2026 IEEE International Conference on Pervasive Computing and Communications (PerCom)

Code Reviewed Dataset Reviewed

Wandatch: Infrastructure-Free Point-to-Command with Smartwatches and Speakers

Lin Chen[1], Yandao Huang[1,2], Minghui Qiu[1], Shuxin Zhong[1*], Jun Chen[1], and Kaishun Wu[1]
The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{lchen297, mqiu585, jchen512}@connect.hkust-gz.edu.cn, yhuangfg@connect.ust.hk
{shuxinzhong, wuks}@hkust-gz.edu.cn

***Abstract*—Imagine walking into a room, flicking your wrist to light a lamp or swirling in the air to shift a fan's speed—an effortless *point-and-command* interaction that remains elusive despite billions of IoT devices in our homes. Decades of work have turned to cameras, wearables, anchors, and arrays, but these designs come at a cost: line-of-sight, intrusive hardware, or tedious calibration—constraints that leave systems brittle and unscalable. We present Wandatch, a system that repurposes everyday smartwatches and speakers into a *wrist ray*—a virtual laser that knows both where the hand points and what gesture it performs. Realizing this vision hinges on two challenges: (C1) localization burdened by hardware overhead and training requirements, and (C2) gestures trapped in a rigid "gesture–pause–gesture" routine. Wandatch addresses these with two designs: *Physics-driven Localization*, which fuses acoustic ranging with Inertial Measurement Unit (IMU) orientation to cast a precise ray for appliance selection; and *Universal Gesture Interaction*, which recognizes the full spectrum of gestures—from subtle micro-gestures for lightweight activation to expressive multi-stroke gestures for rich control. In evaluation, Wandatch achieves 97.6% selection accuracy in sparse deployments and sustains 85.5% accuracy even under dense 40 cm spacing. Compared to smartphone apps and voice control, it reduces interaction time by 12.6–50.1% and significantly improves usability across ten subscales. Together, these results demonstrate *the first practical path toward natural, infrastructure-free point-and-command interaction*.**

***Index Terms*—Smart Appliance, Indoor Positioning, Gesture Recognition, Cross-device Interaction**

I. Introduction

The number of Internet of Things (IoT) devices is skyrocketing, expected to grow from 16.6 billion in 2023 to more than 40 billion by 2030, with the majority embedded in consumer environments [1]. Homes are rapidly becoming dense IoT ecosystems filled with smart speakers, lamps, fans, and countless appliances. Yet interaction remains frustratingly fragmented—apps buried under menus, voice commands that fail at the wrong moment, and remotes scattered across brands [2]–[4]. Now imagine a different experience: you walk into a room, raise your wrist, and simply point (Fig. 1). A flick of your hand lights up the lamp; a swirl in the air shifts the fan speed. No device IDs to recall, no wake words to shout—just *point-and-command*, like casting a spell with a wand.

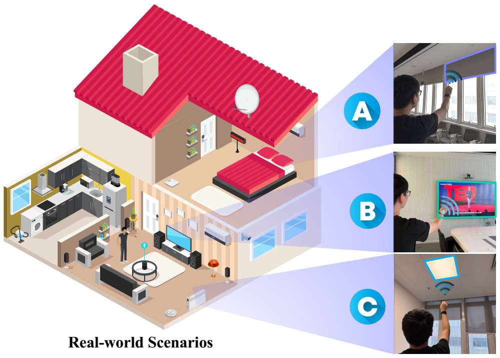

Fig. 1: Smart home interaction with Wandatch: (A) Opening curtain, (B) Playing media, (C) Adjusting light brightness.

A long line of research has chased the vision of pointing-based interaction in smart environments. *Vision-based systems* used cameras and Kinect mounted in rooms [5], handheld projectors tied to photo-sensors on appliances [6], or wearable variants such as vision-based rings that embed cameras [2], and even AR systems that establish an "optical tether" [3]. Yet all of these approaches *demand line-of-sight and bulky, intrusive hardware*. *Instrumentation-based systems* went in the opposite direction, attaching accelerometers, magnetometers, or ultrasonic receivers to every appliance [7], but this becomes *costly and infeasible at scale*. *Wireless and acoustic solutions* took a third path—from UWB anchors [8] to large microphone arrays [9]—but still rely on dense infrastructure and painstaking calibration. In short, despite decades of effort, existing solutions remain *device- and environment-specific*, breaking scalability and preventing the seamless, cross-device experience a truly intuitive smart home demands.

We take a different path: instead of adding new hardware, we repurpose what people already own—a smartwatch on the wrist and a smart speaker in the room. The watch can act as both an ultrasonic beacon and a gesture sensor; the speaker as a spatial anchor. Their acoustic interaction recovers *the watch's 3D position*, while the IMU provides *orientation*. No fingerprints, no per-room training—just physics-driven signals. Unified, these cues project *a virtual ray from the wrist*, like a laser pointer that directly selects and controls appliances, transforming commodity devices into a seamless *point-to-command* interface for the home.

Yet, realizing this vision confronts two barriers. First, hardware overhead and training burden (**C1**). Current localization systems either scatter beacons across every room, each needing

[*] is the corresponding author.

979-8-3315-7613-4/26/$31.00 ©2026 IEEE

a painstaking calibration [10]–[14], or rely on fingerprinting that force room-by-room data collection and retraining [9]. Both make deployments fragile, costly, and fundamentally non-scalable. Second, gesture recognition that fails in the wild (**C2**). Wearables still assume a rigid "gesture–pause–gesture" ritual [15], [16]. Remove the pause and signals blur, classifiers collapse, and accuracy crumbles as soon as the watch is switched to a different wrist or the user changes.

To this end, we design `Wandatch`, a seamless *point-to-command* system that repurposes everyday smartwatches and speakers into a magic wand for the smart home. Its design rests on two components that together turn raw acoustic and inertial signals into reliable control. For **C1**, *Physics-driven Localization* distills noisy acoustic and IMU signals into a clean virtual ray that unifies position and orientation. This unfolds as a four-stage pipeline: acoustic ranging anchors distance, position estimation fuses orientation, target inference casts a ray to pinpoint the intended appliance, and deployment support automates device management to ensure practicality in real homes. For **C2**, *Universal Gesture Interaction* unifies gestures from quick wrist taps to multi-stroke trajectories into a single interface, enabling calibration-free recognition that remains robust across users, contexts, and time.

In summary, the main contributions are summarized as:

- We establish the first infrastructure-free point-to-command interaction, transforming commodity smartwatches and speakers into a universal wrist-ray interface—no cameras, no tags, no calibration.
- `Wandatch` consists of two technical components: i) *Physics-driven Localization*, which distills raw acoustic and IMU signals into a clean wrist-ray abstraction through four stages (ranging, position estimation, target inference, deployment support), achieving training-free, scalable localization; and ii) *Universal Gesture Interaction*, an interface spanning micro- and macro-gestures, enabling both lightweight activation and expressive control while remaining robust across users, contexts, and time.
- We conduct extensive user studies with real appliances, showing that `Wandatch` achieves 97.60% selection accuracy in sparse deployments and maintains 85.47% accuracy even at 40 cm spacing. Compared to smartphone apps and voice control, it cuts interaction time by 12.62%–50.09% and consistently ranks higher across ten usability sub-scales.

