EchoMotion: Towards Full-Body Pose Tracking with a Single Smartwatch Using Active Acoustic Sensing

SAIF MAHMUD*, Cornell University, USA
JIWAN KIM, KAIST, Republic of Korea
TIANHONG CATHERINE YU, Cornell University, USA
QIKANG LIANG, Cornell University, USA
VINEET PARIKH, Cornell University, USA
FRANCOIS GUIMBRETIERE, Cornell University, USA
CHENG ZHANG, Cornell University, USA

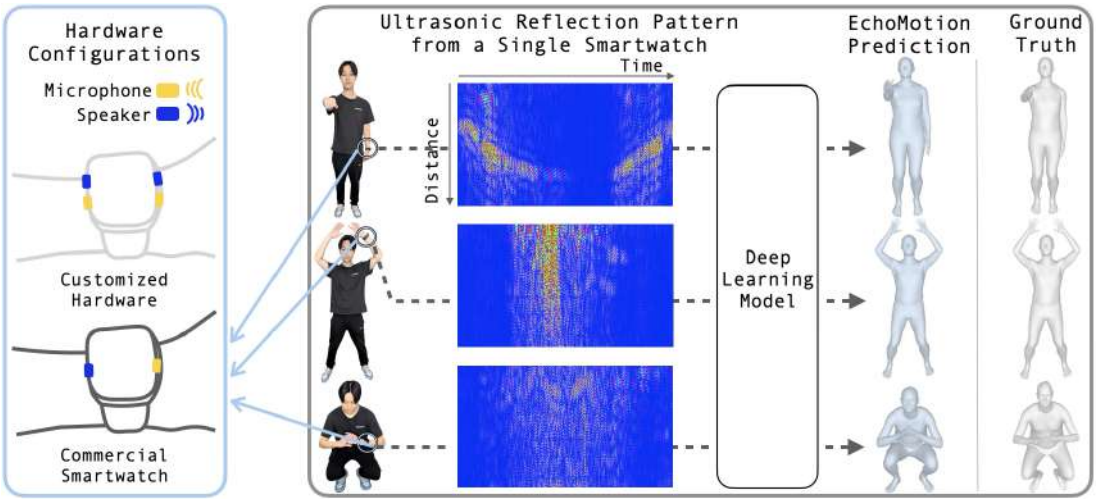


Fig. 1. EchoMotion features a single smartwatch for full-body pose tracking using active acoustic sensing. Ultrasonic reflection patterns from a single smartwatch are inputs to a customized deep learning model that estimates full-body poses in three dimensions. In this paper, as shown on the left, we demonstrated the feasibility with 2 hardware configurations: (1) a customized hardware with 2 pairs of speakers and microphones, and (2) an off-the-shelf smartwatch with only 1 speaker-microphone pair.

*Corresponding author.

Authors' Contact Information: Saif Mahmud (corresponding author), Cornell University, Ithaca, NY, USA, sm2446@cornell.edu; Jiwan Kim, KAIST, Daejeon, Republic of Korea, jiwankim@kaist.ac.kr; Tianhong Catherine Yu, Cornell University, Ithaca, NY, USA, ty274@cornell.edu; Qikang Liang, Cornell University, Ithaca, NY, USA, ql75@cornell.edu; Vineet Parikh, Cornell University, Ithaca, NY, USA, vap43@cornell.edu; Francois Guimbretiere, Cornell University, Ithaca, NY, USA, francois@cs.cornell.edu; Cheng Zhang, Cornell University, Ithaca, NY, USA, chengzhang@cornell.edu.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
© 2026 Copyright held by the owner/author(s).
ACM 2474-9567/2026/9-ART139
https://doi.org/10.1145/3832001

Full-body pose tracking on wearables enables fitness tracking, rehabilitation, and activity logging on the go, but remains challenging as it requires capturing movements across the torso, arms, legs, and head. Existing solutions often rely on specialized suits or multiple sensors, limiting practicality. We present EchoMotion, an active acoustic sensing system that enables full-body pose tracking using only a wrist-worn smartwatch. By leveraging the built-in speaker and microphone to emit inaudible ultrasonic signals and process their reflections, EchoMotion employs a lightweight deep learning pipeline to infer the 3D positions of 12 body joints. We first validated the approach with 10 participants using a custom hardware prototype, achieving a mean per-joint position error (MPJPE) of 5.36 cm and rotation error (MPJRE) of 11.68°. We then confirmed feasibility with 15 participants on a commercial smartwatch (MPJPE 5.99 cm, MPJRE 11.92°). Finally, in a semi-in-the-wild study with 10 participants, EchoMotion maintained comparable performance (MPJPE 6.41 cm, MPJRE 13.18°), suggesting its potential for future real-world use.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

ACM Reference Format:
Saif Mahmud, Jiwan Kim, Tianhong Catherine Yu, Qikang Liang, Vineet Parikh, Francois Guimbretiere, and Cheng Zhang. 2026. EchoMotion: Towards Full-Body Pose Tracking with a Single Smartwatch Using Active Acoustic Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 139 (September 2026), 36 pages. https://doi.org/10.1145/3832001

1 INTRODUCTION

Full-body pose tracking using everyday wearable devices offers a promising pathway to reliably capturing the movement of different body parts without compromising mobility. This capability underpins a wide range of applications, from fitness monitoring [42, 112] and activity analysis [57] to rehabilitation [70] and AR/VR interaction [6, 67]. For instance, accurately capturing a user's pose during a workout not only enables precise exercise tracking but also supports personalized feedback from intelligent assistants. However, requiring users to wear a specialized suit or carry multiple devices for full-body tracking is impractical for on-the-go use.

Smartwatches stand among the most popular wearable devices, offering capabilities to count steps, measure vital signs, and monitor sleep. Despite these strengths, a single off-the-shelf smartwatch, with its current sensing capabilities, struggles to track full-body pose reliably. The primary limitation stems from the nature of Inertial Measurement Units (IMUs) - the predominant sensing technology in modern smartwatches due to their ubiquity and low cost. These IMU-based systems can only sense movement at their direct point of instrumentation. Consequently, full-body pose tracking traditionally requires additional data from other devices (smartphones [3, 68], earbuds [68], or VR headsets [36]) using IMUs or alternative sensing modalities like ultrawideband (UWB) [9, 22]. This challenge is compounded by the typical placement of smartwatches on the passive or non-dominant hand, making it particularly difficult to track the movements of multiple active body parts.

While integrating high-fidelity sensors such as cameras [55] or mm-wave radars [106] into smartwatches could theoretically enhance tracking capabilities - as these technologies have proven effective in stationary settings - such additions present significant drawbacks. These include increased computational and power demands, delayed mass deployment due to hardware modifications, and perhaps most critically, substantial privacy concerns with body-worn cameras.

In this paper, we aim to address the following research question:

- *Is it possible to track full-body pose on a single wrist-mounted device using a low-power and privacy-aware sensing approach, with the potential to be deployed on future smartwatches?*

A positive answer could allow seamless, on-the-go full-body pose estimation without the need for specialized suits or multiple devices. Such a capability could unlock a wide range of new applications, from real-time feedback for fitness and rehabilitation exercises to precise logging of daily activities, using only a single wrist-mounted device.

To address this research question, we present EchoMotion, an intelligent active acoustic sensing system on a wrist-mounted device that can track full-body poses continuously. It only requires microphones and speakers, which are lightweight, minimally obtrusive, privacy-aware, and also widely available in modern smartwatches/wristbands. Our system transmits inaudible encoded ultrasonic waves from the speaker and analyzes their reflections coming from the torso, arms, legs, and head captured by the microphone. Through a biomechanics-informed lightweight neural network capable of extracting movement and posture information of different body parts from the reflected acoustic signals, EchoMotion can infer the 3D positions of 12 body joints on the head, legs, torso, and arms.

To evaluate the performance of our proposed method, we conducted user studies with 35 participants across two implementation phases: (1) a customized wristband with two pairs of microphones and speakers, and (2) a commercial off-the-shelf (COTS) smartwatch, the Samsung Galaxy Watch 6 [84], which has only one built-in speaker–microphone pair. In the first phase, we built a prototype wristband with two pairs of acoustic sensors embedded in a 3D-printed watch replica. This prototype achieved a mean per-joint position error (MPJPE) of 5.36 cm and a mean per-joint rotation error (MPJRE) of 11.68° when predicting 3D coordinates for 12 full-body joints across 10 participants performing 18 different exercises. Building on these promising results, we developed a software-only implementation of our sensing system for the Samsung Galaxy Watch 6 and evaluated it with another 15 participants. Despite relying on only one built-in speaker–microphone pair on this off-the-shelf watch, this implementation achieved an MPJPE of 5.99 cm and an MPJRE of 11.92° in a user-adaptive evaluation, demonstrating the potential feasibility of deploying our approach on a single COTS smartwatch.

To further evaluate robustness in naturalistic settings, we conducted a semi-in-the-wild study with 10 additional participants. Unlike controlled lab experiments, this study introduced real-world variability, including diverse room and furniture layouts, human movement in a building hallway, and environmental acoustic noise near a roadside curb. Under these conditions, our user-independent evaluation achieved an MPJPE of 6.41 cm and an MPJRE of 13.18°. Building on EchoMotion, we further demonstrate a potential application scenario of exercise tracking on a COTS smartwatch and evaluated it in a pilot study with 5 participants performing gym workouts, achieving a 95.4% F1-score. Together, these results highlight the promise of EchoMotion's active acoustic sensing for enabling full-body pose tracking from a single wrist-mounted device, while also underscoring remaining challenges for real-world deployment. We discuss these challenges in detail, including generalizability across smartwatch hardware, participants, and diverse body movements.

The key contributions of this work are as follows:

- We are the first to demonstrate the feasibility of full-body pose tracking using only acoustic sensing from a single wrist-mounted device worn on the non-dominant hand.
- We are the first to show that full-body pose tracking can be achieved on an unmodified commercial off-the-shelf smartwatch, using only its built-in speaker–microphone pair and a software-only implementation, suggesting the potential for future deployment on widely available devices.
- We conducted user studies with 35 participants across both a customized wristband prototype and a COTS smartwatch, evaluating performance in diverse conditions, including controlled laboratory settings and semi-in-the-wild environments with varied room and furniture layouts, nearby human movement in a building hallway, and ambient acoustic noise at a roadside curb, and demonstrated the feasibility of tracking 12 body joints.

2 RELATED WORK

In this section, we review existing works related to full-body pose estimation. We will first discuss techniques utilizing external infrastructure for full-body pose tracking, then examine tracking systems implemented on

wearables, beginning with the most popular sensor type, Inertial Measurement Units (IMUs), and subsequently exploring alternative sensing modalities.

Table 1. Comparison of full-body pose tracking systems implemented on wearable form factors. Note that the Mean Per Joint Position Error (MPJPE) and Mean Per Joint Rotation Error (MPJRE) are not calculated on the same dataset or following the same train-test scheme.

System	Sensing Modality	Sensor Placement	# of Instr. Points	Consumer Device	MPJPE (cm) ↓	MPJRE (°) ↓
SIP [96]	IMU	Forearms, Lower Legs, Head, Waist	6	×	6.7	8.7
DIP [32]	IMU	Forearms, Lower Legs, Head, Waist	6	×	7.3	15.1
TransPose [108]	IMU	Forearms, Lower Legs, Head, Waist	6	×	7.1	8.8
BodyTrak [55]	RGB Camera	Wristband	1	×	6.9	-
AvatarPoser [36]	IMU	VR Headset, Handheld Controllers	3	✓	4.1	3.2
IMUPoser [68]	IMU	Smartwatch, Smartphone, Earbud	1-3	✓	9.7	23.9
MobilePoser [103]	IMU	Smartwatch, Smartphone, Earbud	1-3	✓	10.6	23.7
Ultra Inertial Poser [9]	IMU + UWB	Forearms, Lower Legs, Head, Waist	6	×	5.1	10.7
UltraPoser [53]	IMU + Ultrasound	Smartwatch, Smartphone, Earbud	3	✓	3.2	7.4
EchoMotion	Active Acoustic	Smartwatch	1	✓	5.3	11.7

2.1 Pose Estimation Using External Infrastructure

A wide range of approaches relies on external sensing modalities to track full-body pose. Commercial motion capture systems like OptiTrack [76] and Vicon [95] represent the gold standard in precision tracking. These systems use infrared cameras to track retroreflective markers placed on multiple body locations such as hands, ankles, and head. After a calibration step, they can achieve millimeter-level accuracy, making them ideal for movie special effects, gaming character animation, and scientific research. However, their immobility and high cost limit their adoption in everyday consumer scenarios such as fitness tracking, activity logging, and rehabilitation.

Consumer-grade systems, including Microsoft Kinect [1], LeapMotion [31, 93], and Intel RealSense [34], combine depth and infrared sensing with RGB camera input to track full-body pose. In parallel, researchers have explored deep learning–based computer vision approaches [15, 23, 29, 52, 77, 97, 102, 105, 114, 121] that estimate full-body pose from frames captured by monocular RGB cameras [37, 40, 79, 80, 109, 118, 125]. Despite their fine-grained accuracy, these vision-based systems often struggle with challenges such as occlusion and lighting variations. Moreover, because they depend on external camera infrastructure, they are not well-suited for everyday mobile applications.

Beyond visual sensing, external sensing modalities such as radio frequency (RF) signals [119, 120], mmWave radars [2, 46], acoustic sensing [89], WiFi signals [38, 81, 82], tactile sensing [16], pressure sensor [18, 60], and electromagnetic sensing [117] have been applied to full-body pose tracking. Built-in smartphone cameras and depth sensors also show promise in approximating human body movements [3, 11]. However, these external infrastructure-based systems typically require the installation of multiple devices in a controlled environment, making them unsuitable for mobile use. Additionally, the reliance on continuous camera recording not only imposes high computational and power costs but also raises significant privacy concerns in both public and private settings.

