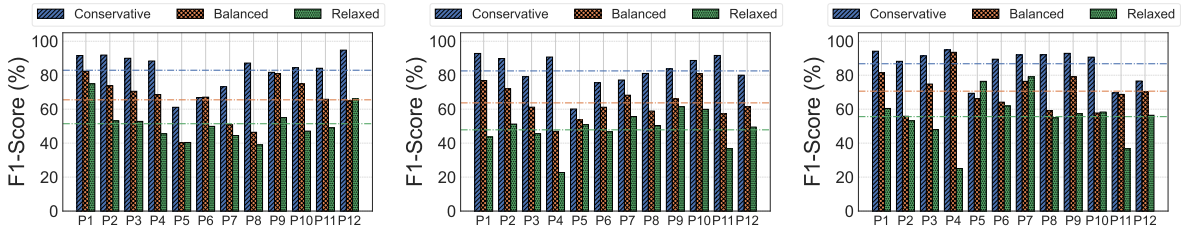
In this section, we examine how HAR models trained on our discovered labels perform across different sensor configurations. Regarding this, the label discovery evaluation in the above revealed low F1-scores (72-75%) in the Relaxed setting, indicating imperfect alignment between some discovered fine-grained labels and ground truth. This misalignment would affect HAR models trained on these labels, where the system attempts to distinguish between highly similar activities. This informs us that the Conservative and Balanced settings ($\lambda = 0.4$ and $\lambda = 0.3$) represent more realistic deployment scenarios for OrganicHAR with current sensing capabilities. In other words, the Relaxed setting results ($\lambda = 0.2$) should be interpreted as exploratory findings that illustrate current technical boundaries rather than definitive performance benchmarks.

7.2.1 Overall Performance. Table 2 shows our trained HAR models' performance across sensor configurations. The HAR models demonstrate robust performance with 70-88% accuracy and 48-87% F1 scores, varying by



(a) Ambient(Basic) Only (b) Ambient(Basic)+Wearable(IMU) (c) Ambient(Advanced)+Wearable(IMU)

Fig. 7. Per-participant accuracy across three sensing configurations. Kitchen 1 participants (P1-P4) show consistently high accuracy even with basic ambient sensing, Kitchen 2 participants (P5-P8) show greater degradation with finer granularity, and Kitchen 3 participants (P9-P12) benefit most from wearable IMU. Horizontal dashed lines indicate average performance per granularity setting.

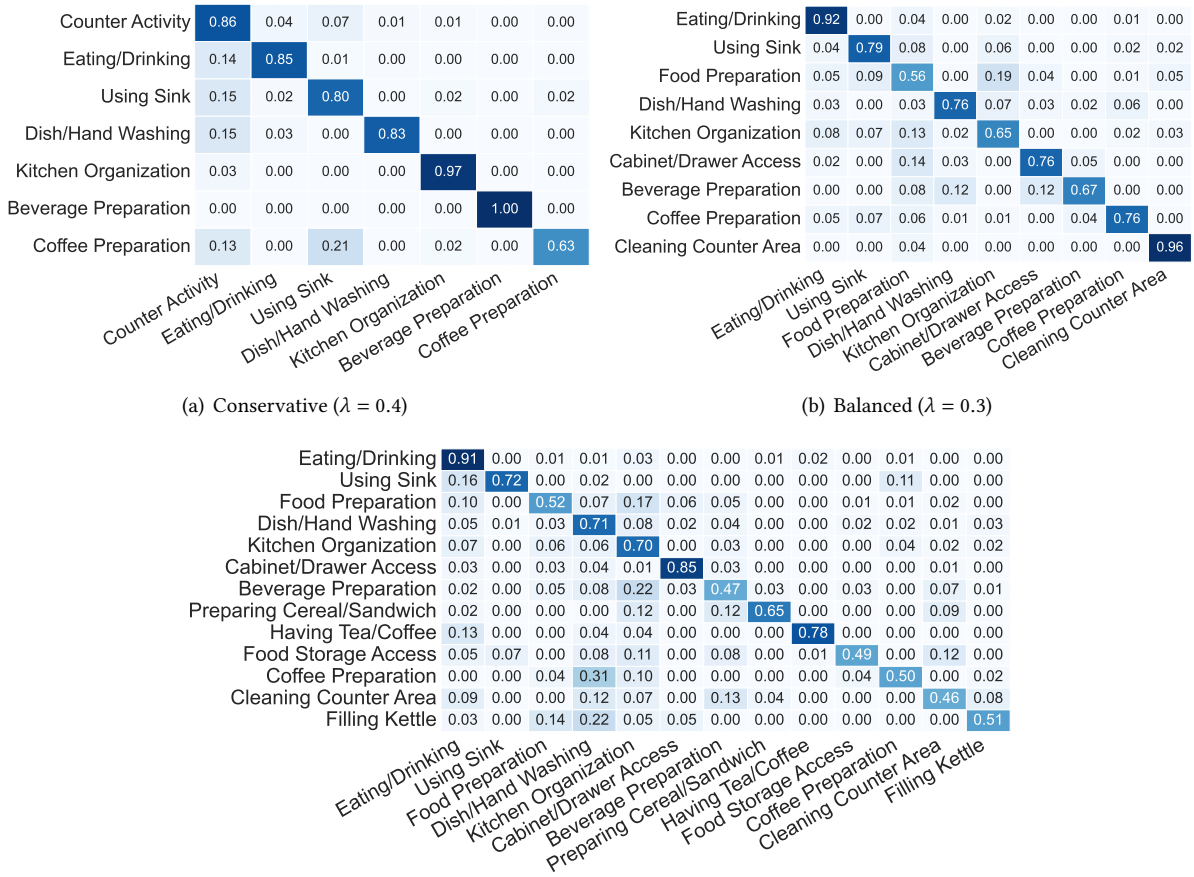


(a) Ambient(Basic) Only (b) Ambient(Basic)+Wearable(IMU) (c) Ambient(Advanced)+Wearable(IMU)

Fig. 8. Per-participant F1 scores across sensing configurations, revealing sharper performance drops than accuracy metrics with increasing granularity. F1 scores for the Relaxed setting (green) show particularly large variance between participants and kitchens, illustrating the impact of environmental layout on the precision-recall balance in fine-grained activity recognition.

sensor configuration and semantic granularity. Performance decreases with increasing granularity: Conservative ($\lambda = 0.4$) achieves 84-88% accuracy and 83-87% F1 scores across all configurations, Balanced ($\lambda = 0.3$) reaches 72-79% accuracy and 64-71% F1 scores, and Relaxed ($\lambda = 0.2$) attains 69-73% accuracy and 48-56% F1 scores. The *Ambient(Advanced)+Wearable(IMU)* configuration consistently outperforms others, particularly for fine-grained activities, achieving 73% accuracy and 56% F1 score even in the challenging Relaxed setting. Additional spatial awareness from pose estimation and depth sensing helps resolve ambiguities between semantically similar activities. Interestingly, *Ambient(Basic) Only* occasionally outperforms *Ambient(Basic)+Wearable(IMU)*, indicating wearable data can introduce noise when movements aren't consistently captured.