II. Related Works

A. Pointing-based Interaction System

Decades of systems have chased the vision of pointing as the most natural command. Early designs like XWand [17] combined cameras and IMUs, while iThrow [18] and MagicPhone [19] turned to UWB and phone orientation. Kinect-based approaches lifted this into full-body skeleton [5], and PICOntrol pushed hardware into appliances themselves with embedded photo-sensors [6]. More recent efforts have pushed novel hardware—rings with cameras (IRIS [2]), AR glasses with optical tethers (EchoSight [3]) or arrays of accelerometers, magnetometers, and ultrasonic receivers mounted on appliances [7]. Others scaled up to large arrays, such as Soundr's head-orientation–based selection [9], but demanded tedious, room-specific calibration. *These approaches reveal a tradeoff: every gain in sensing fidelity is paid for with bespoke hardware, dense instrumentation, or brittle calibration—costs that make real-world deployment untenable.*

B. Ultrasonic Indoor Positioning

A rich line of work has advanced ultrasonic positioning (UPS) as a foundation for indoor interaction. Early systems such as BAT [10] demonstrated meter-scale ranging with dedicated ultrasonic beacons, and ALPS [11] improved accuracy using TDoA chirps synchronized over BLE. Later efforts sought to broaden deployment, but still required specialized setups: CAT [12] and MilliSonic [13] relied on multiple speakers or microphone arrays for 3D tracking, UPS+ [14] customized beacon hardware for room-scale coverage, and SyncEcho [20] exploited higher-order echoes but degraded in sound-absorbing environments. *Collectively, these systems highlight a fundamental tradeoff: precise UPS often demands bespoke transducers, dense arrays, or controlled environments, limiting their practicality at scale.*

C. IMU-based Gesture Recognition

Wearable inertial sensors have long been explored for gesture-driven interaction. Early systems such as SHOW [16] and AirContour [15] reconstructed 3D trajectories for handwriting, but were confined to discrete characters—forcing pauses between strokes and breaking natural flow. More recent work pushed toward continuity: SignSpeaker [21], ViFin [22], and WearSign [23] leveraged deep sequence models (e.g., LSTM, CTC) to recognize air-writing and even sign language continuously. Yet, this continuity comes at a cost: such models hinge on extensive training data and often require per-user calibration, leaving them brittle and impractical to scale.

Summary. In contrast, `Wandatch` transforms COTS smartwatches and speakers into a universal substrate for positioning and pointing, removing the need for per-device sensors or calibration and realizing seamless, room-scale *point-to-command*.

III. System Overview

A. Design Principles

The design of `Wandatch` is driven by a single goal: *to make smart-home control as natural as pointing a finger*. To this end, we distill four guiding principles:

- **Ubiquity.** Leverage only commodity hardware already in homes—IMU and acoustic sensors on smartwatches, and off-the-shelf smart speakers—without requiring UWB chips or per-device tags [2], [3], [18], [19].
- **Deployability.** Replace brittle fingerprinting [9] with physics-driven localization that works out-of-the-box, needing only a speaker and appliance binding.
- **User Independence.** Avoid per-user calibration [21], [23] by designing recognition that generalizes across individuals, so newcomers can gesture immediately without retraining.

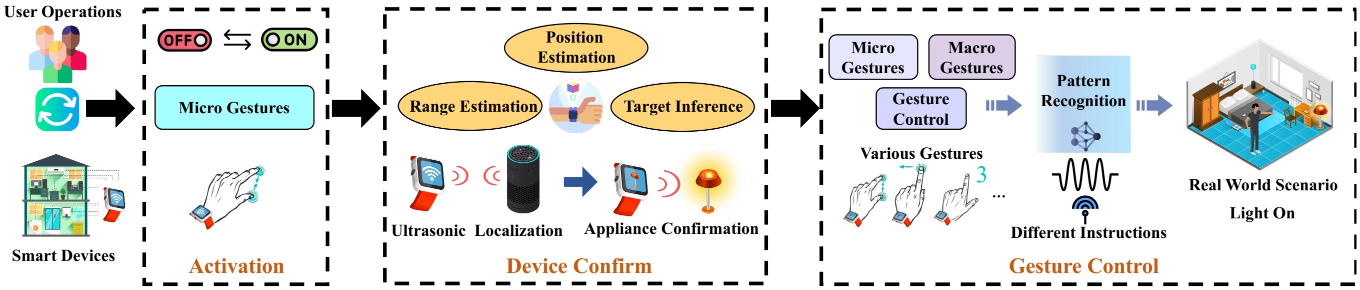

Fig. 2: System Overview of `Wandatch`.

- **Precision and Efficiency.** Deliver fine-grained localization and robust recognition so appliances are controlled correctly on the first attempt, saving time and avoiding frustration.

B. System Workflow

As shown in Fig. 2, `Wandatch` turns smart-home control into a fluid three-step ritual, with each step powered by a core component. In **Step 1 (Activation)**, the user simply raises their smartwatch-wearing hand and double-taps—a lightweight micro-gesture handled by *Universal Gesture Interaction*, enabling instant, calibration-free wake-up. In **Step 2 (Device Confirmation)**, the watch and speaker perform a brief acoustic handshake; fused with IMU cues, *Physics-driven Localization* reconstructs a 3D wrist-ray to resolve the intended appliance, confirmed via haptic and voice feedback. In **Step 3 (Gesture Control)**, the user performs in-air gestures—ranging from quick swipes to multi-stroke digits—to command functions. Here, *Universal Gesture Interaction* ensures gestures remain robust across users and contexts, unifying micro- and macro-movements into expressive control.

C. Core Components

Behind this workflow are two tightly integrated components:

- **Physics-driven Localization** transforms raw acoustics and IMU readings into a clean wrist-ray abstraction. It measures ranges, reconstructs 3D pose, and projects a virtual ray to pinpoint appliances, while deployment support ensures robustness in real homes—delivering precise, training-free localization (detailed in Sec. IV).
- **Universal Gesture Interaction** spans the full spectrum of human motion, from quick taps to multi-stroke trajectories, unifying them into one interface. It delivers calibration-free recognition for both activation and control, staying robust across users, contexts, and time (detailed in Sec.V).

IV. Physics-driven Localization

`Wandatch` relies on the microphone arrays found in standard smart speakers. This design choice aligns with commercial off-the-shelf devices; for example, the Amazon Echo [24] and Apple HomePod [25] are typically equipped with circular arrays of 4-7 microphones, which offer sufficient spatial resolution for acoustic sensing.

Physics-driven Localization unfolds as a four-stage journey from raw sound into seamless control. First, *Range Estimation* measures the acoustic distance between watch and speaker, laying the foundation. Second, *Position Estimation* recovers the watch's 3D position by fusing these ranges. Third, *Target Inference* reconstructs the watch's 3D pose and transforms the pose into a virtual ray to pinpoint the target appliance. Finally, *Database Maintenance* keep the system practical, automating database upkeep for real homes.