2.2 Pose Estimation Using Wearables

Compared to external infrastructure-based high-fidelity motion capture systems, wearable solutions enable more portable and flexible pose estimation without necessitating the instrumentation of the environment. Table 1 summarizes the full-body pose tracking systems implemented on different wearable devices.

2.2.1 Using Inertial Measurement Units (IMUs). The XSens [71] system, a commercially available solution, represents state-of-the-art technology that employs 17 IMUs embedded in a specialized suit to track a comprehensive range of full-body movements. This system relies on classical filtering-based approaches such as Complementary filters [12] or Kalman filters [66] to infer body pose. Similarly, Rokoko [83] smart suits utilize comparable sensing techniques with IMUs and sensor fusion to offer even finer-grained full-body pose tracking.

With the advent of efficient deep learning-based modeling of IMU data [25, 35, 65, 87, 92, 123], recent systems can now track full-body movements using fewer IMUs by processing multiple data frames simultaneously. Sparse Inertial Poser (SIP) [96], a seminal work in this field, utilized optimization-based methods to track full-body pose with six IMUs placed at the forearms, lower legs, head, and waist. Building upon SIP's success [96], subsequent systems, including Deep Inertial Poser (DIP) [32], TransPose [108], and Physical Inertial Poser (PIP) [107] implemented bidirectional Recurrent Neural Network (RNN) [30]-based modeling, yielding more precise pose predictions.

It should be noted that these systems predominantly rely on high-quality sensors providing higher frame rates, components typically absent from consumer mobile or wearable devices due to their higher cost. Addressing this limitation, IMUPoser [68] and MobilePoser [103] utilized data captured from everyday devices such as smartphones, smartwatches, and earbuds to produce accurate poses through efficient machine learning modeling and Inverse Kinematics-based joint rotation refinement. Similarly, AvatarPoser [36] captures motion input from the user's head and hands using a VR headset and handheld controllers. Furthermore, Loose Inertial Poser [126] integrates four IMUs into loose-fitting garments to accurately track human motion.

Beyond full-body tracking systems, several works have focused on more specialized applications. ArmTrak [88] uses a smartwatch IMU to track movements of the instrumented arm. SmartPoser [22] extends this by fusing ultrawideband (UWB) ranging data between a smartwatch and a pocket-placed smartphone with IMU data to capture finer-grained arm poses. Ultra Inertial Poser (UIP) [9] pushes further by combining UWB ranging with IMU data from six placement locations, including the forearms, lower legs, head, and waist, similar to multi-sensor setups like SIP [96], DIP [32], or PIP [107]. Finally, UltraPoser [53], conceptually closer to EchoMotion, leverages ultrasonic transmission and reception together with IMU data from consumer-grade smartwatches, smartphones, and earbuds to estimate full-body pose.

Despite being the most popular sensing modality in wearable devices and demonstrating efficiency in reliably tracking full-body pose, the primary limitation of IMU-based systems is their capacity to track only the points of instrumentation, thus requiring placing IMU sensors at multiple locations on the human body to achieve comprehensive full-body pose tracking. However, wearing multiple IMUs or wearables during everyday activities presents a higher barrier and is more challenging to deploy in real-world settings.

2.2.2 Using Sensing Modalities Other Than IMUs. Researchers have explored various sensing modalities to capture the movement and body postures on different body parts, beyond IMUs across a wide range of wearable form factors. For example, BodyTrak [55] leverages a miniature camera placed on a wristband to track full-body pose from a partial view of the body. Egocentric cameras have also been placed on hats [104], VR headsets [91, 99], handheld controllers [4], or chest mounts [33, 75] to capture full-body pose. However, akin to externally placed cameras (discussed in 2.1), body-worn cameras suffer in scenarios involving occlusion and low-light conditions and, more importantly, raise serious privacy concerns for both users and bystanders.

Although primarily designed for upper-body tracking rather than full-body tracking, systems such as SeamPose [111] and MoCaPose [122] utilize capacitive sensors embedded in everyday garments to capture upper-body pose. Similarly, electromagnetic sensors [41] and flex sensors [17] have been attached to tight-fitting clothing for human pose estimation. In another approach, MI-Poser [8] employs electromagnetic field (EMF) sensors on AR glasses and handheld devices to estimate upper-body pose, while PoseSonic [62] leverages active acoustic sensing on customized eyeglasses to track upper-body motion.

Given that smartwatches are among the most popular wearable devices, with a global user base exceeding 500 million [26] and widely regarded as the most acceptable form factor for digital health data sharing across age groups [86], it is noteworthy that sensing modalities beyond inertial measurement units (IMUs) have rarely been explored on smartwatches for full-body pose tracking. Existing wrist-mounted pose-tracking methods suffer from at least one of the following limitations: 1) they require multiple wearable devices beyond a single wristband [9, 22, 68]; 2) they rely on power-hungry sensors (e.g., cameras) that are impractical for continuous use [55]; or 3) they can only track the posture of the instrumented arm rather than full-body poses [88]. To the best of our knowledge, no prior work has explored active acoustic sensing on a single wristband form factor for full-body pose tracking.

3 DESIGN RATIONALE AND SENSING PRINCIPLE OF ECHOMOTION

Our goal in designing EchoMotion is to enable full-body pose tracking using a single wrist-mounted device (e.g., a smartwatch or wristband). This objective introduces two tightly coupled challenges. From a *design* standpoint, the system must remain minimally obtrusive, low-power, and privacy-preserving to be practical for everyday wearable use. From a *sensing* standpoint, a device placed on the non-dominant wrist must capture motion information from multiple body parts, including the arms, legs, torso, and head, despite being physically localized; many of these body parts are also partially or fully occluded from the device's line of sight by the body itself.

These constraints significantly narrow the design space. Introducing new sensing modalities such as cameras or mmWave radar is impractical: they are not commonly available on consumer smartwatches, increase power consumption, and, in the case of cameras, raise immediate privacy concerns. Moreover, substantial hardware modifications would compromise the form factor and usability of the device.

Inertial Measurement Units (IMUs), which are standard in smartwatches, satisfy the requirements of low power, unobtrusiveness, and privacy preservation. However, IMUs fundamentally capture local motion and therefore struggle to infer movements of distal body parts. This limitation makes them insufficient for full-body pose tracking from a single wrist-mounted location.

In contrast, active acoustic sensing has demonstrated the ability to capture movements of multiple body parts without requiring direct instrumentation, by leveraging reflections of acoustic signals. Prior work has validated this capability across a range of form factors, including eyeglasses [49, 50, 62, 64, 78, 115], earbuds [39, 51], wristbands [43, 47, 48, 61, 63], ring [54, 110] and robot platforms [56]. This makes it a compelling candidate for our setting.

3.1 Active Acoustic Sensing Primer

The key advantage of active acoustic sensing is its ability to infer spatial and motion information from a single point of instrumentation. Combined with the fact that speakers and microphones are already integrated into commodity smartwatches, this paradigm naturally aligns with our design constraints. EchoMotion builds on this principle to enable fine-grained, on-body sensing using only wrist-mounted hardware.

In active acoustic sensing, the device emits high-frequency acoustic signals, typically around 20 kHz, constrained by the Nyquist limit of wearable audio hardware, through a speaker, while a colocated microphone captures the reflected signals. These reflections encode information about the geometry and motion of the human body and

the surrounding environment. Importantly, acoustic signals travel not only along the direct path between the device and a body part, but also along indirect paths, reflecting off the floor, surrounding surfaces, and the body itself before returning to the microphone. The received signal therefore also carries information about body parts that are not in the device's direct line of sight.

EchoMotion adopts a C-FMCW [100] based sensing approach due to its strong performance in fine-grained, range-resolved tracking [39, 56, 62]. Compared to alternative approaches such as channel impulse response (CIR) with GSM sequences [113, 116] or Doppler-based methods [10], C-FMCW produces higher-resolution cross-correlation features, which are critical for distinguishing subtle movements of multiple body parts from a single sensing location.

Operationally, EchoMotion continuously emits ultrasonic C-FMCW chirps while simultaneously recording the reflected signals. The received signal is first filtered using a Butterworth bandpass filter to remove out-of-band noise. We then compute the cross-correlation between the transmitted chirp and the filtered received signal to estimate propagation delays. The strongest correlation peak corresponds to the direct path between the speaker and microphone and serves as a temporal reference. Additional peaks correspond to reflections from different body parts and environmental surfaces, enabling range-aware sensing relative to the wrist.

3.2 Signal Processing and Echo Profile

To extract motion information across different distances from the instrumentation point, we compute the cross-correlation between transmitted and received signals for each chirp. This process reveals phase shifts corresponding to changes in propagation distance. Each chirp produces an *echo frame*, where peaks at specific indices indicate motion within corresponding range bins.

For the C-FMCW [100] approach used in EchoMotion, the theoretical range R is given by:

$$R = \frac{C \cdot Lag}{2F_s} - vt \tag{1}$$

where C is the speed of sound (343 m/s at 20°C), F_s is the sampling frequency, Lag is the sample offset, v is the velocity of the reflecting body part, and t is elapsed time.

Following prior work [100], we approximate $v \approx 0$ for human motion, since even fast movements (e.g., up to 11.5 m/s [44]) are negligible compared to the speed of sound. As a result, the motion-induced term vt has minimal impact, and the range resolution simplifies to:

$$\delta R = \frac{C \cdot \delta Lag}{2F_s} \tag{2}$$

Since cross-correlation is computed at single-sample granularity, we set $\delta Lag = 1$.

As chirps are transmitted continuously, successive echo frames are stacked over time to form a range–time representation, referred to as the *echo profile*, where the horizontal axis represents time and the vertical axis represents distance from the sensor.

To isolate dynamic motion and suppress static reflections from the surrounding environment, we compute differences between consecutive echo frames, producing a *differential echo profile*. This representation is analogous to optical flow [13], capturing motion patterns across range bins while attenuating static components. It also mitigates artifacts caused by the motion of the wrist-mounted device itself. Finally, sliding windows of the differential echo profile are used as input to the EchoMotion deep learning pipeline (Section 4.2).

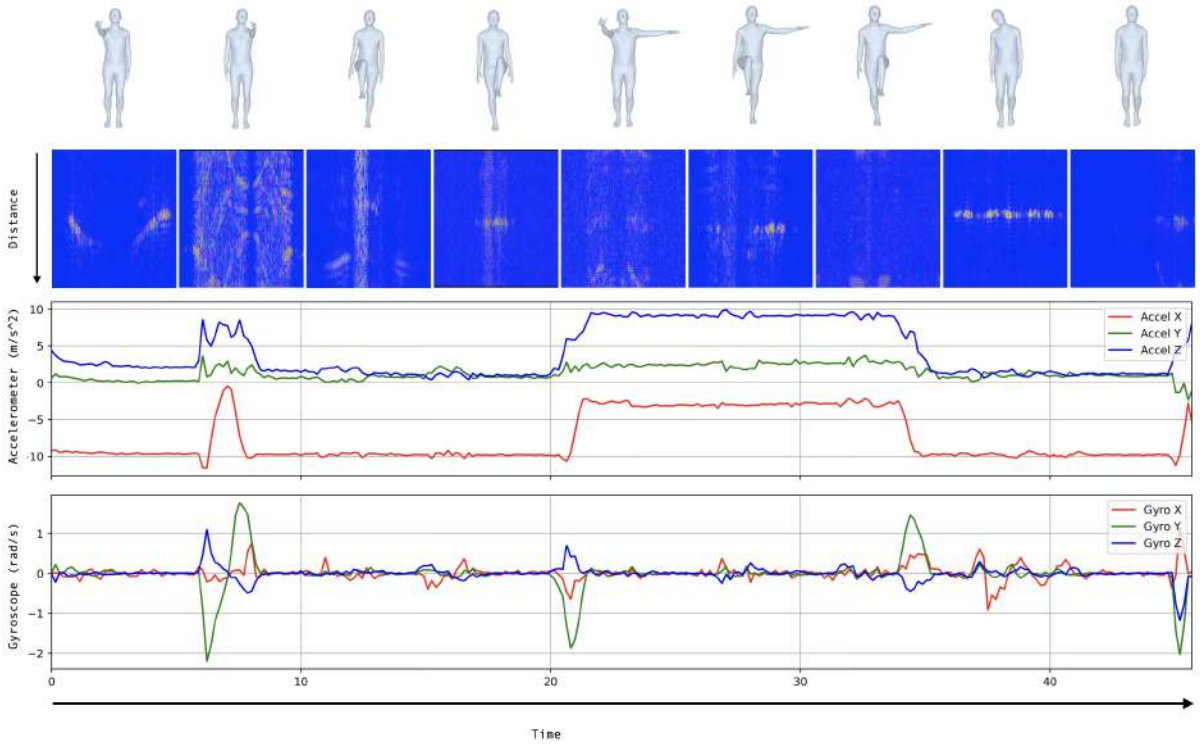


Fig. 2. Exploratory analysis of active acoustic sensing of EchoMotion for full-body pose estimation. Top: Representative full-body poses over time. Middle: Corresponding echo profiles captured from a smartwatch worn on the left wrist. Distinct motion patterns from occluded, uninstrumented body parts (e.g., right arm and both legs) are observable despite sensing from a single wrist-mounted device. Wrist motion appears as a shift in the echo profile, while movements from other body parts remain separable. Bottom: Synchronized IMU signals (accelerometer and gyroscope), which primarily reflect motion of the instrumented wrist and provide limited information about other body parts. These observations demonstrate that active acoustic sensing captures motion beyond the instrumented limb, including body parts outside the direct line of sight, and remains informative even when the sensing device itself is in motion.