7.2.2 User-level Performance Variability. Figures 7 and 8 show accuracy and F1 scores across participants for three sensor configurations and granularity levels. Participants P1-P4 (Kitchen 1) maintain high performance across granularity settings with basic ambient sensing, while P5-P8 (Kitchen 2) show significant degradation with increased granularity. This suggests environment layout impacts recognition—Kitchen 1's constrained layout enables more reliable activity differentiation than Kitchen 2's open layout. For P9-P12 (Kitchen 3), wearable IMU data notably improves performance, providing valuable complementary information in certain environmental configurations. F1 scores prove more sensitive to sensor configuration and granularity due to label imbalance,



(a) Conservative ($\lambda = 0.4$)

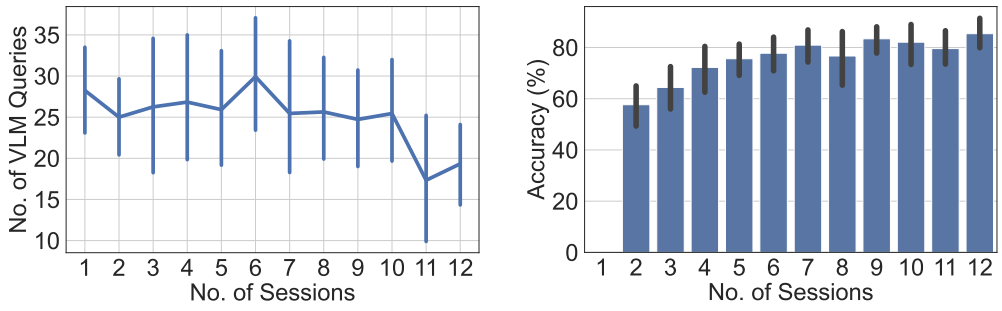
(b) Balanced ($\lambda = 0.3$)

(c) Relaxed ($\lambda = 0.2$)

Fig. 9. Confusion matrices showing the HAR performance using basic ambient sensors across three granularity settings. As granularity increases from Conservative (a) to Relaxed (c), more fine-grained activities emerge, revealing specific confusion patterns between semantically or spatially related activities.

dropping 31-35 percentage points from Conservative to Relaxed settings versus 13-16 points for accuracy. This steeper decline highlights challenges in maintaining precision and recall when distinguishing semantically similar activities with limited sensing. The *Ambient(Advanced)+Wearable(IMU)* configuration shows a smaller Conservative-to-Relaxed performance gap (15% accuracy difference) than other configurations, indicating richer sensing enables fine-grained recognition without sacrificing coarse-grained performance. These results suggest that OrganicHAR achieves robust performance across various hardware configurations by adapting recognition granularity to available sensing capabilities rather than forcing predetermined labels,

7.2.3 Activity Confusion Analysis. Figure 9 shows confusion matrices for the *Ambient(Basic) Only*configuration across three granularity settings. The Conservative setting (Figure 9a) achieves high accuracy with minimal cross-category confusion. "Counter Activity" confuses with other activities as it encompasses multiple fine-grained



(a) VLM queries required across sessions

(b) Recognition accuracy across sessions

Fig. 10. Incremental training analysis of OrganicHAR: (a) Count of VLM queries after incorporating new training session. (b) Recognition accuracy measured on all future sessions. Error bars represent 95% confidence intervals across users.

activities; "Coffee Preparation" confuses with "Using Sink" due to spatial proximity. The Balanced setting (Figure 9b) shows increased confusion with finer distinctions. "Food Preparation" has the lowest accuracy, confusing with "Kitchen Organization" and "Using Sink" due to similar ambient signatures. "Beverage Preparation" significantly confuses with cabinet access and washing activities since preparation incorporates these actions. In the Relaxed setting (Figure 9c), "Beverage Preparation" degrades further, confusing primarily with "Kitchen Organization" as ambient sensors cannot distinguish fine manipulations in similar locations. Kettle-related and washing activities show substantial confusion, illustrating difficulty differentiating tasks with similar postures and locations using only ambient sensors. Despite increasing confusion at finer granularities, "Eating/Drinking" maintains excellent performance across all settings with distinctive ambient signatures. "Cabinet/Drawer Access" similarly maintains high accuracy through distinctive movement patterns captured by basic ambient sensors.

These patterns validate our sensor-first approach—OrganicHAR adapts activity granularity based on reliable detection capabilities rather than forcing potentially indistinguishable predetermined categories. This analysis identifies which sensor configurations suit specific recognition goals, enabling informed privacy-capability tradeoffs (Appendix A.4 covers other configurations).

7.2.4 Incremental Training Analysis. To understand how cloud VLM query requirements and recognition accuracy evolve with increasing training data, we conducted an incremental training analysis simulating sequential session processing. We used *Ambient(Advanced)+Wearable(IMU)* as it has the highest VLM usage among configurations. With each new session, OrganicHAR identifies interesting segments across all data, applies the current model, invokes VLM analysis for low-confidence predictions or novel patterns, and retrains with the expanded dataset. Figure 10a shows total VLM queries required after incorporating each session, including queries for new data and previously-seen data that becomes significant due to pattern repetition across sessions. Query counts remain stable (25-30 per session) for the first 9 sessions despite the growing dataset, as the system continuously reassesses historical data alongside new observations, sometimes identifying patterns in earlier sessions not initially flagged when the dataset was smaller. The downward trend in sessions 10-11 (17-19 queries) is notable, though extended data collection is needed to confirm this pattern. Figure 10b shows recognition accuracy (measured on sessions not included in training) improving from 57.6% after the first session to 85.4% after 11 sessions, with most gains in early sessions. These results demonstrate promising trends: steady accuracy improvement coupled with

(a) Home-1 (b) Home-2 (c) Home-3 (d) Home-4 (e) Home-5

Fig. 11. Kitchen environments used in real-world home deployments: (Home-1) compact galley layout captured from overhead diagonal perspective, (Home-2) medium-sized kitchen with distinct functional zones, (Home-3) narrow galley configuration from elevated angle, (Home-4) spacious kitchen with multiple countertops from horizontal perspective, and (Home-5) linear wall-mounted kitchen with integrated appliances from horizontal viewpoint.

potential VLM query reduction in later sessions, suggesting OrganicHAR could become both more effective and computationally efficient over time.

7.3 Real-world Deployment Evaluation

We extended our evaluation to examine OrganicHAR's performance in actual home environments, where users engage in daily routines without experimental constraints. We deployed OrganicHAR in 5 homes for 7 days, with participants collecting 6-8 sessions each based on their kitchen usage frequency. Participants used their kitchens normally—cooking, cleaning, snacking, socializing—without task restrictions or researcher guidance, capturing the complexity of real environments including activity interruptions and varying lighting conditions (See Figure 11). The deployment generated approximately 11 hours of sensor data across diverse home layouts and activity patterns. For accuracy evaluation, we sampled 250 five-second segments per participant using temporal stratified random sampling: 60 from early sessions (1-2), 60 from middle sessions (3-5), and 130 from late sessions (6-8), enabling analysis of learning progression and mature system performance. We annotated ground truth using our established methodology (§6), mapping labels to discovered activities. Based on controlled evaluation results, we used the Balanced granularity setting ($\lambda = 0.3$) and *Ambient(Advanced)+Wearable(IMU)* sensor configuration for all real-world deployment analyses.

7.3.1 Overall Performance. Figure 12 presents the overall recognition accuracy across homes using leave-one-session-out cross-validation. OrganicHAR achieves an average accuracy of 74.2%±14.8% across the five homes, with performance ranging from 60% to 95%. This represents a moderate decrease compared to the controlled evaluation results (78.8%±8.9% for the same sensor configuration and granularity setting), reflecting the additional complexity of unstructured home environments. Performance for individual homes varied significantly, with Home-1 achieving the highest accuracy at 95%, followed by Home-2 at 85%. Home-3, Home-4, and Home-5 achieved lower performance at 69%, 63%, and 60% respectively. This variability appears partially related to both kitchen layout constraints and camera positioning challenges (See Figure 11). Home-4 and Home-5 utilized more lateral camera placements due to their linear kitchen configurations, which limited the VLM's ability to capture comprehensive activity context compared to the elevated diagonal perspectives used in Home-1 and Home-2. Additionally, Home-4's consistently dim lighting conditions throughout most sessions further degraded VLM performance during key moment analysis. These real-world constraints, which are difficult to replicate in controlled settings, highlight the importance of strategic camera positioning and adequate lighting for effective VLM-based activity discovery.