A. Range Estimation

We build on **Double-Sided Two-Way Ranging** (DS-TWR) [26]. The principle is simple: by exchanging chirps in both directions and averaging their delays, DS-TWR cancels clock offsets while retaining precise propagation timing.

Concretely, the smartwatch emits an up-chirp (17–23 kHz) timestamped locally at t_w^1 and at the smart speaker at t_b^1. The speaker then replies with a down-chirp (23–17 kHz), logged at t_b^2 and t_w^2. Each chirp lasts 40 ms (1920 samples at 48 kHz), balancing noise resilience with low overhead. The round-trip distance D is computed as $D = \frac{(t_b^1 - t_w^1) + (t_w^2 - t_b^2)}{2} \times c$, where c is the speed of sound. This design achieves centimeter-level precision with $\sim$880 ms latency, sufficient for real time.

A key challenge in real homes, however, is *multipath reflections*, which produce delayed replicas that obscure the true line-of-sight (LoS) path. To overcome this, `Wandatch` introduces a **Multipath-Resilient Delay Estimation** module consisting of two steps:

Delay Estimation. The received signal $y(t)$ is transformed into a complex analytic form using a Hilbert transform [27]:

$$\hat{y}(t) = \alpha \exp\Big(i\big[2\pi(f_0(t-\tau) + \tfrac{B}{2T}(t-\tau)^2)\big]\Big), \quad (1)$$

where α is the attenuation factor and τ is the propagation delay. Mixing $\hat{y}(t)$ with the reference waveform yields:

$$\begin{aligned} m(t) &= \hat{y}(t)\exp\Big(-i2\pi(f_0 t + \tfrac{B}{2T}t^2)\Big) \\ &= \alpha \exp\Big(i2\pi\big(-f_0\tau + \tfrac{B}{2T}\tau^2 - \tfrac{B}{T}\tau t\big)\Big), \end{aligned} \quad (2)$$

whose spectrum exposes a frequency bin proportional to $-\frac{B\tau}{T}$. The delay τ is obtained by locating the dominant spectral peak.

Robust LoS Detection. To suppress multipath artifacts, we first use a noise-only frame to estimate the baseline energy distribution, which provides a dynamic threshold. Subsequent frames yield candidate peaks, among which we select the one consistent with the physical constraint that the LoS path must correspond to the shortest delay (i.e., the highest-frequency component).

B. Position Estimation

As illustrated in Fig. 3 (a), the watch-to-speaker distance D alone is insufficient; we also need the orientation of the smartwatch relative to the microphone array.

A natural approach is angle-of-arrival (AoA) estimation, which fundamentally relies on resolving the inter-microphone time delays. The bandwidth B of ultrasonic chirps constrains the achievable delay resolution to $\Delta\tau \approx \frac{1}{2B}$, corresponding to a distance resolution of $\frac{c}{2B} \approx 2.83$ cm at $B = 6$ kHz. To enhance the practical estimation accuracy, we enlarge the FFT size when processing the mixed signal $m(t)$, which refines the spectral representation and enables more accurate localization of the delay peaks, thereby improving the robustness of AoA estimation in practice.

Formally, we construct the observed matrix M, where each element a_{ij} is the inter-microphone delay. For each candidate orientation (θ, ϕ) consistent with D, we compute a predicted matrix $\hat{M}(\theta, \phi)$ from the geometric model. The orientation is then obtained by minimizing the Frobenius norm:

$$(\theta, \phi) = \arg \min_{(\theta_i, \phi_i) \in S} ||M - \hat{M}(\theta_i, \phi_i)||_F, \tag{3}$$

where S enumerates candidate orientations with 1° resolution, yielding 91×360 pairs. The smartwatch's 3D position in the array's local coordinate system is then derived as (Fig. 3 (b)):

$$p_w = \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = D \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}, \tag{4}$$

which is further transformed into the global coordinate system via a rotation matrix R provided by the Android API.

C. Target Inference

To derive the watch's pointing direction in the microphone array coordinate system, we transform its reference vector in the watch frame through two coordinate systems:

$$\vec{v_w} = R_e^m * R_w^e * v_w^{watch}, \tag{5}$$

where R_w^e rotates the vector from the watch to the earth frame, and R_e^m further maps it to the microphone array frame. The reference vector v_w^{watch} is set to $\begin{bmatrix} -1 & 0 & 0 \end{bmatrix}^T$ when the watch is worn on the right hand and $v_w^{watch} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$ when worn on the left hand. All rotation matrices are obtained via Android's `SensorManager.getRotationMatrix`.

In practical smart-home environments, appliances may lie close together or even align along the same pointing direction, creating ambiguity in identifying the intended target. To resolve such ambiguity, we compute the likelihood of appliance i being selected as the cosine similarity between the pointing vector v_w and displacement vector v_d^i from the watch to the device (Fig. 4):

$$\begin{cases} prob(i) = \frac{v_w \cdot v_d^i}{||v_w|| \cdot ||v_d^i||} \\ v_d^i = p_d^i - p_w, \end{cases} \tag{6}$$

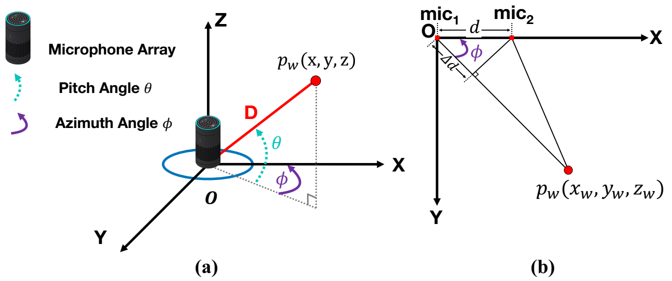

Fig. 3: (a) Angle notations; (b) Illustration of AoA model.

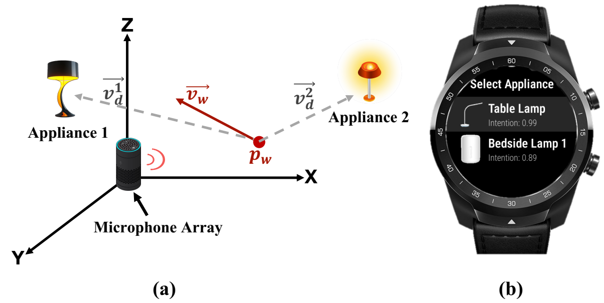

Fig. 4: (a) Illustration of vectors used in device selection; (b) The user interface of appliance confirmation.

where p_d^i and p_w denote the positions of the watch and device i, respectively. Devices with similarity above threshold (0.85 empirically) are retained as candidates. If none remain, the operation is ignored; if exactly one remains, it is selected automatically. When multiple candidates persist, `Wandatch` enters a confirmation phase: candidates are listed on the watch (Fig. 4 (b)) and disambiguated either via touchscreen input or lightweight gestures (UP/DOWN to navigate, TAP to confirm). Upon confirmation, a connection is established and can be terminated either explicitly (ROTATE gesture) or implicitly via a 5s timeout, reducing unintended operations.