3.3 Exploratory Study on Active Acoustic Sensing for Full-body Pose Estimation

We first develop a proof-of-concept hardware prototype (Section 4) to implement active acoustic sensing on a custom wristband resembling a consumer smartwatch. This exploratory study aims to assess the feasibility of full-body pose estimation using a single wrist-mounted sensing system.

Two key challenges arise in this setting. First, the system must capture motion from occluded body parts: regions such as the contralateral arm, the legs, and the head are not only uninstrumented but are also frequently blocked from the sensor's direct line of sight by the torso, clothing, or the instrumented arm itself. Second, it must remain robust when the sensing device itself is in motion, as the wrist frequently moves during everyday interactions.

Occlusion, however, does not completely remove a body part's contribution to the received signal. Because acoustic signals reach the microphone through both direct and indirect paths, an occluded body part still modulates some of these paths, and its motion appears in the differential echo profile, albeit with weaker intensity. Moreover,

human motion is highly coordinated: the movement of one body part is often correlated with that of others (e.g., the arms and legs swing in coordination during walking). A learning model can exploit such correlations to infer the pose of an occluded body part from the echoes of observable ones, which motivates the biomechanics-informed design of EchoMotion's neural network (Section 4.2). Together, these two factors (indirect acoustic paths and kinematic correlations) suggest that occluded body parts can still be tracked, although generally with lower accuracy than unoccluded ones.

To analyze these challenges, we conceptually divide the body into four quadrants: left upper, right upper, left lower, and right lower. The sensing device is worn on the left wrist, instrumenting only the left upper quadrant. The goal is to evaluate whether the system can capture motion in the remaining three quadrants, which are uninstrumented and often occluded.

As illustrated in Figure 2, the acoustic sensing system captures distinct motion patterns corresponding to movements of the right hand, right leg, and left leg, even though these body parts are neither instrumented nor consistently within the sensor's direct line of sight. In contrast, IMU signals exhibit strong responses primarily for movements of the instrumented wrist, providing limited information about distal body parts.

We further examine robustness to self-motion of the sensing device. While IMU signals are dominated by wrist motion, the acoustic echo profiles retain distinguishable patterns corresponding to movements of other body parts, even when the wrist itself is moving. Importantly, wrist motion manifests as a shift along the distance axis in the echo profile, while movements of other body parts produce separable patterns. This shift can be effectively handled by translation-invariant convolutional networks, which learn to recognize motion patterns regardless of their absolute position in the echo profile.

To validate whether these echo profiles are sufficient for full-body pose estimation, we train a lightweight convolutional model based on MobileNetV2 [85] to predict 12 OpenPose [14] joints (Section 4.2.1). This model achieves a mean per-joint position error (MPJPE) of 10.435 cm and a PCK@0.2 of 83.71%, where PCK measures the percentage of predicted joints within 20% of the torso diameter from the ground truth.

These results provide empirical evidence that active acoustic sensing can capture motion from occluded, uninstrumented body parts and remains effective despite self-motion of the wrist-mounted device, enabled in part by the translation-invariant properties of the learned model. Together, they establish the feasibility of using echo profiles for full-body pose estimation from a single smartwatch form factor.

4 ECHOMOTION ON A CUSTOMIZED WRISTBAND PROTOTYPE

We implemented the EchoMotion system on a custom wristband prototype equipped with two speaker-microphone pairs to evaluate the feasibility of full-body pose tracking via active acoustic sensing on a wrist-mounted device. In this section, we first describe the custom hardware design, and then outline the deep learning framework (Section 4.2) used to infer full-body pose from the captured acoustic data.

4.1 Hardware Design

4.1.1 Sensing and Controller Board. The core sensing unit of EchoMotion is the speaker-microphone pairs. As illustrated in Figure 3, we designed customized sensing boards equipped with an OWR-05049T-38D speaker and an ICS-43434 microphone. The sensing boards connect to a custom controller board that houses a low-power nRF52840 microcontroller and two MAX98357A audio amplifiers for transmitting ultrasonic C-FMCW [100] chirps. The controller board features an SD card interface, and we store raw audio data on a 32 GB SanDisk Extreme microSD card [5]. We optimized the firmware to minimize SD card accesses, enabling swift and power-efficient system operation. Flexible Printed Circuit (FPC) cables, copper shielded to reduce potential interference, connect the sensing boards to the controller unit.

4.1.2 3D-printed Watch Replica and Board Placement. As shown in Figure 3, we designed and 3D-printed an enclosure for the EchoMotion sensing system, modeled after the Apple Watch 6 [7] chassis, and used the original

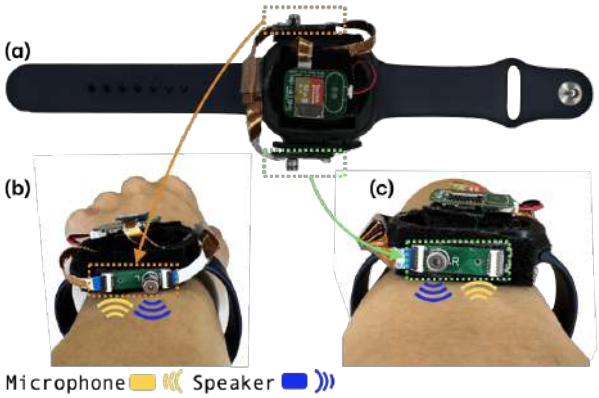


Fig. 3. Customized hardware prototype of EchoMotion: (a) Wristband form factor. (b-c) Side views of EchoMotion worn on the user's wrist, with each speaker and microphone's location highlighted.

watch strap to complete the form factor. The enclosure houses both the controller board and the battery. We mounted sensor boards incorporating a speaker-microphone pair on both the left and right sides of the watch replica, respectively.

4.1.3 Configuration of Active Acoustic Sensing Parameters. Given that the microphone on the EchoMotion sensing board samples at 50 kHz, we ensured that the frequency of transmitted chirps from both boards remained below the Nyquist frequency of 25 kHz. Specifically, the left speaker transmits C-FMCW chirps sweeping from 18 kHz to 21 kHz, while the right speaker transmits in the range of 21.5 kHz to 24.5 kHz. Each speaker transmits C-FMCW chirps with $N = 600$ samples per chirp. With a 50 kHz sampling rate, the sweep period for each chirp is $600/50000 = 0.012$ sec (12 milliseconds). Assuming a speed of sound of 343 m/s at room temperature (20°C), and following the formula provided in 2, the resolution of the EchoMotion sensing system is approximately 0.343 cm (or 3.43 mm), and the sensing range is about 205.8 cm (approximately 2 m).

4.1.4 Power Signature. The controller board is powered by a 290 mAh LiPo battery. Using a CurrentRanger [59], we measured that EchoMotion draws 100.4 mW (4.0 V, 25.1 mA) when the sensing components are active and data are being written to the SD card. With the sensing components physically disconnected, the draw falls to 88.8 mW (4.0 V, 22.2 mA), implying that the main sensing subsystem, two speaker–microphone pairs, consumes roughly 11.6 mW.

In contrast, BodyTrak [55], a wearable camera-based system for full-body pose tracking with offline processing, reports a camera subsystem draw of 460 mW (3.3 V, 140 mA). Likewise, as discussed in Section 2.2.1, IMU-based approaches (e.g. DIP [32]) typically employ arrays of Xsens sensors [71]. For example, the MTi-1 [72] and MTi-610 IMUs [73] draw about 44 mW and 340 mW per sensor, respectively. Since full-body tracking typically requires multiple IMUs, the total power scales by the number of units deployed. Together, these figures underscore that EchoMotion's active acoustic sensing approach can achieve lower power consumption compared to conventional imaging and inertial sensing techniques.

4.2 Deep Learning Framework

The acoustic sensing system of EchoMotion generates a differential echo profile, a range-spectrogram–like representation that encodes body movements across multiple distances. Our empirical evaluation confirms that distinct movements of different body parts produce characteristic patterns in these profiles. However,

reconstructing the full-body pose from these complex acoustic reflections requires a specialized inference mechanism. To address this challenge, we developed a lightweight, customized deep learning pipeline for EchoMotion. By processing sliding windows of the differential echo profile, our framework jointly infers SMPL [58] human mesh parameters and the 3D coordinates of full-body joints. In this section, we first describe our SMPL ground-truth acquisition procedure, then present the model's input construction, architecture, custom loss functions, and training hyperparameters.

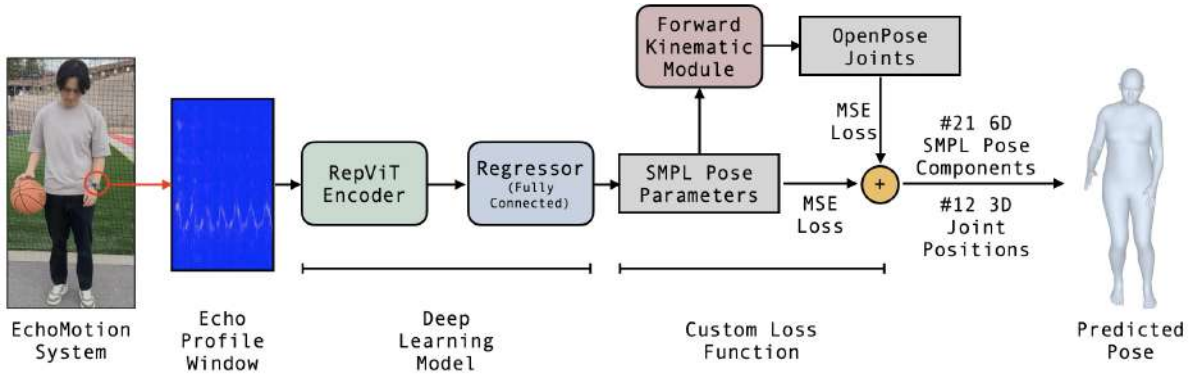


Fig. 4. Deep learning framework of EchoMotion: a RepViT [98] encoder processes 0.75-second context windows of differential echo profiles. The resulting feature representation is fed to a fully connected regressor to predict 6D rotations for 21 SMPL [58] pose parameters. A forward kinematics module then converts these SMPL parameters into 3D coordinates for 12 OpenPose [14] joints. The sum of the MSE losses on both the SMPL parameters and the joint coordinates against the ground truth is minimized to train the model in an end-to-end fashion.

4.2.1 Ground Truth Acquisition. The evaluation of EchoMotion involves user studies conducted across diverse environments (detailed in Sec 8), including outdoor spaces, indoor rooms with varying layouts, and building hallways. Due to the need to support data collection in these unconstrained settings, we adopt a vision-based pose proxy for supervision, following prior wearable pose tracking works that similarly rely on camera-based pose estimation in non-laboratory environments [8, 16, 24, 53, 111, 122]. While marker-based optical motion capture systems (e.g., OptiTrack [76], Vicon [95]) provide higher-fidelity pose measurements, their deployment is limited to controlled environments and does not scale to the breadth of scenarios considered in the evaluation of EchoMotion.

Specifically, we employ 4DHumans [27], a state-of-the-art vision-based human pose estimation framework. Each frame is first processed using a Mask R-CNN detector [28] implemented in Detectron2 [101] to localize the subject. The cropped region is then passed to HMR 2.0 [27], a Vision Transformer (ViT)-based model that estimates human pose parameters in the SMPL [58] representation.

To validate the reliability of this vision-based pose proxy, we conducted a calibration study in which 3 subjects performed all 18 full-body movements (outlined in Sec 5.1.2) from the EchoMotion action set, following the same protocol as our first user study (discussed in Sec 5) and yielding 60 minutes of data per participant, while being simultaneously recorded using a ZED 2i Stereo [90] RGB-D camera and an OptiTrack [76] marker-based motion capture system. We align the root joints of the estimated poses from both systems and compute standard pose accuracy metrics. The vision-based pose estimates achieve a mean joint position error of 0.41 ± 0.16 cm and a mean angular error of 0.83 ± 0.18 degrees with respect to the OptiTrack reference. These results indicate that,

while not a replacement for optical motion capture, the vision-based pose estimates provide a reasonable and quantifiable approximation of full-body motion suitable for large-scale and in-the-wild evaluation.

SMPL [58] (Skinned Multi-Person Linear Model) is a realistic 3D model of the human body that incorporates skinning and blend shapes, accounting for human anatomy and biomechanics. SMPL provides a body mesh with 6890 vertices and includes shape components β in a 10-dimensional vector. It also outputs a (3×3) rotation matrix for 24 joints, represented as a 216-dimensional vector. Among these joints, the rotation of the pelvis provides the global orientation (i.e., the direction the body is facing), while the remaining 23 joint rotations are defined relative to their parent joints in the kinematic tree. SMPL further offers a forward kinematic module that computes OpenPose [14] joint positions based on these rotations.

We extract two types of ground truth poses for training the EchoMotion deep learning pipeline:

i **SMPL Pose Parameter:** We utilize the SMPL pose components for 21 joints. The global orientation is excluded since we normalize the EchoMotion pose ground truth by setting the pelvis as the origin. Additionally, we omit the left and right foot joints because the left and right ankles, already included in the list, sufficiently represent foot poses, and toe tracking is outside the scope of EchoMotion. Prior works [36, 68, 111, 124] have demonstrated that using a 6D representation of rotation angles (obtained by discarding the last row of the (3×3) rotation matrix) produces smoother and more robust predictions. Therefore, we adopt a 6D representation for the rotations of the selected 21 joints, resulting in a ground truth SMPL pose parameter with a shape of $(21 \times 6) = 126$. Note that EchoMotion does not predict SMPL shape parameters.

ii **OpenPose Joint Coordinates:** We derive 3D coordinates for 12 OpenPose joints from the ground truth SMPL pose parameters. The joints include the nose (representing the head), neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right knee, right ankle, left knee, and left ankle. We select this joint subset to capture the primary kinematic structure of the upper and lower body while avoiding joints that are either redundant (e.g., hips, which closely follow the pelvis) or unreliable in unconstrained settings due to occlusion and loose clothing (e.g., toes and fine facial keypoints). We re-center the 3D coordinates to the pelvis of the user, and then scale them to physical units (centimeters) using the measured length of the user's non-dominant arm.