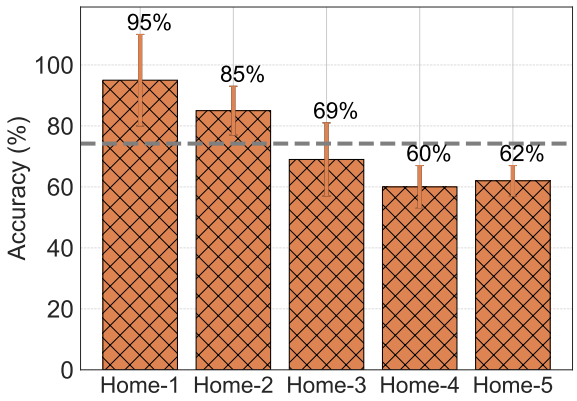


Fig. 12. Overall recognition accuracy in real-world home deployments using leave-one-session-out cross-validation with Balanced granularity setting and *Ambient(Advanced)+Wearable(IMU)* configuration. Each bar represents average performance when training on all other available sessions (including future ones) and testing on one held-out session. The horizontal dashed line indicates average accuracy (74.2%) across all participants.

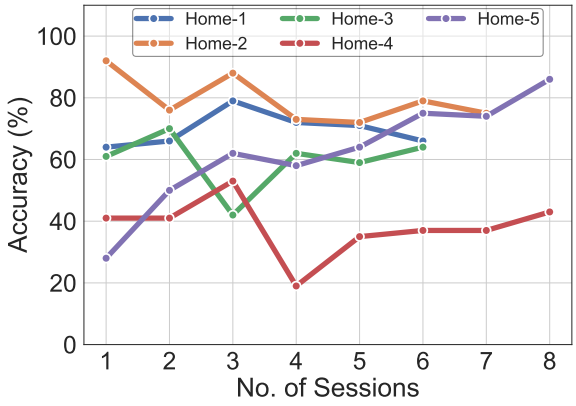


Fig. 13. Incremental training trajectories showing forward-looking evaluation in chronological session order, where the system trains on sessions 1 through N and tests on all remaining future sessions (N+1 onwards). *Note:* Performance patterns differ from Figure 12 due to temporal dependencies, varying amounts of training/test data, and chronological activity pattern evolution—reflecting realistic deployment conditions where future data is unavailable.

7.3.2 Discovered Activity Patterns. Across all 5 homes, OrganicHAR identified 4 common kitchen activities: "stovetop cooking", "refrigerator interaction", "sink area tasks", and "countertop food preparation". These represent behaviors that occur typically across these kitchen environments. Beyond these commonly occurring behaviors, we discovered activities specific to usage in each home. Home-4 had "coffee preparation" routines with patterns for coffee station tasks and coffee machine operation. Home-5 showed "rice cooker usage" with separate patterns for preparation and serving phases, along with "dishwasher interaction" patterns. Home-3 was the only home with "waste disposal" detected as a standalone recurring activity. Home-2 demonstrated granular activity detection, with "refrigerator door interactions" separate from "general refrigerator access". This dual-layer discovery—common activities plus individual specific patterns—shows that users can expect baseline functionality while the system adapts to their specific routines.

7.3.3 Incremental Training Analysis. Figure 13 shows learning trajectories using forward-looking evaluation in chronological session order, where accuracy represents the system's performance on all future sessions after training exclusively on sessions 1 through N. The performance patterns may differ from Figure 12 due to several temporal factors: (1) the amount of training data increases with each session while test data decreases, and (2) activity patterns evolve differently, making early sessions good/poor predictors of later behavior based on what kind of activities is performed in later sessions. While computationally expensive (and cost-prohibitive) combinatorial approaches (training on all possible session combinations) could provide alternative insights, this chronological evaluation reflects actual deployment scenarios. The results demonstrate varied adaptation across participants. Home-5 showed substantial improvement (28% to 86% accuracy), indicating successful learning of evolving activity patterns. Home-2 maintained consistently high performance (92% to 72%), reflecting stable routines that enable reliable prediction. Home-1 achieved stable moderate performance (64-79% range). Home-3 and Home-4 exhibited variable patterns, suggesting the system requires additional training or user assistance when participants have irregular routines. These trajectories indicate that OrganicHAR adapts effectively to

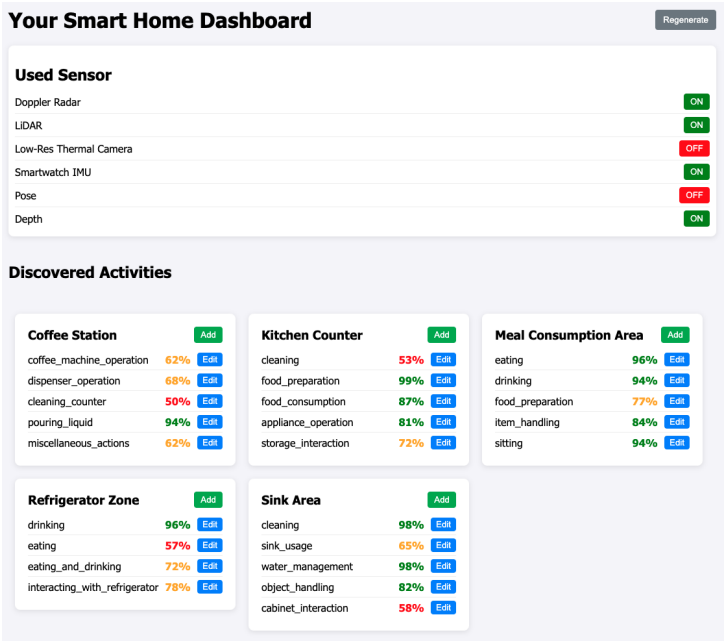


Fig. 14. Interface for customizing the sensors to be used and activity labels. The percentage(%) values show how well a deployed sensing configuration can predict activities when compared with discovered activity labels.

individual activity patterns, with performance stability correlating to routine consistency—regular kitchen usage enables more stable recognition, while variable patterns present ongoing adaptation challenges.

8 DISCUSSION

8.1 User Story: From Initial Adoption to Customization

OrganicHAR is designed to incrementally adapt to users and their environments while preserving privacy. First, by continuing data collection and analysis, we can surface representative examples of detectable activities for each environment, allowing users to form realistic expectations before deciding to deploy our system. The set of activities that we can typically detect in each location, *e.g.*, a kitchen, can be told to the user to manage their expectations (see §7.3.2). Upon installation, the system begins by identifying activities that are reliably detectable in the user's home using only privacy-preserving sensors. As demonstrated in our deployment study, even with a few days of data, OrganicHAR can detect 4–5 coarse activity categories using ambient sensors alone, and up to 8–9 fine-grained categories with additional modalities like IMU, depth, and pose.

After this initial discovery phase, users can interact with our prototype customization interface (Figure 14), which allows them to merge similar activity labels (*e.g.*, "washing dishes", "rinsing hands") into broader categories like "sink-related tasks", or define entirely new labels they want OrganicHAR to detect. While this interface has not yet been formally evaluated, it is technically feasible by reusing components from our label clustering and refinement pipeline. Our contribution is not the interface itself, but the underlying paradigm shift, *i.e.*, OrganicHAR surfaces candidate activity labels that are grounded in what the sensors can actually detect in each environment, which can then be reconciled with each user's preferences and goals. Investigating how users make sense of, reconfigure, or reject such sensor-driven candidates represents a promising direction for future HCI work.