D. Database Maintenance

Smart-home environments are inherently dynamic: appliances may be frequently relocated, and even the smart speaker anchor itself may be moved as furniture is rearranged. To deal with this, `Wandatch` maintains an up-to-date appliance database that adapts to both device- and anchor-level changes. For *ordinary appliances*, new entries are logged by bringing the smartwatch close to the device and performing an activation gesture, while relocated devices are directly remeasured and their coordinates updated. When the *smart speaker anchor* itself is moved, however, its displacement d and rotation R relative to the old coordinate system are estimated; all appliance coordinates are then transformed as $p_d^i = R(\bar{p}_d^i - d)$, ensuring that spatial mappings remain globally consistent.

V. Universal Gesture Interaction

To enable expressive yet lightweight control, `Wandatch` introduces *Universal Gesture Interaction*, which systematically recognizes the full spectrum of gestures: **micro gestures**,

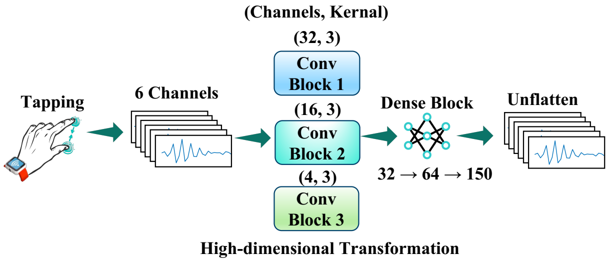

Fig. 5: Convolutional autoencoder architecture.

relying on subtle wrist motion for lightweight *activation* (e.g., tapping), and **macro gestures**, leveraging full arm movement for rich *control*, spanning from *atomic single-stroke* to *composite multi-stroke* gestures.

A. Micro Gestures Recognition

Micro gestures, such as tapping, are well-suited for activation due to their subtlety, yet their motion signatures are easily obscured by incidental wrist activity and sensor noise. To address this, `Wandatch` adopts a three-stage progressive pipeline that incrementally refines detections—pruning candidates from *possible* to *authentic* to *reliably valid* gestures. We adopt *double-tap* as the activation gesture.

- **Detection.** Micro gestures typically appear as short, high-frequency bursts distinct from background wrist motion. We therefore apply a 30 Hz high-pass Butterworth filter on 3-axis accelerometer signals to suppress gravity and drift, followed by frame energy computation in a 50 ms sliding window. Peaks exceeding a threshold mark candidate taps, and segments of ±120 ms around each peak are extracted to capture the full gesture.
- **Filtering.** Incidental wrist activity often produces spurious bursts that resemble taps. To filter these out, we use a lightweight convolutional autoencoder (CAE, 0.06 MB) trained exclusively on tapping signals (Fig. 5). The CAE captures the spectral–temporal signatures common to tapping, yielding low reconstruction error for genuine taps. Incidental motions that lack these universal patterns produce high error and are rejected, enabling cross-user generalization without per-user calibration.
- **Verification.** To minimize false activations in noisy environments, we leverage the behavioral regularity that intentional taps are often performed as double taps with consistent temporal structure. We assess the similarity between consecutive segments using dynamic time warping (DTW), and only pairs passing this check are accepted as valid activations.

B. Macro Gestures Recognition

To support diverse interaction needs, `Wandatch` employs a three-layer hierarchy that progressively transforms raw IMU signals into semantic gestures: (1) *Motion Normalization* applies roll-compensated rotation to produce pose-invariant yaw and pitch trajectories; (2) *Single-stroke Gesture Recognition* uses out-and-return modeling with gyro–acc fusion to robustly identify directional gestures across both hands; and (3) *Multi-stroke Gesture Recognition* combines transition-resilient segmentation—removing spurious transitions through stroke decomposition and progressive pruning—with a dual-modality representation that fuses spatial maps and temporal sequences for robust, user-independent recognition.

1) Motion Normalization: Raw gyroscope signals are reported in the local watch frame, which depends on how the device is worn on the wrist. As a result, identical gestures (e.g., raising the arm upward) produce inconsistent three-axis patterns under different wrist poses. To eliminate this ambiguity, we normalize motion by transforming gyroscope readings from the watch frame into a global reference frame.

We first estimate the watch's roll angle relative to gravity using accelerometer measurements:

$$roll = arctan2(a_y, a_z). \tag{7}$$

The roll estimate is then used to rotate gyroscope signals in the y–z plane via a standard rotation matrix (Fig. 6):

$$\begin{bmatrix} \hat{g_y} \\ \hat{g_z} \end{bmatrix} = \begin{bmatrix} \cos(roll) & -\sin(roll) \\ \sin(roll) & \cos(roll) \end{bmatrix} \begin{bmatrix} g_y \\ g_z \end{bmatrix}. \tag{8}$$

This transformation cancels the influence of wrist orientation, yielding corrected signals $\hat{g_y}$ and $\hat{g_z}$ that are consistently aligned with global horizontal and vertical axes. By integrating these normalized angular velocities, we obtain yaw and pitch trajectories that are invariant to the user's initial arm pose.

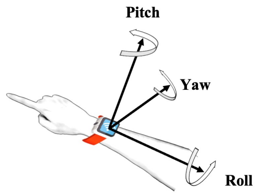

Fig. 6: Smartwatch coordinate system.

2) Single-stroke Gesture Recognition: Building on motion normalization, `Wandatch` recognizes five single-stroke gestures—**UP**, **DOWN**, **LEFT**, **RIGHT**, and **ROTATE** (Fig. 7). To improve robustness, gestures are formulated as out-and-return movements rather than single displacements. This constraint induces closed convex or concave patterns in gyroscope traces (Fig. 8), enabling reliable separation from casual hand motions or low-frequency drift. `Wandatch` detects such gestures by analyzing gyroscope trajectories for monotonic rotations: a valid gesture requires a sufficient excursion in one direction followed by a comparable return.

A key challenge is **hand asymmetry**: the same motion (e.g., raising the arm) produces inverted yaw or pitch traces when the watch switches hands. To address this, `Wandatch` fuses gyroscope trajectories with accelerometer cues. Upward motion consistently causes a transient increase in the z-axis acceleration, whereas downward motion yields the opposite. Leveraging these inertial signatures ensures consistent UP/DOWN recognition across both hands while preserving high discriminability against incidental arm movements.

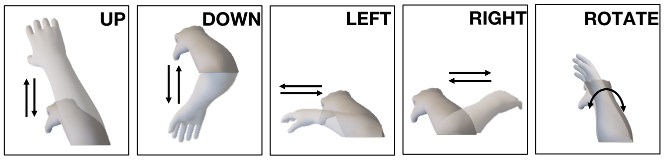

Fig. 7: The sketches of five single-stroke gestures.

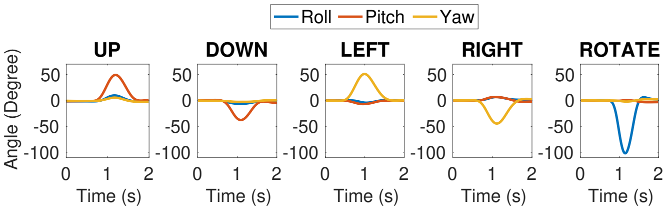

Fig. 8: Gyroscope readings for single-stroke gestures with the watch on the right hand.