4.2.2 How Does Active Acoustic Feature Input Look Like? The deep learning model of EchoMotion processes sliding windows extracted from differential echo profiles. In these profiles, the x-axis represents time, while the y-axis corresponds to spatial information derived from the sensing system's range bins, a higher value in a given range bin indicates more significant movement at that distance. Each ground truth pose is aligned with the corresponding index in the synchronized echo profile, ensuring temporal consistency.

The length of the sliding window defines the amount of contextual information available for predicting the full-body pose at the current timestamp. Our hyperparameter tuning showed that shorter windows (less than 1.5 seconds) yield better predictions due to the instantaneous nature of human body movement. In particular, a window of 0.75 seconds, represented by 60 consecutive samples along the temporal axis, delivered optimal performance and was chosen for model training.

The EchoMotion prototype employs a customized hardware implementation on a watch replica, which includes speaker-microphone pairs on both the left and right sides. As a result, there are four possible acoustic paths for the transmitted chirps (left speaker to left microphone, left speaker to right microphone, right speaker to left microphone, and right speaker to right microphone), yielding an echo profile with four channels. Furthermore, our tuning experiments determined that a sensing range of approximately 1.5 meters, represented by 200 pixels along the spatial axis, produces the most stable full-body pose predictions.

In summary, the input to the EchoMotion deep learning model is a three-dimensional tensor with dimensions corresponding to the number of channels, context window length, and sensing range, that is, a tensor of shape $(4 \times 60 \times 200)$.

4.2.3 Backbone Selection for Feature Extraction. In the case of EchoMotion's deep learning framework, each input sliding window can be viewed as a multi-channel "image," where each channel encodes spatial and temporal aspects of the tracked movement. To extract rich spatiotemporal features, we can therefore adapt standard 2D convolutional neural networks (CNNs) or vision transformers (ViTs) [23] to accept four-channel inputs. ViTs excel at treating image-like inputs as sequences of overlapping patches, yielding powerful representations of complex spatiotemporal patterns. However, their high computational cost can limit real-time performance on resource-constrained devices.

Conversely, CNNs, especially those optimized for mobile inference [85], offer significantly lower latency. Our design goal for EchoMotion is to efficiently extract movement features from echo profiles with minimal computational overhead, potentially enabling real-time full-body pose inference. To meet this requirement, we adopt RepViT M1 [98], a CNN architecture that incorporates insights from lightweight ViTs while retaining the speed advantages of MobileNet-style networks [85]. RepViT achieves this by using reparameterizable convolutions arranged in a ViT-like topology, delivering both the modeling capacity of transformers and sub-millisecond inference latency.

Because RepViT M1 [98] satisfies our criteria for low latency, compact footprint, and robust spatio-temporal feature extraction, and outperforms other lightweight neural architectures (e.g. FastViT [94]) on our active acoustic pose-tracking benchmarks, we select it as the encoder backbone for EchoMotion's deep learning framework.

4.2.4 Model Architecture. The input sliding windows of the differential echo profile are first normalized. This involves clipping the echo profile values using hyperparameter-tuned minimum and maximum thresholds, followed by a linear mapping to the $[0, 1]$ range. As illustrated in Figure 4, the normalized input is then passed through the RepViT encoder [98], which extracts a compact and informative feature vector.

This feature vector is subsequently fed into a regressor network, implemented as a fully connected layer, to predict the SMPL pose parameters (as described in Section 4.2.1). The predicted SMPL parameters are then used by a forward kinematics module to infer the 3D joint coordinates corresponding to the OpenPose [14] skeleton.

To train the system, we supervise it using two complementary losses. First, we compute the Mean Squared Error (MSE) between the predicted SMPL parameters $\hat{y}_s$ and the ground-truth parameters y_s, both represented as 6D rotation matrices: $l_s = \|y_s - \hat{y}_s\|^2$.

Next, using the SMPL's [58] forward kinematics module, we transform the predicted pose into 3D joint positions. These are further scaled based on a physical body constraint, the distance between the wrist and shoulder, to obtain the predicted OpenPose joint coordinates $\hat{y}_j$. We then compute a second MSE loss between the predicted and ground-truth joint positions: $l_j = \left\|y_j - \hat{y}_j\right\|^2$.

The final objective is a custom loss function that combines both terms: $l_m = l_s + l_j$, which we use to optimize the entire network end-to-end.

4.2.5 Training Details. We implemented the deep learning framework using PyTorch Lightning. The model was trained on NVIDIA Tesla V100 GPUs for 25 epochs with a batch size of 128. We used the Adam optimizer [45] and applied a cosine annealing learning rate scheduler, starting with an initial learning rate of 5×10^{-3}.

5 USER STUDY 1: EVALUATING ECHOMOTION IMPLEMENTATION ON CUSTOMIZED WRISTBAND

To evaluate the full-body pose tracking performance of the customized EchoMotion hardware, we conducted a user study with ten participants. In this section, we first outline the study details and then present our analysis of the evaluation results.

5.1 Study Protocol

5.1.1 Participants. The user study protocol was approved by the Institutional Review Board (IRB) at our organization. After receiving approval, we recruited 10 participants aged between 21 and 27 years, with an average age of 24.56 ± 2.01 years. Among the participants, 5 identified as male, 4 as female, and 1 preferred not to disclose their gender. The average height and weight of the participants were 170.47 ± 8.48 cm and 69.16 ± 13.65 kg, respectively.

Participants were asked about their non-dominant hand (9 participants responded that their left hand is the non-dominant one) and where they would prefer wearing a smartwatch; all chose their left wrist. We measured the length of each participant's left arm (from the left shoulder to the left wrist), which was used to normalize and scale the ground truth OpenPose [14] joint coordinates, with the pelvis as the origin of the coordinate system. The average left arm length was 68.38 ± 5.72 cm.

The study took place in a large experimental room with an active HVAC system. The total duration of the study was 75 minutes. Upon completion of the study, each participant received a $20 gift card as compensation.

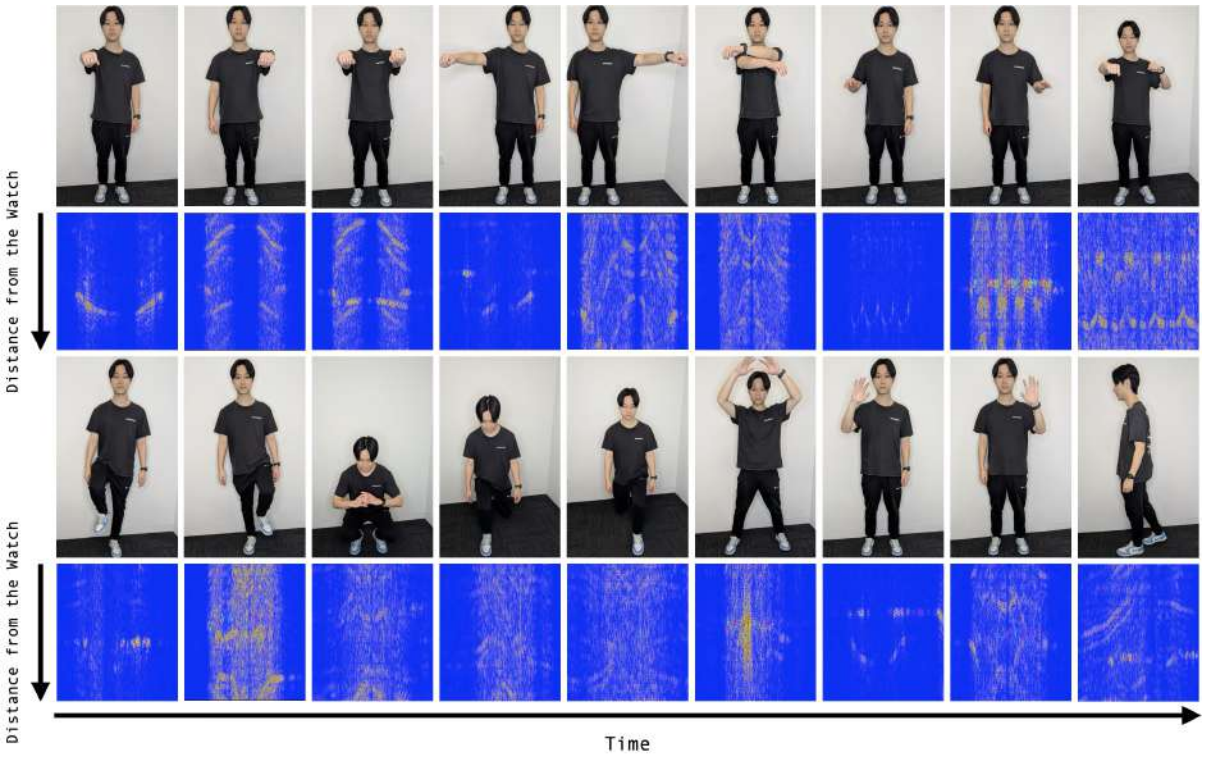


Fig. 5. The 18 full-body movements utilized in the EchoMotion user study. We also show the corresponding differential echo profiles captured using the EchoMotion sensing system mounted on the user's left wrist.

5.1.2 Data Collection Procedure. We recruited 10 individuals to evaluate the performance of the EchoMotion pose tracking system implemented on a customized wristband prototype. At the outset, participants were informed of the experiment's purpose and provided their consent. Then, the participants wore the wristband on their

left wrist, adjusted it to their comfort level, and used Apple AirPods Pro to receive audio instructions. Each participant was instructed to stand in a designated area of the room, calibrated with the camera angle, while an Apple MacBook Pro 2022 was positioned on a desk approximately 4 meters away. Visual instructions in both text and video formats were displayed on the laptop screen.

Ground truth data were recorded as 30 fps video using ZED 2i Stereo [90] RGB-D camera while participants executed various body movements wearing the EchoMotion prototype. Note that the participants performed movements at their own natural pace. The acoustic sensing data were saved to an SD card; Audio data from the smartwatch and video ground truth were synchronized using a clapping action by the user, which created a distinct audio pattern that could be matched to the corresponding video frame.

For each participant, we collected data over 10 sessions, each lasting 6 minutes. Within each session, participants performed 5 repetitions of 18 different body movements, presented in a random order. Between sessions, they removed and re-worn the smartwatch to simulate a typical remounting scenario. As EchoMotion explores the feasibility of tracking full-body pose with a single smartwatch, we focused on a representative subset of possible full-body movements instead of covering every possible movement. Based on prior work [22, 36, 53, 68, 111, 126] and with the goal of introducing variance in movement patterns and body sides, we selected 18 full-body movements (illustrated in Figure 5) categorized into three groups:

- **Upper-body Movements [8]:** right arm raise, left arm raise, both arm raise, right arm swing, left arm swing, both arm swing, dribbling basketball with right arm, dribbling basketball with left arm.
- **Lower-body Movements [5]:** right leg raise, left leg raise, squat, lunges with right leg, lunges with left leg.
- **Freestyle and Locomotion [5]:** jumping jack, waving with right arm, waving with left arm, boxing with alternate arms, walking.

Overall, we collected 60 minutes of data from each participant, amounting to 108,000 data instances per participant (60 minutes $\times$ 60 seconds/minute $\times$ 30 fps). In total, this resulted in 10 hours of data from the EchoMotion customized hardware prototype.

5.2 Evaluation Metrics

We adopt two widely used evaluation metrics for human pose reconstruction, based on prior work [9, 36, 68], to quantitatively evaluate the performance of the EchoMotion full-body pose estimation system:

i **Mean Per Joint Position Error (MPJPE):** MPJPE is computed as the average Euclidean distance between the predicted and ground truth joint positions, with the pelvis aligned as the origin of the coordinate system. MPJPE is reported in centimeters (cm).

ii **Mean Per Joint Rotation Error (MPJRE):** MPJRE is computed as the average angular discrepancy between the predicted and ground truth joint rotations. MPJRE is reported in degrees (°).

5.3 Train-Test Scheme

We evaluate the performance of the EchoMotion system using three different train-test strategies:

i **Across-Session Evaluation:** Each participant contributes data from 10 distinct recording sessions, where devices were remounted between sessions. In this evaluation strategy, data from one session is used as the test set, while the remaining nine sessions are used for training and validation. This process is repeated for all sessions, and we report the average performance across sessions for each participant.

ii **Leave-One-Participant-Out Evaluation:** Here, the data from one participant is held out as the test set, while data from the remaining nine participants are used to train a user-independent model. This setup forms a 10-fold cross-validation scheme, where each participant's data serves as one fold. We report performance

metrics, specifically, MPJPE (Mean Per Joint Position Error) and MPJRE (Mean Per Joint Rotation Error), for each participant.

iii **Leave-One-Action-Out Evaluation:** The EchoMotion user study includes 18 distinct movement types. In this scheme, all instances of a specific action (e.g., jumping jacks) are used as the test set, while the model is trained on data from the remaining 17 actions. This process is repeated for each action, enabling action-wise cross-validation.

For each train-test scheme, we additionally report the prediction error for each full-body joint estimated by EchoMotion, providing a detailed breakdown of model performance across different joints and motions.

5.4 Results

Table 2. Across-session evaluation of the EchoMotion custom wristband-form-factor hardware prototype across 10 participants in the user study.