Notably, OrganicHAR does not require users to install all sensors we tested. The framework is designed to be modality-agnostic, which allows flexibility to any combination of available sensors without algorithmic changes. Users can select sensor combinations that balance privacy requirements, budget constraints, and recognition needs. The three-tier configuration framework tested in our evaluation can inform such cost considerations.

8.2 Applications in Healthcare and Assistive Technologies

As we described in the Background (§2), OrganicHAR is particularly promising for healthcare applications, where continuous monitoring can support individuals with special needs while respecting privacy. For people with dementia and their caregivers, the system can notify household members when critical actions are forgotten (*e.g.*, leaving the refrigerator open) or summarize high-level activities for clinicians without invasive video monitoring [10, 20]. Procedural task assistance is another example we are applying OrganicHAR to, for instance, to support wound care procedure for post-operative skin cancer patients [5, 67]. Unlike existing approaches constrained by manually-predefined step sequences, OrganicHAR could make procedural tracking more flexible by automatically identifying meaningful steps based on sensor capabilities rather than forcing predetermined sequences. The hierarchical nature of our discovered activities (from coarse to fine-grained) enables adaptive assistance based on user needs, from general awareness of functional zone usage for those requiring minimal support, to detailed step-by-step guidance for those needing more comprehensive assistance.

8.3 Potential Privacy Concerns and Mitigation Strategies

While OrganicHAR uses privacy-preserving sensors during deployment, using VLMs during training can raise privacy concerns about transmitting domestic visual information to third-party cloud services. While OrganicHAR processes only 9-11% of video data during key moments rather than continuous monitoring, users may worry about data misuse, unintended inferences about their private lives, and potential access by service providers or authorities [33, 79, 81]. Several techniques can mitigate these concerns without compromising labeling effectiveness. For example, since our framework focuses on actions, locations, and object interactions rather than personal identification, automatic face blurring, background anonymization, and removal of personally identifiable information can be applied. Additionally, user agency can be enhanced by providing transparent previews of what video segments would be transmitted to cloud services and enabling users to approve or decline each training session before any data is sent. Similarly, researchers have improved handling of sensitive content through targeted fine-tuning of VLMs [63]. We will also explore emerging approaches such as federated learning and edge-based VLMs to minimize cloud dependencies while maintaining semantic understanding capabilities.

8.4 Limitations and Future Work

8.4.1 Towards In-the-Wild Deployment. Our real-world deployment revealed practical challenges: camera positioning constraints in linear kitchen layouts limited VLM effectiveness, varying lighting conditions affected recognition consistency, and participants' evolving routines created temporal adaptation challenges requiring ongoing refinement. Future work explores ways to improve OrganicHAR by considering these barriers, which will offer practical deployment guideline.

8.4.2 Beyond Discrete Activity Labels. Given challenges maintaining consistent activity labels across heterogeneous users and environments, we can explore alternatives to discrete classification. One approach involves direct translations between sensor streams and natural language descriptions. Rather than forcing classification between "washing dishes" or "washing hands," the system could generate descriptions like "using the sink with small movements, likely washing hands." This direction becomes increasingly feasible as small language models become more capable and deployable on edge devices [70]. While these compact models cannot perform complex reasoning, they could strategically interpret sensor patterns and generate contextual descriptions locally.

8.4.3 Improving Fine-Grained Activity Recognition. Our evaluation reveals insights about current VLM capabilities and their downstream effects on HAR. While Conservative and Balanced settings achieve strong performance (72-89% accuracy), the Relaxed setting (F1 scores of 48-72%) explores current boundaries of fine-grained activity recognition. Since VLM labeling quality influences training data for privacy-preserving HAR models, this performance partially reflects VLM capabilities in distinguishing subtle activity differences. Analysis across three state-of-the-art VLMs—GPT-4.1, Gemini 2.5, and Claude Sonnet 4 (See Appendix A.3)—shows consistent patterns, representing general characteristics of current VLM technology rather than framework-specific limitations. This cascading relationship positions OrganicHAR's sensor-first architecture to benefit from advances in VLM capabilities, particularly improved spatiotemporal reasoning and fine-grained visual understanding.

8.4.4 Hybrid Approaches: Combining Supervised and Discovery Methods. Our clustering approach faces inherent tradeoffs: users must wait for sufficient data collection before meaningful clusters emerge (the "cold start" problem), and initial recognition performance remains lower until adequate training data accumulates. These limitations suggest promising hybrid approaches where supervised and discovery methods play complementary roles. Our real-world deployment revealed that certain activities occur consistently across diverse home environments, while others remain highly environment-specific. This observation suggests that supervised models could provide robust baseline recognition for common activities that generalize across homes, while unsupervised discovery captures the personalized routines that make each home unique. Future work should explore domain adaptation techniques that identify environment-invariant features for transferable activities while preserving the flexibility to adapt to environment-specific variations, potentially combining supervised representation learning with our sensor-first discovery paradigm.

9 CONCLUSION

This work introduces OrganicHAR, a framework for at-home human activity recognition (HAR) that removes the reliance on pre-defined labels and continuous video processing by leveraging local, privacy-preserving sensors to identify key moments for targeted video bootstrapping and activity discovery. This approach minimizes computational overhead and cloud dependency while maintaining robust recognition accuracy across varying levels of sensor granularity, as demonstrated in our evaluations across diverse kitchen environments. By integrating multimodal sensor data with VLMs during an efficient training phase, our hierarchical bootstrapping technique successfully translates rich semantic information into discrete, sensor-compatible activity labels. Ultimately, our results validate the feasibility of scalable, privacy-aware HAR systems that can adapt to different sensing configurations, paving the way for more efficient and user-centric smart environments.

Acknowledgments

This work was partially supported by NSF Awards SaTC-1801472 and CSR-1526237, the Translational Fellowship from CMU's Center for Machine Learning and Health (CMLH), and the Presidential Fellowship from CMU's CyLab Security and Privacy Institute. We gratefully acknowledge the generous gift from Trane Technologies supporting smart buildings research at Carnegie Mellon University. We extend our sincere thanks to the Longitudinal Prize for Dementia Foundation and the Lived Experience Advisory Panel (LEAP) for their valuable feedback on the overall project direction. We are deeply grateful to our Autonomous team for their assistance with data collection and their constructive feedback throughout the multiple iterations of the hardware development. We would also like to thank Ben Weinshel, Anu Sitaraman, Haozhe Zhou, Vimal Mollyn, and Shreya Bali for their insightful comments on the paper and invaluable input on the system design throughout this research. Finally, we express our appreciation to the anonymous reviewers for their thoughtful and constructive feedback that significantly improved this work.

References

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. 2019. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, 451–466. https://www.usenix.org/conference/soups2019/presentation/abdi

[2] Antonio A Aguileta, Ramon F Brena, Oscar Mayora, Erik Molino-Minero-Re, and Luis A Trejo. 2019. Multi-sensor fusion for activity recognition—A survey. *Sensors* 19, 17 (2019), 3808.