3) Multi-stroke Gesture Recognition: While single-stroke gestures provide robust atomic commands, real interaction demands more expressive input. By composing multiple strokes, `Wandatch` supports fine-grained symbolic gestures, but this introduces two key challenges: mitigating transition movements between consecutive strokes and ensuring user-independent recognition under diverse writing styles.

- **Transition-Resilient Segmentation.** Energy-based methods [15], [28] forces users into awkward pauses and still misclassify transition movements as part of a gesture. We propose a *transition-resilient segmentation strategy* that segment gestures into strokes, treating the final stroke as a natural anchor (Fig. 9). All strokes (up to ten) are first fed to the classifier; if recognition fails, earlier strokes are progressively pruned until a valid result is obtained. This stroke-driven design preserves semantics while stripping away noisy transitions.
- **User-Independent Gesture Classification.** Raw IMU signals are unstable—affected by writing speed, amplitude, and stroke order—making user-independent recognition difficult. Existing approaches [4], [15], [28] rely either on spatial trajectory maps, which capture coarse shapes but ignore stroke order, or on temporal sequences, which preserve ordering but remain fragile to speed and amplitude variations. To overcome these limitations, we introduce a *dual-modality representation*: spatial maps (32 × 32) provide shape-level invariance across users, while temporal sequences (32 × 2) preserve ordering cues to resolve confusable digits (e.g., "5" vs. "6"). These complementary views are fused through a lightweight CNN–GRU hybrid (Fig. 10), enabling robust, user-independent recognition under real-world variability while remaining efficient enough for on-device inference.

VI. Evaluation

A. Implementation

We implement `Wandatch` as a runtime-efficient application on a COTS smartwatch (TicWatch Pro 2020, WF12106) with a 1.2 GHz quad-core CPU, 512 MB RAM, 415 mAh battery, and Wear OS. Using system APIs, we collect motion data from the built-in 6-axis IMU (LSM6DSM) and 3-axis magnetometer (MMC3630), and leverage the smartwatch's loudspeaker and microphone to emit 17–23 kHz chirps and capture echoes.

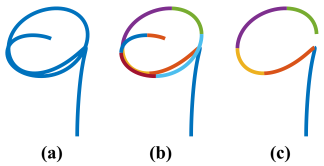

Fig. 9: Writing '9' produces a raw trajectory (a). Stroke segmentation (b) exposes its structure, and pruning redundant segments (c) yields cleaner input for recognition.

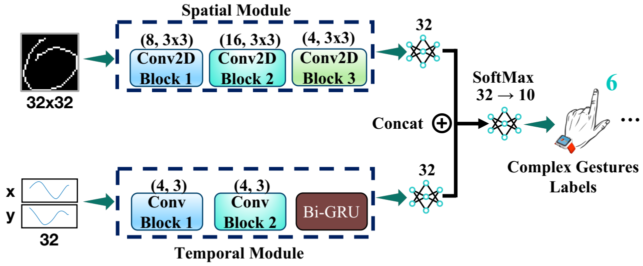

Fig. 10: Architecture for multi-stroke gesture recognition.

The smart speaker prototype combines an 7-microphone array (UMA-8, 4.15 cm radius) and a loudspeaker (EDIFIER R10U), with all computation performed in real time on a Raspberry Pi 4B. Communication with the smartwatch is handled via Wi-Fi. For evaluation, we control four Xiaomi smart appliances (air purifier, tower fan, bedside lamp, and table lamp) using the MiIO protocol, implemented on Android through the open-source python-miio library. [1]

B. System Performance on Appliance Selection

1) Setup: We evaluated appliance selection in two office deployment. In the sparse setup, 5 appliances were deployed across a $5 \times 10m$ office (Fig. 11), while in the dense setup, the same appliances were arranged in an "L" shape with intervals from 100 cm down to 40 cm (Fig. 12).

2) Performance of `Wandatch`: Results show that `Wandatch` achieves an average selection accuracy of 97.6% in the sparse setting (Fig. 13) and 94.1% in the dense scenario. Even when the spacing shrinks to 40 cm, accuracy remained above 82.4% (Fig. 14). These results demonstrate that `Wandatch` maintains high selection accuracy under both sparse and dense deployments, with only slight degradation in crowded settings.

C. Interaction Efficiency of `Wandatch`

1) Setup: We compared `Wandatch` with smartphone apps with voice control in terms of interaction time. Table I gives the steps to control appliances with Wandatch. Ten participants controlled three appliances (air purifier, lamp, and tower fan) from six positions (P1–P6), performing each task 12 times

[1] https://github.com/rytilahti/python-miio

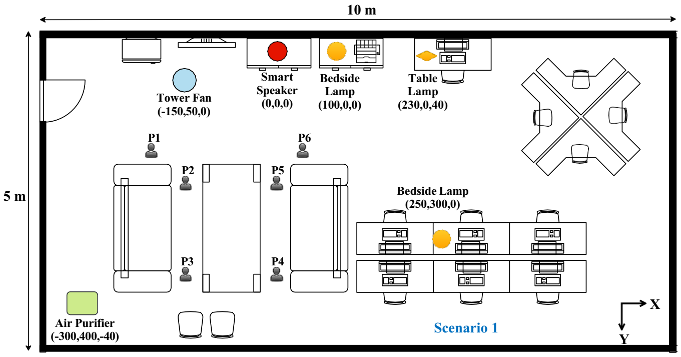

Fig. 11: Sparse deployment of appliances in a 5 × 10 m office.

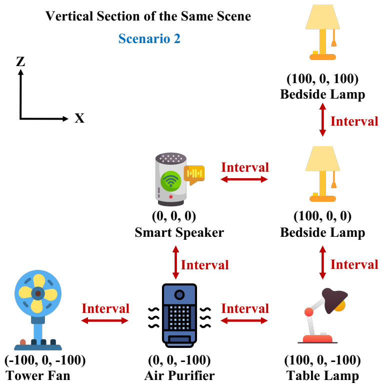

Fig. 12: Dense deployment of appliances with equal spacing.

per method. For Wandatch, participants used activation and control gestures; for the app, they navigated the phone UI; and for the voice, they issued commands until recognition was complete. The interaction time was measured from the start of each method to task completion, including retries when recognition failed. Tasks included turning on the air purifier, adjusting the lamp brightness, and setting the fan speed.

2) Performance: Wandatch achieved a 91.6% success rate (without retries) and consistently outperformed the other methods, reducing the interaction time by 12.6%–50.1% (Fig. 15). These results demonstrate that Wandatch enables faster, more intuitive appliance control, with advantages that become more pronounced as the number of devices increases.

D. Energy Efficiency of Wandatch

Since smartwatch battery capacity is limited, we measured Wandatch's power consumption using Google Battery Historian under 5 states (Fig. 16): (1) idle with display on, (2) activation detection only, (3) full Wandatch pipeline, (4) playing the game 2048 [29], and (5) GPS positioning with Sogou Map. Based on the 415 mAh battery and 3.7 V voltage, the power consumption are 135.48 mW, 290.93 mW, 552.78 mW, 386.56 mW, and 690.97 mW, respectively. Thus, activation detection consumes less power than 2048 by 95.63 mW, while running the full pipeline requires an additional 261.85 mW. In practical use, activation detection is only enabled when the display is already on to save battery.