	MPJPE (cm)	**MPJRE (°)**
Upper-body Movements	4.01 ± 1.39	11.04 ± 1.94
Lower-body Movements	4.38 ± 1.05	10.17 ± 1.76
Freestyle and Locomotion	4.57 ± 1.66	10.91 ± 2.69

5.4.1 Across-Session Evaluation. Given the natural variability in full-body movement, even for the same user, we simulate the performance of EchoMotion in a new session without any form of calibration data using this train-test scheme. Notably, participants remounted the device between each session of the user study. In this evaluation protocol, we train the model on all but one session of a user's data, holding out the remaining session for testing. This procedure is repeated across all sessions, and the aggregated results for all 10 participants are reported in Table 2.

We observed a mean per joint position error (MPJPE) of 4.323 cm (SD = 1.367) and a mean per joint rotation error (MPJRE) of 10.707 degrees (SD = 2.128) across sessions. Additionally, we provide a breakdown of MPJPE and MPJRE by movement category in Table 2 and find that the numerical errors remain consistent across different types of movements within the same user, even when the device was remounted between sessions.

5.4.2 Leave-One-Participant-Out Evaluation. An ideal full-body pose reconstruction system should work seamlessly out of the box for unseen users, enabling intuitive and immediate interaction. To assess EchoMotion's performance under this scenario, we adopt a leave-one-participant-out evaluation strategy. Given that movement patterns vary significantly across individuals, this setup represents the most challenging evaluation condition. In this scheme, we train a user-independent model on data from all participants except one, using the held-out participant's data exclusively for testing. This process is repeated for each of the 10 participants in the user study.

We report the Mean Per Joint Position Error (MPJPE) and Mean Per Joint Rotation Error (MPJRE) for each participant in Figure 6(a), respectively. Compared to the across-session evaluation, both positional and rotational errors increase under this user-independent setting. The average MPJPE across the 10 users is 8.644 cm (SD = 1.917), and the average MPJRE is 15.814 degrees (SD = 2.651). Although this user-independent performance is comparable to prior work [22, 68, 111], the decline in accuracy can be attributed to the inherent variability in motion patterns across unseen users. This challenge may be addressed by curating a large-scale, diverse dataset of active acoustic pose data.

While a calibration-free solution is ideal, it is common for wearable pose tracking systems, including commercial solutions with multiple IMUs such as Xsens [71] and Rokoko [83], to include a calibration phase for new users. To

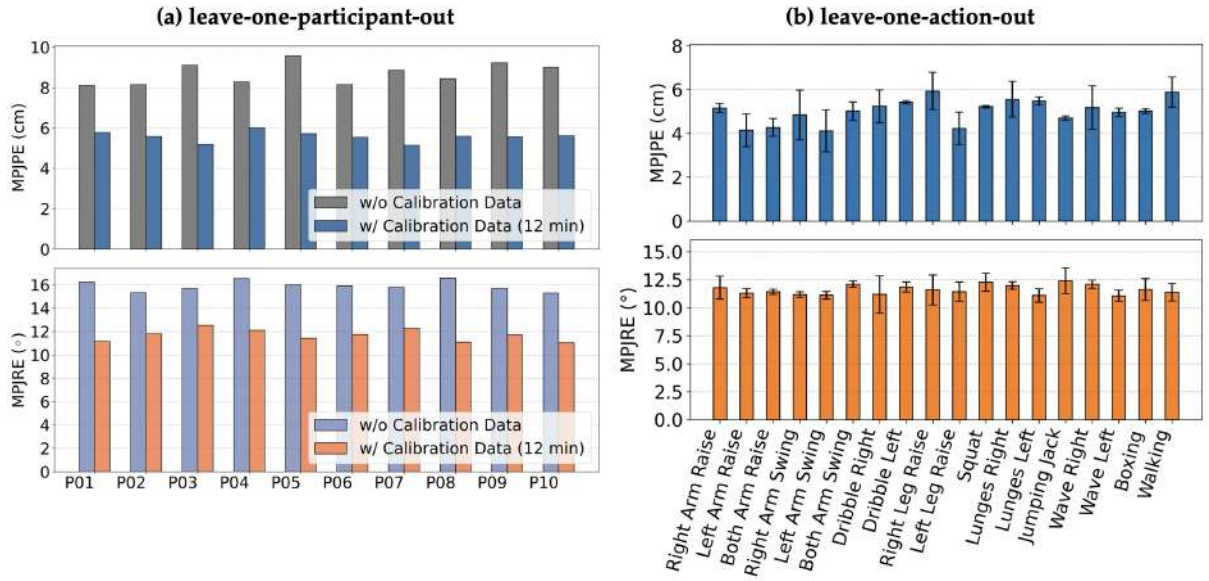


Fig. 6. **(a)** Leave-One-Participant-Out evaluation of EchoMotion's customized hardware prototype across 10 participants in the first user study, showing MPJPE (cm) and MPJRE (°) in scenarios with no calibration data and with fine-tuning using 12 minutes of calibration data. **(b)** Leave-One-Action-Out evaluation of the EchoMotion implementation on a custom hardware prototype, aggregated across all 10 participants in the first user study.

evaluate this practical scenario, we fine-tune the user-independent model using a small amount of calibration data collected from the same test user. As shown in Figure 6(a), this calibration significantly reduces both positional and rotational errors.

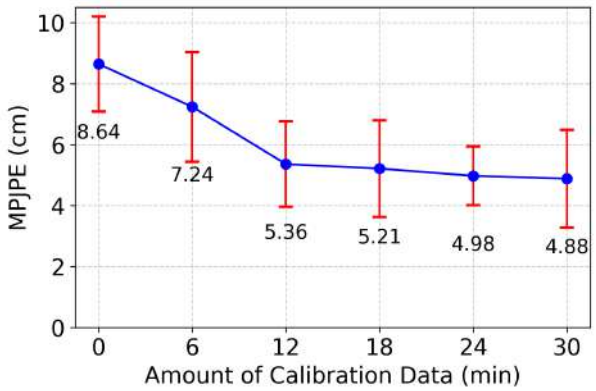


Fig. 7. Impact of using calibration data in training a user-adaptive model. First, a user-independent model is trained without any calibration data (denoted as 0 on the x-axis). Then, we fine-tune the model with increasing amounts of calibration data from the same user in the test set.

Figure 7 further illustrates the impact of the amount of calibration data on performance. With just 12 minutes of calibration data, the MPJPE improves from 8.64 cm to 5.36 cm in the user-adaptive model. The results also show that while additional calibration data continues to enhance performance, the most substantial gains occur within the first 12 minutes of fine-tuning.

5.4.3 Leave-One-Action-Out Evaluation. We conducted this evaluation to simulate a scenario where the EchoMotion full-body pose reconstruction system encounters previously unseen actions. In this protocol, the model is trained on data from all but one movement in the dataset, aggregated across all participants in the user study. The excluded movement is used as the test set. We report the leave-one-action-out performance of the EchoMotion customized hardware prototype in Figure 6(b).

The results show a mean per joint position error (MPJPE) of 5.018 cm (SD = 1.425) and a mean per joint rotation error (MPJRE) of 11.599 degrees (SD = 2.391) across all actions. Notably, the performance degradation for unseen actions is less severe than that observed in the unseen user evaluation. This can be attributed to the fact that, while the specific test action may not appear in the training set—for instance, a right arm raise—similar motions such as a left arm raise are included. As a result, the variability introduced by unseen actions is lower than that introduced by unseen users.

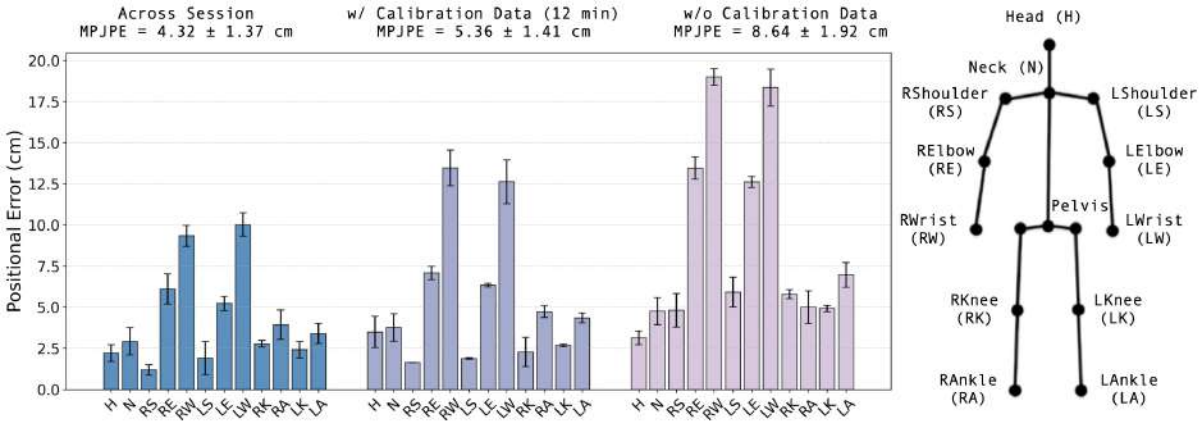


Fig. 8. Positional error (in cm) of each joint tracked by the EchoMotion implementation on a customized hardware prototype across all participants. We report the joint-wise breakdown for different train-test schemes (cross-session and user-independent with and without calibration data) that we adopt to evaluate the EchoMotion system's performance.

5.4.4 Accuracy Across Different Body Joints. We also analyze the breakdown of positional error across different body joints tracked by EchoMotion. As expected, pose estimation error tends to accumulate along the kinematic chain, resulting in higher errors at the end-effectors or leaf nodes of the kinematic tree—namely, the left wrist, right wrist, left ankle, and right ankle (illustrated in Figure 8). This can be partially attributed to the fact that the wrists were involved in the most diverse set of movements in our user study. The combination of high movement variability and error accumulation along the kinematic chain leads to the highest errors being observed at the wrists.

We further observe that these errors decrease when fine-tuning the user-independent model with 12 minutes of user-specific calibration data. For instance, in the user-adaptive model, the left wrist error is 12.621 cm (SD = 1.337), while the right wrist error is 13.466 cm (SD = 1.086). Although the right wrist exhibits slightly higher error, the difference between left and right side joints across all three evaluation scenarios, across-session, user-independent,

and user-adaptive, is negligible. This is notable considering that only the left wrist was instrumented with the EchoMotion smartwatch prototype, further highlighting the robustness of the model.

Additionally, we notice a degradation in performance when tracking shoulder joints under the user-independent evaluation scheme. This may be due to variations in users' body shapes, which affect how movement information is captured in specific range bins of the echo profile. The shape variability of unseen users introduces additional challenges in accurately estimating distances from the wrist-mounted device to the torso.

In summary, this evaluation demonstrates that user-adaptive training with a small amount of calibration data strikes an effective balance between practical usability and model performance. While EchoMotion maintains robust accuracy across most joints, it remains susceptible to larger errors at the end-effectors, particularly the wrists, due to movement variability and kinematic propagation effects.

6 ECHOMOTION ON A COMMERCIAL SMARTWATCH

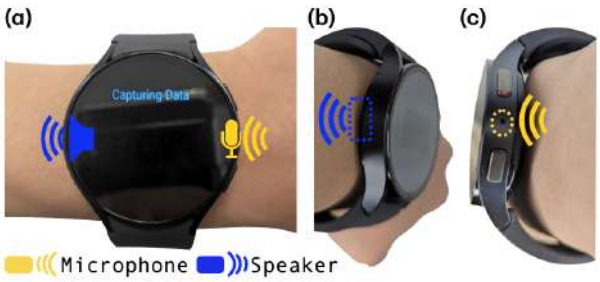


Fig. 9. (a) COTS Samsung Galaxy Watch 6 capturing user data using the WearOS implementation of the EchoMotion sensing system. (b) Left side view of the watch, with speaker location highlighted. (c) Right side view of the watch, with microphone location highlighted.

The first prototype with customized hardware has shown promising performance in the first study on tracking full-body poses. However, although it only requires one form factor (wristband) to track full body poses, it still requires at least two pairs of microphones and speakers to be functional. This is challenging to apply to existing smartwatches, as most watches only have one pair of microphone and speaker. If EchoMotion can perform on off-the-shelf (COTS) smartwatches, it will further lower the barrier of tracking full body poses using wrist-mounted wearables. To explore this feasibility, we apply EchoMotion on a commercial smartwatch and evaluate its performance with 10 participants.

6.1 Details of Software-only Implementation

We implemented the EchoMotion sensing system on an unmodified Samsung Galaxy Watch 6 [84] running WearOS. The device features a 425 mAh Li-ion battery, a left-side speaker, and a right-side microphone, and it supports audio playback and recording at sampling rates of up to 48 kHz. Given that the Nyquist frequency is 24 kHz and the device has only one speaker on the left side, we opted to use C-FMCW [100] chirps with $N = 600$ samples per chirp, sweeping from 18 kHz to 21 kHz, identical to the configuration used on the customized hardware prototype (see Section 4.1.3). This configuration provides an approximate sensing resolution of 0.357 cm ($\approx$ 3.57 mm) and a sensing range of 214.375 cm ($\approx$ 2 m). In this setup, the system simultaneously emits inaudible signals through the left-side speaker and captures their reflections or diffractions from the body using the right-side microphone. The captured data is then transmitted to a host PC over Wi-Fi using the Transmission Control Protocol (TCP).