[3] Karan Ahuja, Yue Jiang, Mayank Goel, and Chris Harrison. 2021. Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 292, 10 pages. https://doi.org/10.1145/3411764.3445138

[4] Riku Arakawa, Jill Fain Lehman, and Mayank Goel. 2024. PrISM-Q&A: Step-Aware Voice Assistant on a Smartwatch Enabled by Multimodal Procedure Tracking and Large Language Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4 (Nov. 2024), 180:1–180:26. https://doi.org/10.1145/3699759

[5] Riku Arakawa, Prasoon Patidar, Will Page, Jill Lehman, and Mayank Goel. 2025. Scaling Context-Aware Task Assistants that Learn from Demonstration and Adapt through Mixed-Initiative Dialogue. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 145, 19 pages. https://doi.org/10.1145/3746059.3747700

[6] Riku Arakawa, Hiromu Yakura, and Mayank Goel. 2024. PrISM-Observer: Intervention Agent to Help Users Perform Everyday Procedures Sensed using a Smartwatch. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3654777.3676350

[7] Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. 2023. PrISM-Tracker: A Framework for Multimodal Procedure Tracking Using Wearable Sensors and State Transition Information with User-Driven Handling of Errors and Uncertainty. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4 (Jan. 2023), 156:1–156:27. https://doi.org/10.1145/3569504

[8] Paola Ariza Colpas, Enrico Vicario, Emiro De-La-Hoz-Franco, Marlon Pineres-Melo, Ana Oviedo-Carrascal, and Fulvio Patara. 2020. Unsupervised Human Activity Recognition Using the Clustering Approach: A Review. *Sensors* 20, 9 (Jan. 2020), 2702. https://doi.org/10.3390/s20092702 Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[9] Luca Arrotta, Claudio Bettini, Gabriele Civitarese, and Michele Fiori. 2024. ContextGPT: Infusing LLMs Knowledge into Neuro-Symbolic Activity Recognition Models. In *2024 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, Osaka, Japan, 55–62.

[10] Autonomous. 2024. AUTONOMOUS; Co-Designing Independence — autonomous-project.com. https://www.autonomous-project.com/. [Accessed 10-10-2025].

[11] Awan-Ur-Rahman. 2023. Understanding Soft Voting and Hard Voting: A Comparative Analysis of Ensemble Learning Methods. https://medium.com/@awanurrahman.cse/understanding-soft-voting-and-hard-voting-a-comparative-analysis-of-ensemble-learning-methods-db0663d2c008

[12] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas. 2014. Window Size Impact in Human Activity Recognition. *Sensors* 14, 4 (April 2014), 6474–6499. https://doi.org/10.3390/s140406474 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[13] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. IMU2Doppler: Cross-Modal Domain Adaptation for Doppler-based Activity Recognition Using IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[14] Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. 2022. Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 42:1–42:28. https://doi.org/10.1145/3534582

[15] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. 2024. An Introduction to Vision-Language Modeling. https://doi.org/10.48550/arXiv.2405.17247 arXiv:2405.17247 [cs].

[16] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit LeDuc, and Ioannis Kanellos. 2021. A Survey of Human Activity Recognition in Smart Homes Based on IoT Sensors Algorithms: Taxonomies, Challenges, and Opportunities with Deep Learning. *Sensors (Basel, Switzerland)* 21, 18 (Sept. 2021), 6037. https://doi.org/10.3390/s21186037

[17] A.J. Bernheim Brush, Bongshin Lee, Ratul Mahajan, Sharad Agarwal, Stefan Saroiu, and Colin Dixon. 2011. Home automation in the wild: challenges and opportunities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2115–2124. https://doi.org/10.1145/1978942.1979249

[18] Timothy I Cannings, Yingying Fan, and Richard J Samworth. 2020. Classification with imperfect training labels. *Biometrika* 107, 2 (2020), 311–330.

[19] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, Honolulu, HI, USA, 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[20] Gabriele Cipriani, Sabrina Danti, Lucia Picchi, Angelo Nuti, and Mario Di Fiorino. 2020. Daily functioning and dementia. *Dementia & Neuropsychologia* 14, 2 (2020), 93–102. https://doi.org/10.1590/1980-57642020dn14-020001

[21] Diane Cook, Narayanan Krishnan, and Parisa Rashidi. 2013. Activity Discovery and Activity Recognition: A New Partnership. *IEEE transactions on cybernetics* 43, 3 (June 2013), 820–828. https://doi.org/10.1109/TSMCB.2012.2216873

[22] Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, and Mario Cifrek. 2012. A brief introduction to OpenCV. In *2012 Proceedings of the 35th International Convention MIPRO*. IEEE, Opatija, Croatia, 1725–1730.

[23] Shohreh Deldari, Hao Xue, Aaqib Saeed, Jiayuan He, Daniel V. Smith, and Flora D. Salim. 2022. Beyond Just Vision: A Review on Self-Supervised Representation Learning on Multimodal and Temporal Data. arXiv:2206.02353 [cs.LG]

[24] Kaikai Deng, Dong Zhao, Zihan Zhang, Shuyue Wang, Wenxin Zheng, and Huadong Ma. 2024. Midas++: Generating Training Data of mmWave Radars From Videos for Privacy-Preserving Human Sensing With Mobility. *IEEE Transactions on Mobile Computing* 23, 6 (June 2024), 6650–6666. https://doi.org/10.1109/TMC.2023.3325399

[25] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. SmartPoser: Arm Pose Estimation with a Smartphone and Smartwatch Using UWB and IMU Data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3586183.3606821

[26] Megan V. Ha, Emma Russell, Haarika A. Reddy, Alexander K. Maytin, Dustin P. DeMeo, Riku Arakawa, Mayank Goel, Jill F. Lehman, and Bryan T. Carroll. 2024. Self-narration for patient monitoring with smartwatch technology in post-operative wound care after dermatologic surgery. *Archives of Dermatological Research* 316, 7 (June 2024), 389. https://doi.org/10.1007/s00403-024-03149-z

[27] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[28] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, 961–970. https://doi.org/10.1109/CVPR.2015.7298698

[29] Shruthi K. Hiremath, Yasutaka Nishimura, Sonia Chernova, and Thomas Plötz. 2022. Bootstrapping Human Activity Recognition Systems for Smart Homes from Scratch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (Sept. 2022), 1–27. https://doi.org/10.1145/3550294

[30] Shruthi K. Hiremath and Thomas Plötz. 2023. The Lifespan of Human Activity Recognition Systems for Smart Homes. *Sensors* 23, 18 (Jan. 2023), 7729. https://doi.org/10.3390/s23187729 Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[31] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. ColloSSL: Collaborative Self-Supervised Learning for Human Activity Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 17 (mar 2022), 28 pages. https://doi.org/10.1145/3517246

[32] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. 2017. A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems. *International Journal of Interactive Multimedia and Artificial Intelligence* 4, Regular Issue (2017), 54–62. https://www.ijimai.org/journal/bibcite/reference/2606

[33] Tianjie Ju, Yi Hua, Hao Fei, Zhenyu Shao, Yubin Zheng, Haodong Zhao, Mong-Li Lee, Wynne Hsu, Zhuosheng Zhang, and Gongshen Liu. 2025. Watch Out Your Album! On the Inadvertent Privacy Memorization in Multi-Modal Large Language Models. https://doi.org/10.48550/arXiv.2503.01208 arXiv:2503.01208 [cs].

[34] Alexander Karpekov, Sonia Chernova, and Thomas Plötz. 2025. DISCOVER: Data-driven Identification of Sub-activities via Clustering and Visualization for Enhanced Activity Recognition in Smart Homes. https://doi.org/10.48550/arXiv.2503.01733 arXiv:2503.01733 [cs].