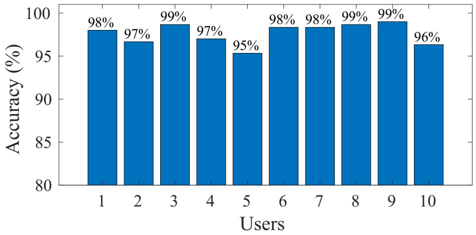

Fig. 13: The appliance selection accuracy of different users.

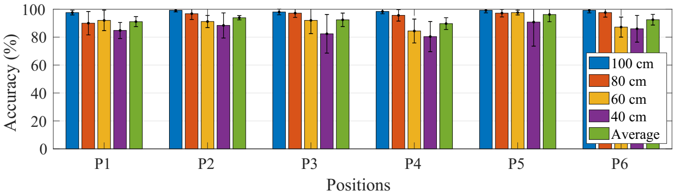

Fig. 14: The device selection accuracy at different positions.

Overall, Wandatch achieves high efficiency, consuming less than half the power of GPS positioning.

E. Performance on Indoor Positioning

1) Setup: We evaluate positioning accuracy in two scenarios: an open area without furniture and an office room with dense equipment. In open area (7.2 $m\times$ 6.0 m), we divide the space into 120 grids of 60 cm, place the speaker at the center, and set the smartwatch at 50 cm, 100 cm, and 150 cm at each grid center, yielding 360 positions. In office (6.0 $m\times$ 4.0 m), we use 24 grids of 100 cm with the speaker at (300 cm, 0, 0) and the same heights, giving 72 positions. The smartwatch is mounted on a tripod, and 20 samples are collected per position, resulting in 8,640 samples in total.

2) Performance: Figs. 17 and 18 present Wandatch's 3D positioning errors in two LoS scenarios. Median errors are 0.16/0.20/0.50 m (x/y/z) in open space and 0.17/0.24/0.50 m in a multipath-rich office. In contrast, state-of-the-art voice-based systems with a single microphone array report 2D errors of 0.44 m [30] and 0.59 m [31]. Thus, Wandatch cuts errors by more than half while supporting full 3D positioning. Errors are larger along z, but 2D suffices for appliance selection; 3D becomes essential when onboarding new devices. To explore how performance varies with configuration changes, Fig. 19 shows the localization CDFs in the open environment with 3 to 7 microphones. The error consistently decreases as more microphones are added, with the most significant gain observed when moving from 3 to 4 mics. Beyond 5 microphones, the performance begins to plateau, indicating diminishing returns in accuracy.

3) Effect of Watch-to-Speaker (WTS) Distance: Intuitively, positioning accuracy degrades as the WTS distance increases due to lower SNR. To quantify this effect, we vary the WTB distance from 75 cm to 475 cm in 50 cm increments in the open environments. As shown in Fig. 20, positioning error grows proportionally with distance, with the median error exceeding 55 cm at 475 cm.

TABLE I: Gesture-based appliance control with `Wandatch`.

Step	Gesture	Function
	Common Activation Step for All Appliances	
1	Double tapping	Activate Wandatch & select appliance
	Air Purifier	
2	Single tapping	Turn on
	Lamp	
2	Double tapping	Enter brightness adjustment mode
3	Raise up his/her hand	Increase brightness
4	Double tapping	Exit brightness adjustment mode
	Tower Fan	
2	UP	Adjust wind speed: Gear 1 → 2
3	UP	Adjust wind speed: Gear 2 → 3

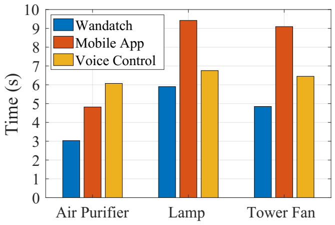

Fig. 15: Time consumption of different interaction methods.

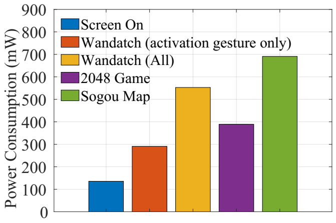

Fig. 16: Comparison of power consumption on smartwatch.

4) Effect of Non-Line-of-Sight (NLoS): In practical scenarios, LoS between the smartwatch and speaker may be blocked by objects or the human body. To evaluate this effect, we conduct experiments in the open environment with a user standing next to the watch at a fixed 2 m distance from the speaker. The user's orientation is varied from 0° to 315° in 45° steps (Fig. 21 (a)). Fig. 21 (b) shows that median positioning errors increase from 0.17 m at 0° to 1.24 m at 180° (NLoS).

F. Performance on Gesture Recognition

1) Setup: We evaluate gesture recognition on 1 micro gesture and 15 macro gestures (5 single-stroke and 10 multi-stroke). We recruited 10 participants (3 female; age 21–27) for controlled gesture collection. Each participant performed the activation gesture 100 times and every other gesture 30 times, resulting in 5,500 samples. This dataset is used to evaluate recognition accuracy across gesture categories, and the same participants also took part in the user study (Section VI-G).

To assess the robustness of the activation gesture, we conducted two additional experiments. In an office scenario, each participant interacted with 5 appliances (Fig. 11) and performed the activation gesture 100 times (20 per appliance) with reverse-counterbalanced order. We also collected 48 hours of daily-life data from 4 participants (1 female; age 23–27) who wore a smartwatch during normal activities. These traces are used to evaluate the false alarm rate (FAR). All participants were informed of the data collection and provided consent.

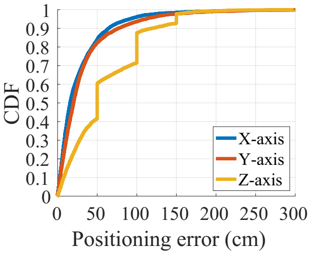

Fig. 17: The CDF of 3D positioning errors in open environment.

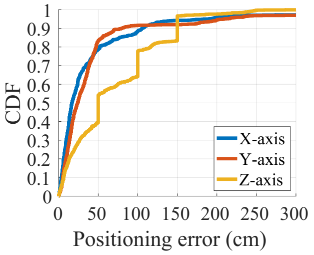

Fig. 18: The CDF of 3D positioning errors in a normal office room.

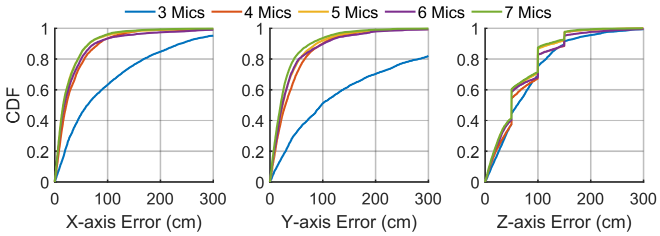

Fig. 19: Localization error CDFs under varying microphone array configurations.