6.2 Power Signature

We measured the power usage of the EchoMotion WearOS app using Android Debug Bridge (adb) [20]'s *batterystats* profiling functionality on the Samsung Galaxy Watch 6 [84]. For a 30-minute (1800-second) run of the sensing system app, the profiler estimated a power drain of 21.7 mAh. This value is derived from Android's internal models, which approximate energy usage based on CPU, network, screen, and sensor activity. During this period, the mean battery level drop for a 425 mAh battery was between 5% and 6%, with the variation due to coarse-grained battery percentage measurements (rounded to approximately 1% increments). Note that all background apps and processes were running, and the system transmitted data to the host computer over TCP while the Android profiler recorded the power signature.

7 USER STUDY 2: EVALUATING ECHOMOTION ON A COMMERCIAL SMARTWATCH

We evaluated the pose-tracking performance of the EchoMotion implementation on a commercial off-the-shelf Samsung Galaxy Watch 6 in a user study with 15 participants. In this section, we first outline the study design, then analyze the results and compare them to those obtained with the customized hardware implementation from our first study.

7.1 Study Protocol

We recruited a new cohort of 15 participants (no overlap with the first study) to evaluate EchoMotion on a commercial off-the-shelf smartwatch. Of these participants, 7 identified as male and 8 as female, with an average age of 24.44 ± 1.45 years (range: 21–29). Their average height was 164.29 ± 3.68 cm, and their average weight was 62.91 ± 3.01 kg. Participants wore the Samsung Galaxy Watch 6 [84] on their left arm, which averaged 67.29 ± 1.14 cm in length.

The study was conducted in the same room and over the same 75-minute duration as the first study (Section 5.1), following an identical data-collection protocol. The only procedural difference was data transfer: with the COTS smartwatch, raw acoustic data streamed over TCP to a laptop, whereas in the first study, it was stored on an SD card. Each participant performed the same set of 18 actions in random order over 10 sessions (6 minutes per session). In total, we collected 15 hours of data, 162,000 samples, from the COTS Samsung smartwatch. Upon completion, each participant received a $20 gift card as compensation.

7.2 Results

We evaluated the COTS-smartwatch implementation of EchoMotion using the same protocol as the first study to ensure a fair comparison. After collecting data from 15 new participants, we trained the EchoMotion deep learning framework with the identical data-splitting strategy. The only change to the model architecture reflects the hardware's acoustic configuration. Our custom prototype used two pairs of speaker–microphones (one on each side), producing an echo profile with four channels. In contrast, the Samsung Galaxy Watch 6 has a speaker and a microphone placed on two sides of the watch as shown in Figure 9, separated by the watch body. If worn on the left wrist, the speaker is towards the body, and the microphone is facing the hand. This different configuration likely introduces different reflection patterns and also more occlusion by the watch body. Since this watch only has one pair of microphone and speaker, the input sliding windows for the second study's echo profiles have a single channel (with the same height and width dimensions). All other training parameters remained unchanged.

7.2.1 Across-Session Evaluation. We evaluated EchoMotion's pose tracking performance on a COTS smartwatch in a new worn-session scenario. Following the same protocol as the first study, participants remounted the smartwatch between each session. For each user, we trained the model on data from nine sessions and held out the tenth for testing. And, we repeat this process for each session to obtain cross-validation performance.

Table 3. Across-session evaluation of the EchoMotion implementation on commercial Samsung Galaxy Watch 6 across 15 participants in the user study.

	MPJPE (cm)	MPJRE (°)
Upper-body Movements	5.71 ± 1.20	10.97 ± 2.83
Lower-body Movements	5.04 ± 1.89	10.66 ± 2.16
Freestyle and Locomotion	5.26 ± 1.16	10.83 ± 2.33

The results of this across-session evaluation are summarized in Table 3. Across all 15 participants, the mean per joint position error (MPJPE) is 5.337 cm (SD = 1.743), and the mean per joint rotation error (MPJRE) is 10.818 degrees (SD = 2.418). A breakdown by movement type shows that lower-body movements incur slightly lower errors compared to other movement categories.

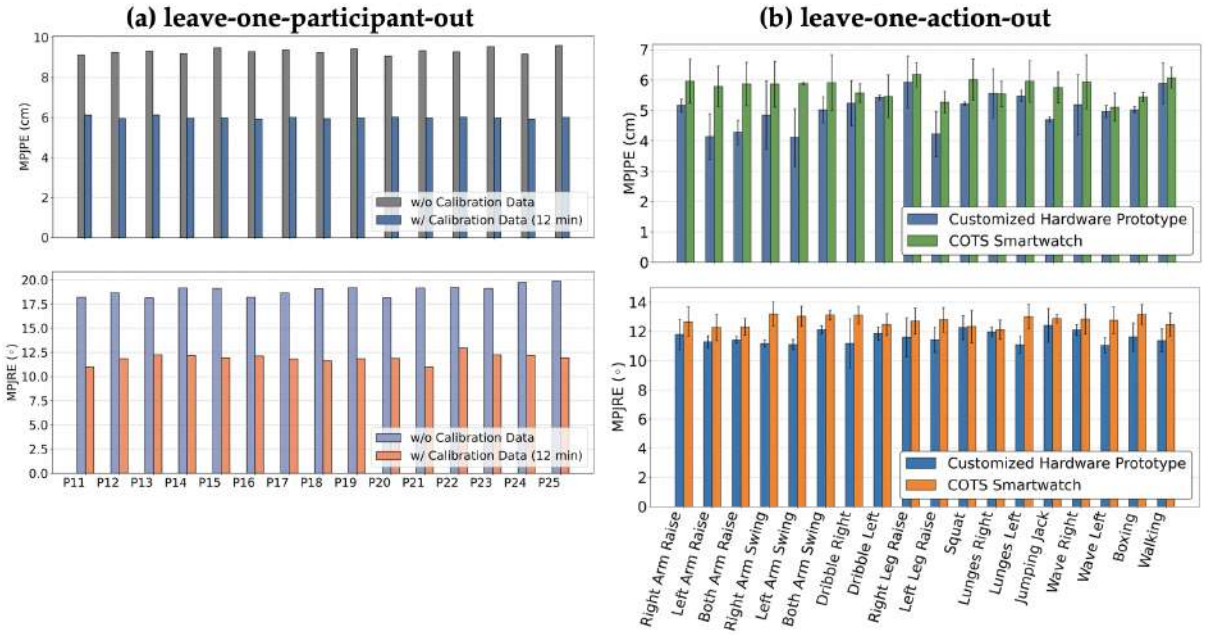


Fig. 10. **(a)** Leave-One-Participant-Out evaluation of EchoMotion implementation on COTS smartwatch across 15 participants in the second user study, showing MPJPE (cm) and MPJRE (°) in scenarios with no calibration data and with fine-tuning using 12 minutes of calibration data. **(b)** Leave-One-Action-Out evaluation of EchoMotion system across 18 activities of the user study, with comparison of performance between the implementation on a customized hardware prototype and on a COTS smartwatch.

7.2.2 Leave-One-Participant-Out Evaluation. We evaluated the COTS smartwatch implementation of EchoMotion in an unseen user scenario using a leave-one-participant-out scheme. For each fold, we trained the model on data from 14 participants and tested it on the held-out participant, repeating this process for all 15 users. As with our custom hardware prototype, this represents the most challenging evaluation due to high inter-user variability. Consequently, performance degrades relative to the across-session results (Figure 10(a)), with an average MPJPE

of 9.302 cm (SD = 1.491) and an average MPJRE of 18.908 degrees (SD = 3.574). These results indicate that EchoMotion can provide reasonable full-body pose estimates for unseen users without any calibration.

However, variability in motion patterns, body geometry, and signal characteristics across users introduces a performance gap. To quantify the benefit of adaptation, we fine-tuned each user-independent model with 12 minutes of calibration data from the test participant. This lightweight calibration yields substantial gains: across all 15 users, MPJPE improves from 9.302 to 5.987 cm (SD = 2.011) and MPJRE from 18.908 to 11.920 degrees (SD = 2.548), as shown in Figures 10(a).

These results suggest that EchoMotion supports two practical operating modes: (1) a calibration-free setting that enables immediate use with moderate accuracy, and (2) a user-adaptive setting that achieves higher accuracy with a brief, one-time calibration. Such calibration procedures are common in wearable, non-vision pose-tracking systems [9, 71, 111]. The improved user-independent performance with 15 participants suggests that increasing dataset diversity and scale can further enhance generalization without calibration.

7.2.3 Leave-One-Action-Out Evaluation. In this evaluation, we train the model on data from all actions and users except one held-out action, repeating the process for each of the 18 actions. Across all actions, we observe an average MPJPE of 5.738 cm (SD = 1.512) and an average MPJRE of 12.239 degrees (SD = 2.747), as shown in Figure 10(b).

Compared to the customized hardware prototype, the leave-one-action-out performance on the COTS smartwatch degrades slightly (Figure 10(b)). This is likely because the custom prototype uses two speaker and microphone pairs, one on each side of the watch, providing broader acoustic coverage, whereas the COTS watch has only one speaker (left) and one microphone (right). Nonetheless, the COTS implementation remains competitive, with only a ≈ 0.75 cm increase in positional error and a ≈ 1.135 degrees increase in rotational error, all without any hardware modifications to the smartwatch.

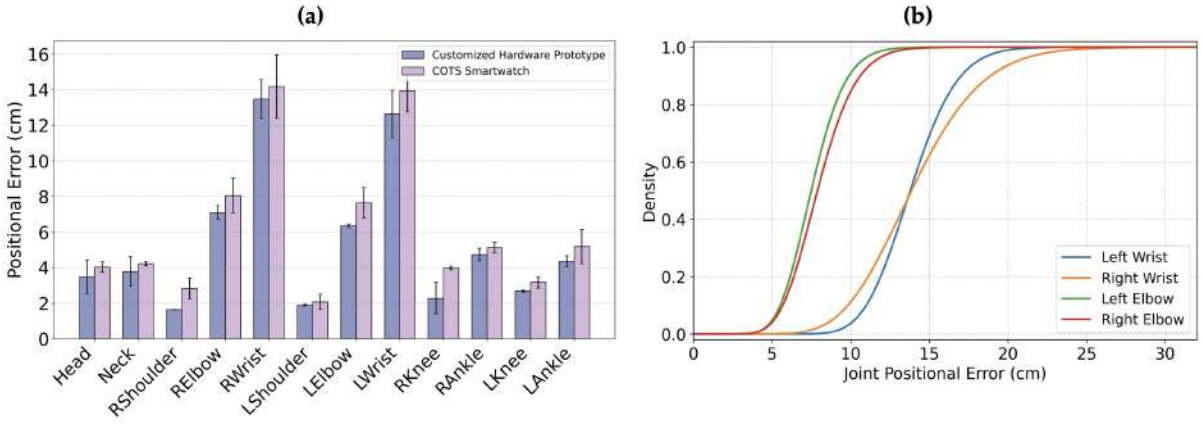


Fig. 11. **(a)** Positional errors (in cm) for each joint tracked by EchoMotion, comparing the customized hardware prototype and the COTS smartwatch implementations. **(b)** Cumulative distribution functions (CDF) of elbow and wrist joint positional errors for the EchoMotion COTS smartwatch implementation, aggregated across 15 participants of the second user study.

7.2.4 Joint-wise Analysis of Result. To examine EchoMotion's full-body tracking in detail, we analyzed positional errors for each joint. Figure 11(a) reveals two key observations: (1) errors increase when using the COTS smartwatch versus the custom prototype, and (2) end-effectors (the wrists) exhibit the highest errors due to kinematic accumulation.

We further plotted the cumulative distribution functions (CDFs) for wrist and elbow positional errors in Figure 11(b). Interestingly, tracking on the left side (both wrist and elbow) outperforms the right side. This is likely because participants wore EchoMotion on their non-dominant (left) wrist, giving the system better acoustic coverage on that side. Crucially, however, right-side joints are still tracked with nearly comparable accuracy, demonstrating that EchoMotion can achieve full-body pose estimation using a single smartwatch on the non-dominant hand, without requiring any other instrumentation or any other sensing system.

7.3 End-to-end Inference on COTS Smartwatches

Table 4. Cross-device performance of EchoMotion. The model is trained on Samsung Galaxy Watch 6 (User Study 2, 15 participants) and evaluated on test data from the same study and a follow-up study (5 participants) on other devices, without retraining.

Test Device	**MPJPE (cm)** ↓	**MPJRE (°)** ↓
Samsung Galaxy Watch 6	9.302	18.908
Samsung Galaxy Watch 7	10.149	18.776
Pixel Watch 3	10.122	19.184

While our data collection pipeline on the Samsung Galaxy Watch 6 streams data to a host PC, EchoMotion is designed to operate fully on-device. We implement the complete pipeline, including sensing, signal processing, and model inference, on commodity smartwatches. The trained model is converted to a traced PyTorch Mobile version (28.1 MB), and runs directly on WearOS. During operation, the device continuously emits and records ultrasonic signals using its built-in speaker and microphone, and a signal processing thread converts the raw audio into echo profiles for real-time pose inference.

We evaluate this system on three commercial off-the-shelf devices: Samsung Galaxy Watch 6, Samsung Galaxy Watch 7, and Google Pixel Watch 3. The model is trained using data from the Samsung Galaxy Watch 6 collected in User Study 2 (15 participants). To assess cross-device generalization, we deploy the same model on the other devices without retraining or calibration, and evaluate it on data collected from a follow-up study with 5 additional participants using the same protocol.

As shown in Table 4, EchoMotion achieves an MPJPE of 9.302 cm and an MPJRE of 18.908° on the Galaxy Watch 6 test set. When evaluated on unseen devices, performance degrades only slightly, with MPJPE increasing to 10.149 cm on the Galaxy Watch 7 and 10.122 cm on the Pixel Watch 3, while MPJRE increases marginally to 18.776° and 18.184°, respectively. Despite both device and user distribution shifts, the performance remains stable, indicating strong cross-device generalization. This robustness stems from the use of cross-correlation-based echo features, which encode relative propagation delays rather than absolute signal magnitudes, making them less sensitive to device-specific variations.