[35] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[36] Gierad Laput and Chris Harrison. 2019. SurfaceSight: A New Spin on Touch, User, and Object Sensing for IoT Experiences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300559

[37] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In *Proc. of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 3986–3999. https://doi.org/10.1145/3025453.3025773

[38] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365.html
[39] Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. 2024. IMUGPT 2.0: Language-Based Cross Modality Transfer for Sensor-Based Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (Aug. 2024), 1–32. https://doi.org/10.1145/3678545
[40] Zikang Leng, Hyeokhyen Kwon, and Thomas Ploetz. 2023. Generating Virtual On-body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers (ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 39–43. https://doi.org/10.1145/3594738.3611361
[41] Zikang Leng, Hyeokhyen Kwon, and Thomas Plötz. 2023. On the Benefit of Generative Foundation Models for Human Activity Recognition. https://doi.org/10.48550/arXiv.2310.12085 arXiv:2310.12085 [cs].
[42] Dawei Liang, Guihong Li, Rebecca Adaimi, Radu Marculescu, and Edison Thomaz. 2022. AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers* (Cambridge, United Kingdom) *(ISWC '22)*. Association for Computing Machinery, New York, NY, USA, 44–48. https://doi.org/10.1145/3544794.3558471
[43] Sicong Liu, Junzhao Du, Anshumali Shrivastava, and Lin Zhong. 2019. Privacy Adversarial Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (dec 2019), 1–18. https://doi.org/10.1145/3369816
[44] Tian-Yu Liu. 2009. EasyEnsemble and Feature Selection for Imbalance Data Sets. In *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*. IEEE, Shanghai, China, 517–520. https://doi.org/10.1109/IJCBS.2009.22
[45] Harsh Lunia. 2024. Can VLMs be used on videos for action recognition? LLMs are Visual Reasoning Coordinators. https://doi.org/10.48550/arXiv.2407.14834 arXiv:2407.14834 [cs] version: 1.
[46] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, 11 (March 2017), 205. https://doi.org/10.21105/joss.00205
[47] Mites.io. 2020. Mites.io: a full-stack ubiquitous sensing platform. https://mites.io/.
[48] MMAction2. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.
[49] MMPose. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.
[50] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing Activities with Motion and Subsampled Audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–19.
[51] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3544548.3581392
[52] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-Based Sensor Fusion Techniques for Multimodal Human Activity Recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(ISWC '17)*. Association for Computing Machinery, New York, NY, USA, 158–165. https://doi.org/10.1145/3123021.3123046
[53] OpenAI. 2025. OpenAI API. https://platform.openai.com/docs/api-reference/ Accessed: 2025-04-29.
[54] The pandas development team. 2020. *pandas-dev/pandas: Pandas*. pandas-dev. https://doi.org/10.5281/zenodo.3509134
[55] Preksha Pareek and Ankit Thakkar. 2021. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review* 54, 3 (2021), 2259–2322.
[56] Prasoon Patidar, Mayank Goel, and Yuvraj Agarwal. 2023. VAX: Using Existing Video and Audio-based Activity Recognition Models to Bootstrap Privacy-Sensitive Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (Sept. 2023), 1–24. https://doi.org/10.1145/3610907
[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[58] Daniel Perazzo, Natalia Souza Soares, Victor Gouveia de Menezes Lyra, Gustavo Camargo Rocha Lima, Alana Elza Fontes da Gama, Joao Marcelo Xavier Natario Teixeira, and Veronica Teichrieb. 2022. OAK-D as a Platform for Human Movement Analysis: A Case Study. In *Proceedings of the 23rd Symposium on Virtual and Augmented Reality* (Virtual Event, Brazil) *(SVR '21)*. Association for Computing Machinery, New York, NY, USA, 167–171. https://doi.org/10.1145/3488162.3488222
[59] Prasoon Patidar, Riku Arakawa, Mayank Goel, Yuvraj Agarwal. 2025. OrganicHAR: Open-source repository for the OrganicHAR. https://github.com/synergylabs/OrganicHAR.
[60] Riccardo Presotto, Gabriele Civitarese, and Claudio Bettini. 2022. Federated Clustering and Semi-Supervised learning: A new partnership for personalized Human Activity Recognition. *Pervasive and Mobile Computing* 88 (2022), 101726.
[61] Suneth Ranasinghe, Fadi Al Machot, and Heinrich C Mayr. 2016. A review on applications of activity recognition systems with regard to performance and evaluation. *International Journal of Distributed Sensor Networks* 12, 8 (2016), 1550147716665520.

[62] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. 2024. Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning. https://doi.org/10.48550/arXiv.2310.12921 arXiv:2310.12921 [cs] version: 2.

[63] Laurens Samson, Nimrod Barazani, Sennay Ghebreab, and Yuki M. Asano. 2025. Little Data, Big Impact: Privacy-Aware Visual Language Models via Minimal Tuning. https://doi.org/10.48550/arXiv.2405.17423 arXiv:2405.17423 [cs].

[64] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 1 (Jan. 2010), 185–197. https://doi.org/10.1109/TSMCA.2009.2029559

[65] Pekka Siirtola and Juha Röning. 2019. Incremental Learning to Personalize Human Activity Recognition Models: The Importance of Human AI Collaboration. *Sensors (Basel, Switzerland)* 19, 23 (Nov. 2019), 5151. https://doi.org/10.3390/s19235151

[66] Adane Nega Tarekegn, Mohib Ullah, Faouzi Alaya Cheikh, and Muhammad Sajjad. 2023. Enhancing Human Activity Recognition Through Sensor Fusion And Hybrid Deep Learning Model. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSPW59220.2023.10193698

[67] Annalise Vaccarello, Alexander K. Maytin, Yash Kumar, Toluwalashe Onamusi, Haarika A. Reddy, Mayank Goel, Riku Arakawa, Jill Fain Lehman, and Bryan T. Carroll. 2024. Barriers to use of digital assistance for postoperative wound care: a single-center survey of dermatologic surgery patients. *Archives of Dermatological Research* 316, 7 (June 2024), 376. https://doi.org/10.1007/s00403-024-03025-w

[68] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[69] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. 2015. A review of human activity recognition methods. *Frontiers in Robotics and AI* 2 (2015), 28.

[70] Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2024. A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. arXiv:2411.03350 [cs.CL] https://arxiv.org/abs/2411.03350

[71] Shuai Wang, Luoyu Mei, Ruofeng Liu, Wenchao Jiang, Zhimeng Yin, Xianjun Deng, and Tian He. 2025. Multi-Modal Fusion Sensing: A Comprehensive Review of Millimeter-Wave Radar and Its Integration With Other Modalities. *IEEE Commun. Surv. Tutorials* 27, 1 (2025), 322–352. https://doi.org/10.1109/COMST.2024.3398004

[72] Pete Warden, Matthew Stewart, Brian Plancher, Colby Banbury, Shvetank Prakash, Emma Chen, Zain Asgar, Sachin Katti, and Vijay Janapa Reddi. 2022. Machine Learning Sensors. https://doi.org/10.48550/ARXIV.2206.03266

[73] Jason Wu, Chris Harrison, Jeffrey P. Bigham, and Gierad Laput. 2020. Automated Class Discovery and One-Shot Interactions for Acoustic Activity Recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376875

[74] Tong Wu, Murtadha Aldeer, Tahiya Chowdhury, Amber Haynes, Fateme Nikseresht, Mahsa Pahlavikhah Varnosfaderani, Jiechao Gao, Arsalan Heydarian, Brad Campbell, and Jorge Ortiz. 2021. The Smart Building Privacy Challenge. In *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation* (Coimbra, Portugal) *(BuildSys '21)*. Association for Computing Machinery, New York, NY, USA, 238–239. https://doi.org/10.1145/3486611.3492234

[75] Chengshuo Xia, Xinrui Fang, Riku Arakawa, and Yuta Sugiura. 2022. VoLearn: A Cross-Modal Operable Motion-Learning System Combined with Virtual Avatar and Auditory Feedback. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 81:1–81:26. https://doi.org/10.1145/3534576

[76] Kenji Yamanishi, Jun'ichi Takeuchi, Graham J. Williams, and Peter Milne. 2004. On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. *Data Mining and Knowledge Discovery* 8, 3 (2004), 275–300. https://doi.org/10.1023/B:DAMI.0000023676.72185.7c

[77] A. Murat Yağcı, Tevfik Aytekin, and Fikret S. Gürgen. 2016. Balanced random forest for imbalanced data streams. In *2016 24th Signal Processing and Communication Application Conference (SIU)*. IEEE, Zonguldak, Turkey, 1065–1068. https://doi.org/10.1109/SIU.2016.7495927

[78] Hyungjun Yoon, Hyeongheon Cha, Hoang C. Nguyen, Taesik Gong, and Sung-Ju Lee. 2024. IMG2IMU: Translating Knowledge from Large-Scale Images to IMU Sensing Applications. https://doi.org/10.48550/arXiv.2209.00945 arXiv:2209.00945 [cs].