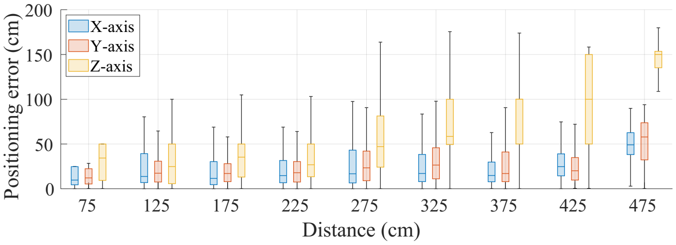

Fig. 20: Positioning error along X, Y, and Z axes as WTS distance increases.

2) Performance of Micro Gesture Recognition: On average, double-tap recognition achieves 92.3% accuracy with a false alarm rate of 0.97/hour. However, accuracy varies widely across participants (82.0%–100%), mainly due to individual behavioral and physiological differences and inconsistent tapping speeds. While requiring an extra double-tap is acceptable in terms of time, improving user-independent recognition remains critical for a better user experience.

3) Performance of Single-stroke Gesture Recognition: Each participant performed five single-stroke gestures (UP, DOWN, LEFT, RIGHT, and ROTATE) for 30 rounds in reverse-counterbalanced order. As shown in Fig. 22 (a), `Wandatch` achieves an average recognition accuracy of 99.67%, with the few errors caused by motion artifacts contaminating UP gestures. These results indicate that `Wandatch` provides highly reliable recognition of directional gestures, though accuracy may degrade when users are in motion.

4) Performance of Multi-stroke Gesture Recognition: Each participant wrote digits '0' to '9' in the air for 30 rounds in reverse-counterbalanced order. We evaluated recognition using leave-one-out validation, where one participant's samples for testing and the rest for training. `Wandatch` achieves an average accuracy of 95.57% (Fig. 22 (b)). Most errors come from confusion between '0' and '6', with about 10% of '0' samples misclassified as '6'.

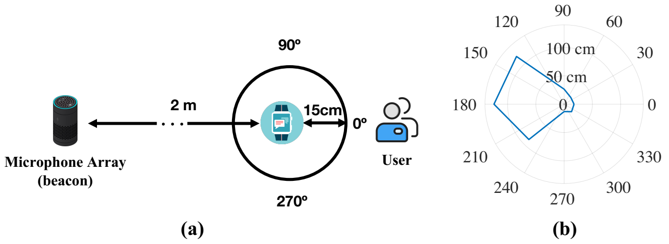

Fig. 21: (a) The NLoS experiment setting; (b) The comparison of NLoS and LoS conditions.

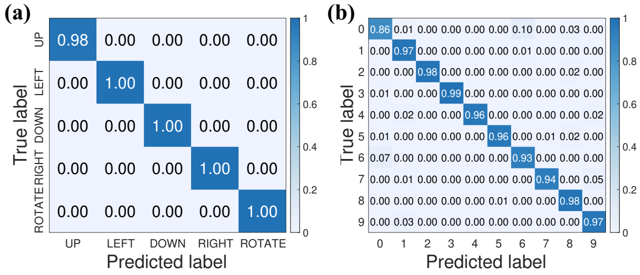

Fig. 22: Confusion matrices of (a) single-stroke and (b) multi stroke gesture.

G. User Experience Analysis

After the interaction experiments, participants completed a questionnaire covering five usability aspects (ease of control, speed, memorability, enjoyment, and preference) and five NASA-TLX subscales [32] (excluding temporal demand due to overlap). In total, ten questions were asked. Results (Fig. 23) show that Wandatch received the highest ratings on seven of the ten factors. Users preferred Wandatch for its speed and low mental demand. Regarding the higher physical demand score, post-hoc interviews revealed this was an artifact of the dense experimental protocol (repeated continuous gestures). Participants noted that for sporadic real-world use, the physical effort was negligible compared to locating a remote.

VII. Discussion and Future Work

- **Usage Scenarios and Necessity.** Beyond performance metrics, Wandatch addresses specific contexts where current modalities fail: (1) Target Ambiguity: Unlike voice, Wandatch uses a "wrist ray" to distinguish identical appliances via spatial context; (2) Environmental Noise: High-frequency acoustic sensing remains robust in noisy rooms where voice assistants fail; and (3) Touchless Utility: In situations involving dirty or wet hands (e.g., cooking), Wandatch enables hygiene-friendly control without touching screens or physical switches.
- **Scalable Design Framework.** To address the burden of memorizing gestures for many devices, we propose a Unified Interaction Framework: micro-gestures (taps) for binary control (on/off), single-stroke gestures for discrete modes, and arm tracking for continuous adjustments. Furthermore, to support scalability, Wandatch supports a Hybrid Interaction Model where the smartwatch screen provides a fail-safe UI with gesture visualizations for unfamiliar tasks. This approach preserves pointing as the primary selector while leveraging the screen to bridge memory gaps.

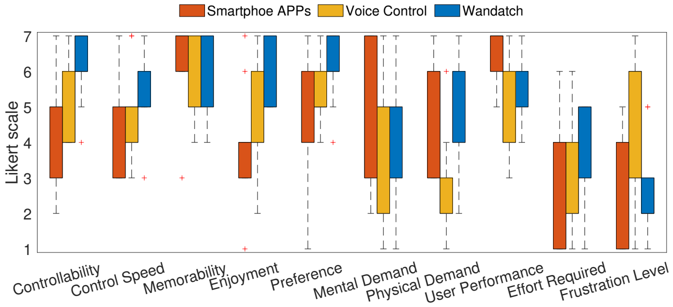

Fig. 23: The questionnaire result of 7-point Likert Scale. 1 represents 'Low/Negative' (e.g., Low Mental Demand), and 7 represents 'High/Positive' (e.g., High Control Speed).

- **Limitations and Privacy.** Current performance relies on Line-of-Sight (LoS); future work will exploit signal reflections to improve NLoS robustness [20], [30], [31]. Regarding privacy, Wandatch is secure by design: microphones are only activated briefly during the explicit activation gesture, and the system relies solely on inaudible ultrasonic chirps rather than speech content.

VIII. Conclusion

In this work, we revisited the long-standing vision of effortless *point-and-command* interaction for smart homes. We introduced Wandatch, a system that transforms commodity smartwatches and speakers into a virtual magic wand, eliminating the need for cameras, anchors, or bespoke hardware. To overcome two fundamental challenges—(**C1**) localization burdened by hardware overhead and site-specific training, and (**C2**) gesture recognition locked in the brittle "gesture–pause–gesture" routine across users and contexts—Wandatch contributes two key designs: *Physics-driven Localization*, which unifies acoustic ranging with IMU orientation to project a virtual wrist ray for appliance selection, and *Universal Gesture Interaction*, which recognizes the full spectrum of micro- and macro-gestures for seamless control. Experimentally, Wandatch achieves 97.6% selection accuracy in sparse deployments and sustains 85.5% under dense 40 cm spacing, while reducing interaction time by 12.6–50.1% over apps and voice, and significantly improving usability across ten sub-scales. Overall, Wandatch demonstrates that natural, infrastructure-free point-and-command interaction is not only feasible but practical, paving the way for the next generation of human–IoT interfaces.