7.4 Real-Time Latency Measurement

We evaluate the runtime performance of EchoMotion to assess its suitability for real-time operation. We first measure the data transmission latency for streaming acoustic data from a Samsung Galaxy Watch 6 to an Apple MacBook Pro M1 (2021) over WiFi using TCP. For a 0.75-second acoustic window (the context required for a single inference), the average transmission latency is 2.6 ms.

We then measure model inference latency on the MacBook. Processing a single window requires 13.1 ms (SD = 2.7) on CPU, and 0.68 ms (SD = 0.09) using GPU and Neural Engine acceleration via CoreML [19].

Finally, we measure fully on-device performance by running both signal processing and model inference on the Samsung Galaxy Watch 6. The average end-to-end latency per prediction for a 0.75-second window is 119.6 ms (SD = 14.1), including 38.1 ms (SD = 3.3) for signal processing and 81.7 ms (SD = 6.8) for model inference. This supports responsive real-time interaction without external infrastructure. All reported inference times exclude SMPL body mesh rendering.

8 USER STUDY 3: EVALUATING ECHOMOTION IN A SEMI-IN-THE-WILD SETTING

The first two user studies assessed EchoMotion on both a custom hardware prototype and a COTS smartwatch, with all experiments conducted in controlled laboratory environments. While these studies established feasibility under ideal conditions, they did not capture the challenges of deployment in everyday contexts. To address this gap, we conducted a third study in a semi-in-the-wild setting, deploying EchoMotion on a COTS smartwatch in naturalistic environments.

Because EchoMotion relies on acoustic reflections from both the user's body and nearby objects within its sensing range, we examined how reflections from environmental structures influence pose tracking performance. We also evaluated robustness to ambient acoustic noise by conducting sessions outside the lab, specifically in a building hallway and at a roadside location. This section describes the study design and presents pose tracking performance under these conditions.

8.1 Study Protocol

We recruited a new group of 10 participants (ages 20–31 years, $M = 24.51$, $SD = 2.33$). Seven were identified as male and three as female. Participants' average height and weight were 168.54 ± 2.26 cm and 64.95 ± 4.06 kg, respectively. None of these participants had taken part in the previous studies.

The study was conducted at three distinct locations:

- **Indoor room (3m × 4m):** furnished with a table, chair, shelf, yoga mat, and two dumbbells. Before each session, the researcher randomly rearranged the furniture to vary reflection patterns. The furniture remained fixed during participant trials. This setting allowed us to test sensitivity to multipath reflections in a cluttered yet contained space.
- **Building hallway (1.5m wide):** located just outside the room, with no restrictions on pedestrian traffic. This environment introduced unpredictable human movement and natural background noise typical of public indoor spaces.
- **Roadside curb:** an outdoor location with moderate vehicular traffic, chosen to test performance under high ambient acoustic noise and less predictable environmental dynamics.

At each location, we collected four sessions of data, each lasting 6 minutes. Following the same protocol as in the prior studies, each session included five repetitions of all 18 actions (Figure 5) presented in randomized order. Ground truth pose data were obtained using the same camera-based acquisition method as in earlier studies. Sensor data was streamed from the COTS smartwatch to a laptop over TCP, consistent with the procedure in User Study 2.

8.2 Results

We adopted an evaluation strategy for this semi-in-the-wild study consistent with the second user study using the COTS smartwatch. Specifically, we assessed EchoMotion under unseen-user and unseen-action scenarios through leave-one-participant-out and leave-one-action-out data splitting, as in studies 1 and 2. In addition, we examined the robustness of EchoMotion to environmental factors. Since the system relies on acoustic reflections, variations in room layout, objects, or people within the sensing range (approximately 1.5 meters) can alter the reflection patterns captured by the device.

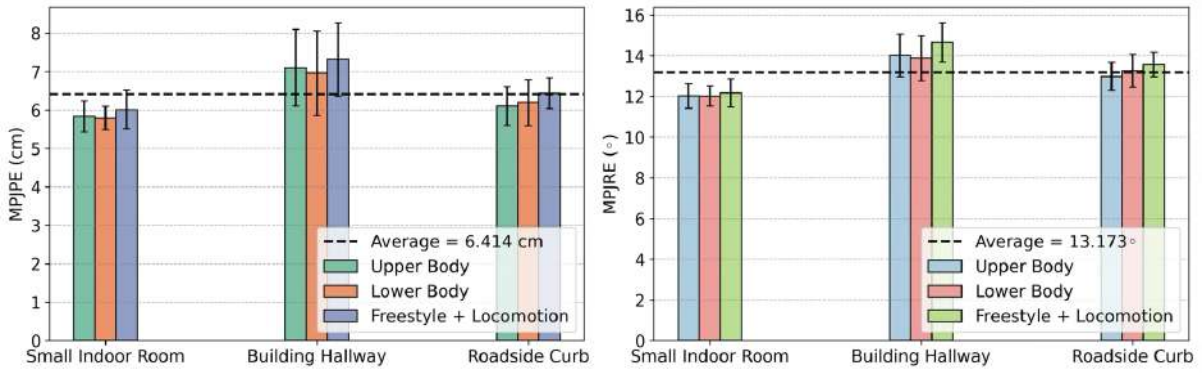


Fig. 12. User-independent evaluation of EchoMotion system in semi-in-the-wild setting.

8.2.1 Leave-One-Participant-Out Evaluation. We evaluated EchoMotion in a semi-in-the-wild setting for the scenario of encountering unseen users. Following the procedure from the previous studies, we trained the model on data from nine participants in each fold and tested on the held-out participant, repeating this process for all 10 users. The results of this user-independent evaluation are shown in Figure 12. The mean per-joint position error (MPJPE) averaged across all test users is 6.41 cm (SD = 1.64), and the mean per-joint rotation error (MPJRE) is 13.17 degrees (SD = 2.78), which is comparable to the COTS smartwatch results from Study 2 (MPJPE 6.18 cm, MPJRE 12.13 degrees), indicating only slight degradation in a semi-naturalistic setting.

Performance decreased more noticeably in the building hallway (MPJPE = 7.12 cm, MPJRE = 14.18 degrees), likely due to dynamic interference from other pedestrians affecting acoustic reflections. In contrast, static furniture in the small room had minimal impact because the differential echo profile input effectively cancels out stationary reflections. Overall, while dynamic human movement around the user introduced some error, EchoMotion maintained consistent performance across unseen users in semi-in-the-wild conditions (Figure 12).

8.2.2 Leave-One-Action-Out Evaluation. The goal of this evaluation was to assess the robustness of EchoMotion in scenarios involving unseen actions. To this end, we trained models on all participants and actions except one held-out action, repeating this process across all 18 actions in the semi-in-the-wild study. Figure 13 presents the per-action results. On average, EchoMotion achieved an MPJPE of 6.19 cm (SD = 1.53) and an MPJRE of 13.41 degrees (SD = 2.71).

We observed slightly higher errors for actions involving the right side of the body (e.g., right arm raise, right-leg lunges, dribbling with the right hand). Since participants wore the Samsung Galaxy Watch 6 on their left wrist, reflections from the contralateral side were weaker due to increased propagation distance and environmental perturbations across the three test locations. In addition, walking produced larger errors, likely due to interference from nearby moving individuals in dynamic environments. These findings highlight a limitation of single-wrist sensing, which could be mitigated by advanced denoising techniques or training larger foundation models on active acoustic signals.

8.2.3 Evaluation of Robustness Across Unseen Environments. The goal of this evaluation is to measure the robustness of EchoMotion in unseen environments. Because EchoMotion relies on acoustic reflections from both the user's body and surrounding objects within its sensing range (approximately 1.5 meters), factors such as environmental clutter, reflective surfaces, and nearby human activity are expected to affect pose tracking performance. To test this, we first trained a base model on COTS smartwatch data from all 15 participants in User Study 2. We then evaluated environmental robustness by holding out four sessions from each environment as

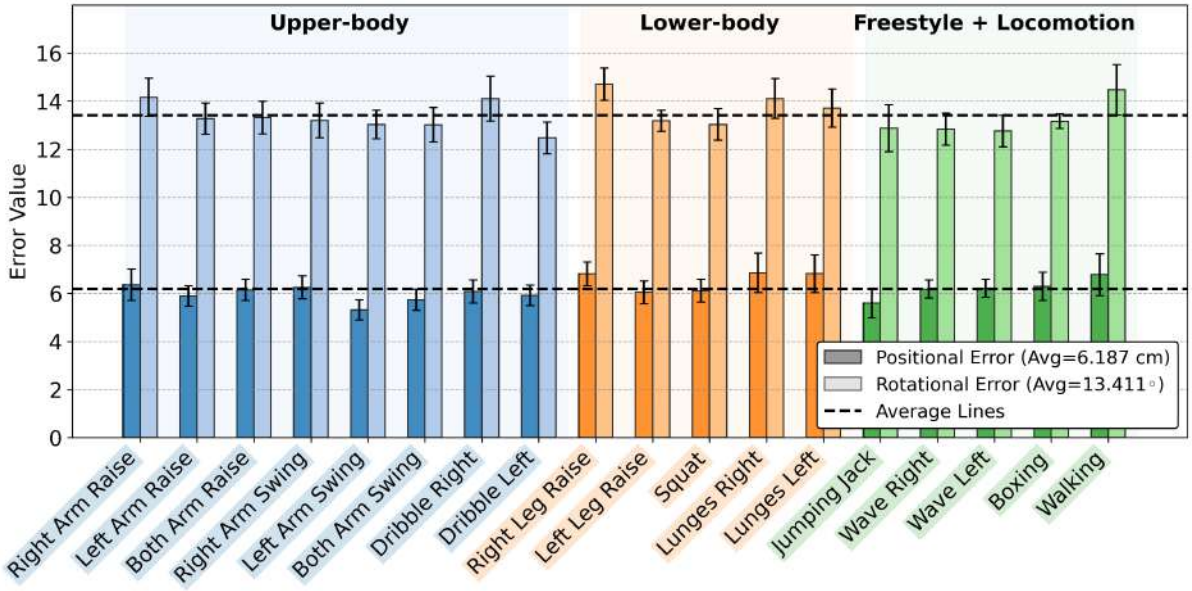


Fig. 13. Leave-one-action-out evaluation of EchoMotion system in semi-in-the-wild setting.

Table 5. Evaluation of EchoMotion's robustness to unseen environments in the semi-in-the-wild study.

	Large Experiment Room (User Study 2)	Small Indoor Room	Building Hallway	Roadside Curb
MPJPE (cm)	6.183 ± 1.286	6.087 ± 1.912	7.005 ± 2.759	6.574 ± 1.865
MPJRE (°)	12.131 ± 2.581	12.169 ± 2.104	13.858 ± 2.667	13.011 ± 2.883

the test set, while using the remaining eight sessions from the other two environments (same participant) to fine-tune the base model. Table 5 reports the mean per-joint position error (MPJPE) and mean per-joint rotation error (MPJRE) across all 10 participants for each environment: small indoor room, building hallway, and roadside curb.

Results show that the small indoor room, furnished with a table, chair, shelf, yoga mat, and two dumbbells, and with layouts rearranged across sessions, had little effect on performance compared to the controlled lab setting. This is likely because static reflections are minimized by the differential echo profile (computed as frame-to-frame differences) used as input to the deep learning model. In contrast, the building hallway, a narrow 1.5 m-wide passage, introduced more error, as multiple people were walking nearby while participants performed full-body poses within the sensing range, generating dynamic reflections that perturbed the acoustic profile. However, these perturbations remained weaker than reflections from the user's body due to their relative distance and shorter propagation time. Finally, we examined the effect of acoustic noise by evaluating at a roadside curb (average 67 dB) compared to the quieter lab setting in User Study 2 (36 dB). A frequency-domain analysis revealed that ambient noise across environments remained below 3.5 kHz, far from EchoMotion's operating band (>18 kHz). Accordingly, Table 5 shows that EchoMotion maintained robust performance under roadside noise.

In summary, environmental clutter and ambient noise had minimal effect on EchoMotion's performance, whereas dynamic human movement within the sensing range introduced the most notable degradation.

9 DISCUSSION

9.1 Comparison to Prior Work

Table 1 provides an overview of full-body pose tracking systems implemented on various wearable devices. As mentioned in Section 2, IMUs are the prevalent sensing modality for tracking full-body pose on smartwatches and other wearable devices. A good point of comparison is the recently published IMUPoser [68], which utilizes IMUs on commercial devices such as smartphones, smartwatches, and earbuds to track full-body pose through efficient data modeling and inverse kinematics-based joint refinement. Although evaluated on the publicly available DIP-IMU [32] dataset using a different train-test scheme, EchoMotion outperforms IMUPoser with lower positional error (5.3 cm MPJPE vs. 9.7 cm) and rotational error (11.7 degrees vs. 23.9 degrees) when trained on similar movement sets.

We can also compare EchoMotion with systems that utilize complementary sensing modalities alongside IMUs to track full-body pose. For example, Ultra Inertial Poser (UIP) [9] and SmartPoser [22] both use UWB ranging for pose tracking. While UIP [9] yields slightly better tracking performance (5.1 cm MPJPE, 10.7 degrees MPJRE), it requires six IMU+UWB multimodal sensing units placed at different body locations to capture precise full-body pose. SmartPoser [22], employing similar sensing techniques, tracks only the elbows and wrists of the instrumented arm with MPJPEs of 10.0 cm (SD=0.9) at the elbow and 15.1 cm (SD=1.2) at the wrist, resulting in a mean MPJPE of 12.6 cm (SD=1.1). UltraPoser [53] combines ultrasonic transmission and reception from the speaker–microphone pairs of a smartwatch and smartphone with IMU data from these devices and from earbuds. When all three devices are used, UltraPoser [53] achieves state-of-the-art accuracy (3.27 cm MPJPE, 7.37° MPJRE). However, excluding the phone leads to a noticeable performance drop, with errors increasing to 5.71 cm MPJPE and 14.43° MPJRE when using acoustic data from the smartwatch together with IMU data from the smartwatch and earbuds, evaluated on a set of 18 actions comparable to EchoMotion's dataset. Notably, UltraPoser [53] does not report full-body pose tracking performance with only a single smartwatch, making direct comparison with EchoMotion difficult.