[79] Sojeong Yun and Youn-kyung Lim. 2025. What If Smart Homes Could See Our Homes?: Exploring DIY Smart Home Building Experiences with VLM-Based Camera Sensors. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3706598.3713265

[80] Shugang Zhang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li. 2017. A Review on Human Activity Recognition Using Vision-Based Method. *Journal of Healthcare Engineering* 2017, 1 (2017), 3090343. https://doi.org/10.1155/2017/3090343

[81] Xian Zhang and Xiang Cheng. 2025. Evaluation of Geolocation Capabilities of Multimodal Large Language Models and Analysis of Associated Privacy Risks. https://doi.org/10.48550/arXiv.2506.23481 arXiv:2506.23481 [cs].

[82] Haozhe Zhou, Riku Arakawa, Yuvraj Agarwal, and Mayank Goel. 2025. IMUCoCo: Enabling Flexible On-Body IMU Placement for Human Pose Estimation and Activity Recognition. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, UIST 2025, Busan, Korea, 28 September 2025 - 1 October 2025*. ACM, New York, NY, USA, 91:1–91:16. https://doi.org/10.1145/3746059.3747695

A APPENDIX

A.1 Featurization Pipelines for Various Sensing Modalities

This appendix provides detailed technical specifications for the featurization pipelines used for specific sensing modalities. The following detailed descriptions outline the specific features extracted from each sensor, and the rationale behind our design choices, providing the implementation details necessary for replicating our multimodal sensing approach.

- **Doppler Radar:** We extract temporal features (velocity statistics, directional changes), spatial features (distance from sensor, movement extent), complexity features (velocity entropy, motion transitions), and motion type (stationary, slow/fast movement *etc.*) distributions. While Doppler provides precise and high fidelity velocity sensing, it only captures motion in the radial direction and struggles with distinguishing movement at the same distance or complex spatial relationships.
- **2D LiDAR:** Our approach first establishes a static boundary model in the horizontal 2D plane, then extracts features from deviations that represent dynamic objects. Features include centroid positions, velocities, point distributions, boundary interactions, and temporal patterns. LiDAR provides precise spatial mapping but with decreasing granularity at greater distances, and it cannot capture fine-grained movements or height information.
- **Low-Resolution Thermal Camera:** From the 10×10 thermal array, we compute spatial features (temperature statistics, gradients) and signature-based features that identify thermal patterns from humans, appliances, and ambient sources. While thermal sensing excels at detecting heat-generating activities and object interactions, it is limited to coarse human movements or thermal changes.
- **IMU (Wearables):** We leverage the pretrained SAMoSA [50] model to extract 128-dimensional motion features from 2.88-second windows with a stride of 0.21 seconds. Wearable IMUs capture detailed hand and arm movements but are limited to the wearer's perspective and require consistent device wearing.
- **Pose:** From the raw 2D pose data over 25 body keypoints (in 640x480 frame) at 6-8Hz, we extract biomechanical features that characterize body positioning and movement, including joint velocities and accelerations, inter-joint configurations, working zones (based on torso and hand movements), and movement complexity. Despite providing rich body movement data, pose estimation is constrained by occlusions and the limited field of view (less than 60 degrees), sometimes missing key interactions outside the camera's perspective.
- **Depth:** We extract motion patterns and statistical features from the 10×10 depth array, establishing boundary maps to identify deviations representing people and objects. While effective for detecting presence and basic movements, depth sensing is limited by its narrow field of view and cannot capture fine-grained interactions.

A.2 Prompts for LLMs and VLMs

This appendix provides the detailed prompts used in the OrganicHAR framework for activity discovery. The pipeline employs five main prompts for Vision Language Models (VLMs) and Large Language Models (LLMs) that work in sequence to translate video clips into consistent activity labels at appropriate granularity levels. The exact prompts can be found in the Supplemental File as a text file. Here, we describe the key design for each prompt to address specific challenges in the pipeline.

(1) The first step leverages a VLM to generate rich semantic descriptions from video frames captured during key moments. This prompt instructs the VLM to analyze short video clips focusing on actions, objects, locations, and activity structure. It specifically directs the model to focus only on clearly observable elements, provide confidence scores, and exclude low-confidence observations.
(2) Since VLM outputs contain diverse location references (*e.g.*, "at sink," "by counter"), our second step employs an LLM to consolidate these into consistent functional zones. This prompt directs the LLM to

cluster locations based on supported activities, ensuring that areas like "sink area," "counter area," and "coffee machine area" are consistently identified. This spatial context is crucial because activities have different meanings in different locations.

(3) After establishing functional zones, we use a two-step process to transform the unstructured VLM descriptions into discrete activity labels. First, we employ the clustering prompt to create initial activity clusters for each functional zone. This prompt directs the LLM to preserve task-specific distinctions while minimizing the overall number of clusters.
(4) Next, we match each VLM description to the appropriate cluster using the matching prompt. This prompt directs the LLM to evaluate matches based on action alignment (60%), object consistency (25%), and location match (15%), ensuring that descriptions are consistently mapped to appropriate activity labels.
(5) Finally, to enable granularity control through the relaxation parameter λ, we analyze each activity's semantic properties across multiple dimensions. This prompt directs the LLM to break down each activity in terms of action type, objects involved, sub-location, purpose, related activities, contrasting activities, adjacent objects, and access patterns. These dimensions form the basis for computing pairwise similarity scores between activities, which are then used with the relaxation parameter λ to determine which activities should be merged at different granularity settings.

A.3 VLM Variability and Configuration Analysis

We conducted additional analyses examining how different VLM models and frame rate configurations impact label generation performance. These evaluations provide insights into design choices and system adaptability across different VLM capabilities.

A.3.1 Performance of Label Discovery pipeline with different VLMs. We evaluated OrganicHAR using three state-of-the-art vision language models as of July 2025: OpenAI's GPT-4.1, Google's Gemini 2.5 Flash, and Anthropic's Claude Sonnet 4. These models represent different architectures and training approaches currently available. For consistency, all models used identical prompts with minor formatting adaptations for model-specific requirements.

Table 15 shows notable variations across models. GPT-4.1 and Gemini 2.5 Flash demonstrate similar accuracy levels in Conservative and Balanced settings (80-85%), while Claude Sonnet 4 shows lower accuracy but reasonable performance in the Balanced setting (60.4% accuracy, 55.4% F1). However, accuracy and F1 scores alone do not capture whether VLMs can consistently provide meaningful activity descriptions. To measure this consistency, we examine the detection rate—the percentage of video segments for which the VLM generates actual location and action information rather than empty responses. GPT-4.1 maintains a high detection rate (92.0%) across all settings, while Gemini 2.5 Flash achieves only 50.2% detection rate despite competitive accuracy when successful, and Claude Sonnet 4 maintains a high detection rate (89.2%) but with lower overall accuracy. For our use case, GPT-4.1 performs the best in terms of both accuracy and reliability, making it the most suitable choice for consistent activity discovery across diverse scenarios.