IX. Acknowledgement

This research is partially supported by China NSFC Grant (62472366), the Project of DEGP (No.2023KCXTD042, 2024GCZX003), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007), "111 Center (No.D25008)", Shenzhen Science and Technology Foundation (ZDSYS20190902092853047).

References

[1] S. Sinha, "State of iot 2024: Number of connected iot devices growing 13% to 18.8 billion globally," https://iot-analytics.com/number-connected-iot-devices/, 2024, last accessed September 3, 2025.

[2] M. Kim, A. Glenn, B. Veluri, Y. Lee, E. Gebre, A. Bagaria, S. Patel, and S. Gollakota, "Iris: Wireless ring for vision-based smart home interaction," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3654777.3676327

[3] J. Li, Q. Yang, K. Xu, Y. Zhang, and C. Xu, "Echosight: Streamlining bidirectional virtual-physical interaction with in-situ optical tethering," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3713925

[4] Z. He, Z. Wang, C. Yu, C. Zhang, X. Shen, and Y. Shi, "Writingring: Enabling natural handwriting input with a single imu ring," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, ser. CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3706598.3714066

[5] M. Budde, M. Berning, C. Baumgärtner, F. Kinn, T. Kopf, S. Ochs, F. Reiche, T. Riedel, and M. Beigl, "Point & control – interaction in smart environments: You only click twice," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, ser. UbiComp '13 Adjunct. New York, NY, USA: Association for Computing Machinery, 2013, p. 303–306. [Online]. Available: https://doi.org/10.1145/2494091.2494184

[6] D. Schmidt, D. Molyneaux, and X. Cao, *PICOntrol: Using a Handheld Projector for Direct Control of Physical Devices through Visible Light*. New York, NY, USA: Association for Computing Machinery, 2012, p. 379–388. [Online]. Available: https://doi.org/10.1145/2380116.2380166

[7] M.-S. Pan and C.-J. Chen, "Intuitive control on electric devices by smartphones for smart home environments," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4281–4294, 2016. [Online]. Available: https://doi.org/10.1109/JSEN.2016.2542260

[8] A. Alanwar, M. Alzantot, B.-J. Ho, P. Martin, and M. Srivastava, "Selecon: Scalable iot device selection and control using hand gestures," in *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*, ser. IoTDI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 47–58. [Online]. Available: https://doi.org/10.1145/3054977.3054981

[9] J. J. Yang, G. Banerjee, V. Gupta, M. S. Lam, and J. A. Landay, "Soundr: Head position and orientation prediction using a microphone array," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376427

[10] A. Ward, A. Jones, and A. Hopper, "A new location technique for the active office," *IEEE Personal Communications*, vol. 4, no. 5, pp. 42–47, 1997. [Online]. Available: https://doi.org/10.1109/98.626982

[11] P. Lazik, N. Rajagopal, O. Shih, B. Sinopoli, and A. Rowe, "Alps: A bluetooth and ultrasound platform for mapping and localization," in *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 73–84. [Online]. Available: https://doi.org/10.1145/2809695.2809727

[12] W. Mao, J. He, and L. Qiu, "Cat: high-precision acoustic motion tracking," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 69–81. [Online]. Available: https://doi.org/10.1145/2973750.2973755

[13] A. Wang and S. Gollakota, "Millisonic: Pushing the limits of acoustic motion tracking," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–11.

[14] Q. Lin, Z. An, and L. Yang, "Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices," in *The 25th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3300061.3300139

[15] Y. Yin, L. Xie, T. Gu, Y. Lu, and S. Lu, "Aircontour: Building contour-based model for in-air writing gesture recognition," *ACM Trans. Sen. Netw.*, vol. 15, no. 4, oct 2019.

[16] X. Lin, Y. Chen, X.-W. Chang, X. Liu, and X. Wang, "SHOW: Smart Handwriting on Watches," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 1–23, jan 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3161412

[17] A. Wilson and S. Shafer, "Xwand: Ui for intelligent spaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 545–552.

[18] J.-W. Yoo, Y.-W. Jeong, Y. Song, J. Lee, S.-H. Lim, K.-W. Park, and K. H. Park, "Ithrow : A new gesture-based wearable input device with target selection algorithm," in *2007 International Conference on Machine Learning and Cybernetics*, vol. 4, 2007, pp. 2083–2088.

[19] J. Wu, G. Pan, D. Zhang, S. Li, and Z. Wu, "Magicphone: Pointing & interacting," in *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing - Adjunct*, ser. UbiComp '10 Adjunct. New York, NY, USA: Association for Computing Machinery, 2010, p. 451–452.

[20] H. Murakami, T. Sasatani, M. Sugimoto, I. Sukeda, Y. Mita, and Y. Kawahara, "Syncecho: Echo-based single speaker time offset estimation for time-of-flight localization," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 718–729.

[21] J. Hou, X.-Y. Li, P. Zhu, Z. Wang, Y. Wang, J. Qian, and P. Yang, "Sign-speaker: A real-time, high-precision smartwatch-based sign language translator," in *The 25th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '19. New York, NY, USA: Association for Computing Machinery, 2019.

[22] W. Chen, L. Chen, M. Ma, F. S. Parizi, S. Patel, and J. Stankovic, "ViFin: Harness Passive Vibration to Continuous Micro Finger Writing with a Commodity Smartwatch," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–25, mar 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3448119

[23] Q. Zhang, J. Jing, D. Wang, and R. Zhao, "Wearsign: Pushing the limit of sign language translation using inertial and emg wearables," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 1, Mar. 2022. [Online]. Available: https://doi.org/10.1145/3517257

[24] S. Goldheart, "Amazon echo teardown," https://www.ifixit.com/Teardown/Amazon+Echo+Teardown/33953, 2014, last accessed December 15, 2025.

[25] J. Suovanen, "Homepod teardown," https://www.ifixit.com/Teardown/HomePod+Teardown/103133, 2018, last accessed December 15, 2025.

[26] Y. Jiang and V. C. Leung, "An asymmetric double sided two-way ranging for crystal offset," in *2007 International Symposium on Signals, Systems and Electronics*, 2007, pp. 525–528.

[27] N. Huang, Z. Shen, S. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. Yen, C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903 – 995, 1998.

[28] Y. Luo, J. Liu, and S. Shimamoto, "Wearable air-writing recognition system employing dynamic time warping," in *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, 2021, pp. 1–6.

[29] fast soft, "2048 for smart watch - apps on google play," https://play.google.com/store/apps/details?id=com.fastsoft.game2048w, 2025, last accessed August 4, 2025.

[30] S. Shen, D. Chen, Y.-L. Wei, Z. Yang, and R. R. Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '20. New York, NY, USA: Association for Computing Machinery, 2020.

[31] M. Wang, W. Sun, and L. Qiu, "MAVL: Multiresolution analysis of voice localization," in *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*. USENIX Association, Apr. 2021, pp. 845–858.

[32] NASA, "Nasa tlx task load index," https://humansystems.arc.nasa.gov/groups/tlx/, 2020, last accessed September 5, 2025.