A.3.2 Frame Rate Impact Analysis. Based on GPT-4.1's combination of high accuracy and detection rate, we selected it for frame rate analysis. Using GPT-4.1, we evaluated performance across three frame rates (1, 3, and 5 FPS) using identical 5-second video clips from our dataset. Table 16 shows an interesting pattern: increasing frame rate from 1 to 5 FPS results in modest performance decreases rather than improvements. Accuracy drops from 84.9% to 82.4% in the Conservative setting, with similar trends observed across all granularity levels. This finding suggests that kitchen activities may operate at coarse timescales where meaningful changes occur over multi-second intervals. Additional frames often capture intermediate postures, repetitive motions, or transitional states that may not provide additional semantic information useful for activity classification. The 1 FPS configuration

VLM Model Comparison				
Granularity	**Model**	**Acc (%)**	**F1 (%)**	**Detection Rate (%)**
Conservative ($\lambda = 0.4$)	GPT-4.1	84.9	72.5	92.0
	Gemini 2.5	80.9	69.7	50.2
	Claude Sonnet 4	59.3	50.8	89.2
Balanced ($\lambda = 0.3$)	GPT-4.1	83.9	69.4	92.0
	Gemini 2.5	79.0	67.2	50.2
	Claude Sonnet 4	60.4	55.4	89.2
Relaxed ($\lambda = 0.2$)	GPT-4.1	72.0	55.4	92.0
	Gemini 2.5	69.0	54.0	50.2
	Claude Sonnet 4	48.7	35.6	89.2

Fig. 15. Label discovery performance across VLM models and semantic granularity settings. Detection rate represents the % of key moments that generate useful activity descriptions.

Frame Rate Impact Analysis				
Frame Rate	**Granularity**	**Acc (%)**	**F1 (%)**	**Detection Rate (%)**
1 FPS	Conservative	84.9	72.5	92.0
	Balanced	83.9	69.4	92.0
	Relaxed	72.0	55.4	92.0
3 FPS	Conservative	82.9	70.5	91.3
	Balanced	80.6	62.5	91.3
	Relaxed	69.0	52.4	91.3
5 FPS	Conservative	82.4	67.5	90.6
	Balanced	79.2	65.1	90.6
	Relaxed	66.7	53.1	90.6

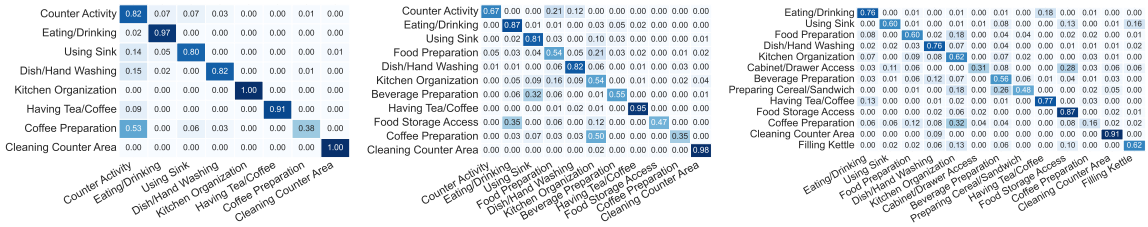
Fig. 16. Impact of frame rate on label discovery performance using GPT-4.1. Detection rates remain consistent across frame rate settings.

appears to capture key semantic moments while avoiding potential temporal noise. Frame rate selection also has practical implications for deployment. Processing at 5 FPS requires five times the input context in the VLM API calls compared to 1 FPS, with corresponding increases in latency, bandwidth usage, and cloud processing costs. In conclusion, our analysis indicates that this increased computational overhead does not translate to improved recognition performance under current VLM capabilities.

A.3.3 Implications for System Design. These evaluations provide several insights for OrganicHAR deployments. The detection rate metric reveals that model selection involves trade-offs between accuracy when successful and reliability of processing, where consistent processing across all key moments may be more important than marginal accuracy improvements on successfully processed segments. The frame rate analysis supports our original configuration choices (1 FPS, 640×480 resolution) as a reasonable balance between performance and computational efficiency. The consistent performance patterns across multiple VLMs indicate that fine-grained activity recognition limitations (F1 scores of 48-72% in Relaxed settings) reflect current VLM capabilities rather than weaknesses in our approach, as similar performance degradation occurs across different model architectures when distinguishing semantically similar activities. OrganicHAR's modular architecture enables adaptation to different VLMs without modifying sensor processing components, allowing the framework to incorporate future advances in VLM temporal reasoning and fine-grained visual understanding while maintaining efficiency for current deployments.

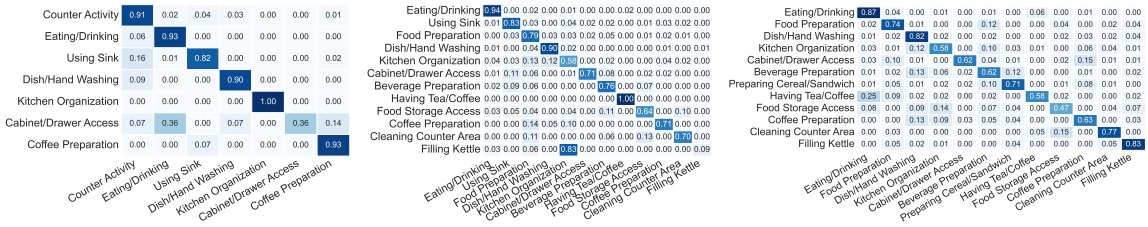
A.4 Activity Confusion Analysis (Other Sensor Configuration)

We present the confusion matrices for the other configurations in our evaluation. First, adding wearable IMU sensing to basic ambient sensors reveals complementary strengths and persistent challenges (Figure 17). Activities with distinctive motion patterns like "Kitchen Organization" and "Cleaning Counter Area" show excellent recognition across granularity settings. However, activities sharing similar motion signatures in different locations remain challenging to differentiate. In Conservative settings, "Coffee Preparation" frequently confuses with



(a) Conservative ($\lambda = 0.4$) (b) Balanced ($\lambda = 0.3$) (c) Relaxed ($\lambda = 0.2$)

Fig. 17. Confusion matrices showing activity recognition performance for *Ambient(Basic)+Wearable(IMU)* configuration across three granularity settings.



(a) Conservative ($\lambda = 0.4$) (b) Balanced ($\lambda = 0.3$) (c) Relaxed ($\lambda = 0.2$)

Fig. 18. Confusion matrices showing activity recognition performance for *Ambient(Advanced)+Wearable(IMU)* configuration across three granularity settings.

broader "Counter Activity" categories. The Balanced setting reveals confusion between "Beverage Preparation" and "Using Sink," while the Relaxed setting shows increased confusion between semantically related activities like "Cabinet/Drawer Access" and "Food Storage Access." These patterns demonstrate that wearable sensing effectively complements ambient sensors for activities with unique motion signatures but cannot fully resolve spatial ambiguities for activities with similar hand movements performed in different contexts.

Secondly, the most advanced configuration combining ambient sensors with wearable IMU, pose estimation, and depth sensing shows notable improvements in activity disambiguation (Figure 18). With pose and depth information, activities that were previously confused due to spatial ambiguity show clearer separation. Certain activities like "Having Tea/Coffee" and "Eating/Drinking" achieve particularly strong recognition rates across settings. However, even with this rich sensing configuration, some challenging distinctions remain—particularly for activities with subtle differences like "Filling Kettle" (which often confuses with general "Kitchen Organization") and "Food Storage Access" (which sometimes confuses with "Cleaning Counter Area"). The advanced configuration particularly excels at disambiguating fine-grained activities that involve distinctive postures or spatial relationships, confirming that spatial awareness through pose and depth information provides a crucial complement to motion and environmental sensing. This highlights how different sensor modalities each contribute unique capabilities to the overall recognition system.