In contrast, EchoMotion provides superior tracking performance while requiring just a single instrumentation point. As outlined in Table 1, EchoMotion demonstrates robust performance compared to other IMU-based systems that require three to six IMUs for reliable performance. EchoMotion makes a significant contribution by enabling smartwatches to track body parts that are not instrumented with any sensing device through active acoustic sensing. This advancement enables emerging applications that utilize pose information on the go to help users with fitness tracking, activity logging, and rehabilitation.

9.2 Application Demonstration: Exercise Tracking

Full-body pose tracking on wearables has been widely explored in prior work. To demonstrate the practical utility of EchoMotion, we implement an exercise-tracking system on a Samsung Galaxy Watch 6 that leverages differential echo profiles for pose estimation and a lightweight feedforward classifier for exercise recognition.

We conducted a pilot study with five participants across two gyms. Each participant performed four sessions per exercise, with 10 repetitions per session. The system supports 20 exercises, Bicep Curl, Chest Press, Crunch/Sit-ups, Glute Bridge, Hip Abduction, Jumping Jack, Lateral Raise, Lat Pull Down, Leg Extension, Leg Press, Lunges, Pec Fly, Plank, Push-ups, Russian Twist, Seated Cable Rows, Seated Leg Curl, Shoulder Press, Side Bends, and Squat, along with a null class for non-exercise activity. Participants performed exercises under naturalistic conditions, alongside other gym users, with no movement constraints and background music averaging 76 dB. Figure 14 shows representative acoustic echo profiles and corresponding IMU signals.

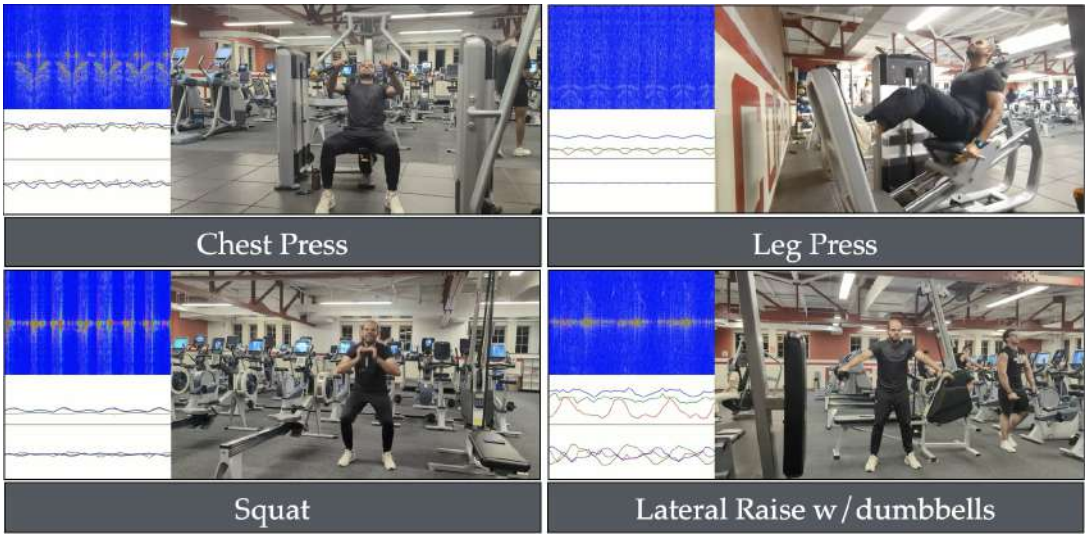


Fig. 14. Demonstration of the exercise-tracking application built on the EchoMotion system, which tracks 20 gym exercises using an off-the-shelf Samsung Galaxy Watch 6. Each panel shows the active acoustic signal (top) alongside accelerometer and gyroscope readings (bottom).

Using leave-one-participant-out evaluation, the system achieves an average F1-score of 0.954 ± 0.022. It also counts repetitions with a mean absolute error of ±0.97 (excluding plank). The underlying pose tracking module attains a mean per-joint position error of 9.75 cm and a rotation error of 17.96°, without requiring user-specific calibration. These results highlight the effectiveness of EchoMotion in enabling accurate, real-time exercise tracking on commodity smartwatches.

Beyond exercise recognition, EchoMotion opens up opportunities for broader wellness applications such as yoga, pilates, and tai chi, where fine-grained pose estimation and real-time feedback are essential. This positions EchoMotion as a foundation for next-generation wearable systems supporting personalized fitness, rehabilitation, and mindful movement in everyday environments.

9.3 Impact of Using IMUs as a Complementary Sensing Modality

As discussed in Section 2.2.1, IMUs are the most common sensing modality in wearable devices for human pose tracking. In our software-only implementation of EchoMotion on the Samsung Galaxy Watch 6, we simultaneously collected IMU data alongside active acoustic data and streamed both to the TCP server. Consequently, the datasets from User Studies 2 and 3 contain synchronized accelerometer, gyroscope, and acoustic data.

To evaluate the impact of IMUs as a complementary modality, we implemented a multimodal deep learning architecture. The model uses the same RepViT encoder (6.4M parameters) for acoustic data and an IMU encoder (1.2M parameters) adapted from IMU2CLIP [69], with late fusion of representations before pose prediction.

Results showed only marginal improvements over the acoustic-only baseline. In User Study 2 (COTS smartwatch, lab setting), leave-one-participant-out evaluation improved MPJPE from 6.183 cm to 6.117 cm and MPJRE from 12.131 degrees to 12.158 degrees (paired t-test, $p = 0.21$, not significant). For leave-one-action-out, MPJPE changed from 5.756 cm to 6.017 cm and MPJRE from 12.733 degrees to 13.213 degrees ($p = 0.34$, not significant). Similarly, in the semi-in-the-wild dataset (User Study 3), leave-one-participant-out evaluation yielded MPJPE of 6.388 cm and MPJRE of 12.981 degrees with acoustic+IMU, compared to 6.414 cm and 13.173 degrees with acoustic-only

($p = 0.27$, not significant). For leave-one-action-out, MPJPE slightly improved from 6.187 cm to 6.159 cm and MPJRE from 13.411 degrees to 13.338 degrees ($p = 0.41$, not significant).

These minimal gains came at the cost of an additional 1.2M parameters for the IMU encoder, leading to increased inference latency and higher computational overhead. We hypothesize that the IMU provides movement information only from the instrumented body part (left wrist), whereas acoustic sensing captures reflections from the entire body, including the left wrist, making IMU contributions largely redundant. Nevertheless, IMUs may still provide value in scenarios where acoustic signals are perturbed, serving as a fallback to maintain robustness. We envision future work exploring more efficient multimodal architectures that can dynamically select between acoustic and IMU streams to balance accuracy, power consumption, and latency for single-smartwatch full-body pose tracking.

9.4 Impact of Watch Placement Variations Across Users

In practice, users wear smartwatches in different configurations, such as on the left or right wrist and with varying orientations (e.g., display facing outward or inward). To evaluate the robustness of EchoMotion under such variations, we conduct a study with 5 new participants (no overlap with training data), following the same protocol as User Study 2 on the Samsung Galaxy Watch 6.

We consider four common wearing configurations: left wrist (outward), left wrist (flipped), right wrist (outward), and right wrist (flipped). Each participant performs 4 sessions per configuration (16 sessions per participant). We train a base model on data from User Study 2 (10 participants) and directly evaluate it on this dataset without retraining.

In the user-independent setting, performance remains consistent across configurations. The average MPJPE / MPJRE are 10.041 cm / 19.135° (left-outward), 9.772 cm / 18.548° (left-flipped), 10.265 cm / 18.972° (right-outward), and 9.608 cm / 18.181° (right-flipped). The small variation across configurations indicates that EchoMotion is largely robust to changes in wrist placement and orientation.

To further isolate the effect of unseen placement, we evaluate a user-adaptive setting. For each configuration, we fine-tune the base model using data from the same user collected under the other three configurations and test on the held-out configuration. This setup simulates adapting to a new placement without requiring direct calibration data for that placement. After adaptation, performance improves to 6.179 cm / 13.296°, 6.038 cm / 12.597°, 6.646 cm / 13.571°, and 5.913 cm / 12.016° for the four configurations, respectively.

Overall, these results show that EchoMotion generalizes well to unseen watch placements, with only minor degradation in the user-independent setting and substantial gains from lightweight adaptation. This robustness arises because the echo profile encodes relative motion with respect to the wrist-mounted sensor, making it less sensitive to global orientation changes. In contrast, generalization across users remains more challenging due to variability in individual motion patterns. We also observe slightly improved performance in flipped configurations, likely due to stronger acoustic reflections when the speaker is oriented toward moving body parts.

9.5 Limitations and Future Work

In this section, we discuss the challenges that need to be tackled before EchoMotion can be applied at scale on different watches, users, and body movements.

9.5.1 Limited Evaluation on Motion Artifacts. While our user studies include scenarios such as walking as part of the freestyle or locomotion subset and nearby pedestrian movement in a hallway in the semi-in-the-wild setting, we do not fully evaluate more dynamic conditions where users move (e.g., walking or running) while simultaneously performing other body movements. Such combined motions can introduce noticeable challenges for EchoMotion's active acoustic–based full-body pose tracking. Specifically, EchoMotion relies on characteristic reflection patterns and joint kinematics to infer the motion of uninstrumented body parts, as well as on the

model's ability to handle wrist motion. When users perform additional full-body translation (e.g., walking or running) alongside exercises such as bicep curls or lunges, these learned reflection patterns can be perturbed, leading to degraded pose estimation accuracy. Addressing this type of motion artifact likely requires larger-scale training data with targeted data augmentation, as well as techniques for disentangling motion components based on signal intensity and direction. While our results demonstrate the feasibility of full-body pose tracking using a single wrist-mounted device, a systematic evaluation of motion artifacts and methods to mitigate them remains an important direction for future work.

9.5.2 Cloth Coverage. As with other camera-based or wireless sensing-based wristband systems [21, 48, 55, 74], EchoMotion's active acoustic sensor cannot acquire noise-free signals when the sensing unit is obscured by clothing, particularly long sleeves. Consequently, users must wear short or rolled-up sleeves to avoid degrading the system's sensing performance. In future work, we plan to investigate denoising algorithms that compensate for cloth attenuation and to integrate complementary sensing modalities that activate when excessive coverage is detected, thereby ensuring robust tracking across diverse real-world scenarios.

10 CONCLUSION

In this paper, we present EchoMotion, an active acoustic sensing system for full-body pose tracking using a single smartwatch worn on the non-dominant hand. Our user studies, conducted in two implementation phases, first with a customized hardware prototype and then with a commercial off-the-shelf smartwatch, demonstrate that the system can accurately track 18 diverse full-body movements with low positional and rotational error across 30 participants in controlled and semi-in-the-wild settings. While future work will evaluate the system's robustness over a broader range of movements and real-world scenarios, EchoMotion provides a proof of concept for reliable full-body tracking on a single smartwatch without any hardware modifications to existing devices.

Acknowledgments

This project was supported by the National Science Foundation under Grant No. 2239569. Large language models (LLMs) were used to polish the writing of this paper.

References

[1] 2021. Kinect2 for Windows. https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows

[2] 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

[3] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-Go: Approximating User Pose with Smartphone Sensor Fusion and Inverse Kinematics. 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 9, 12 pages. https://doi.org/10.1145/3411764.3445582

[4] Karan Ahuja, Vivian Shen, Cathy Mengying Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. 2022. ControllerPose: Inside-Out Body Capture with VR Controller Cameras. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 108, 13 pages. https://doi.org/10.1145/3491102.3502105

[5] Amazon.com. [n. d.]. SanDisk Extreme Pro 32GB SDHC UHS-I Card (SDSDXXG-032G-GN4IN). https://www.amazon.com/SanDisk-Extreme-32GB-UHS-I-SDSDXXG-032G-GN4IN/dp/B01J5RHBQ4. [Online; accessed 21-May-2024].

[6] Apple. [n. d.]. Apple Vision Pro. https://www.apple.com/apple-vision-pro/. [Online; accessed 24-Mar-2025].

[7] Apple. [n. d.]. Apple Watch. https://www.apple.com/lae/watch/. [Online; accessed 24-Mar-2025].

[8] Riku Arakawa, Bing Zhou, Gurunandan Krishnan, Mayank Goel, and Shree K. Nayar. 2023. MI-Poser: Human Body Pose Tracking Using Magnetic and Inertial Sensor Fusion with Metal Interference Mitigation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 85 (Sept. 2023), 24 pages. https://doi.org/10.1145/3610891

[9] Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. 2024. Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 51, 11 pages. https://doi.org/10.1145/3641519.3657465

[10] Md Tanvir Islam Aumi, Sidhant Gupta, Mayank Goel, Eric Larson, and Shwetak Patel. 2013. DopLink: using the doppler effect for multi-device interaction. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 583–586. https://doi.org/10.1145/2493432.2493515

[11] Teo Babic, Florian Perteneder, Harald Reiterer, and Michael Haller. 2020. Simo: Interactions with Distant Displays by Smartphones with Simultaneous Face and World Tracking. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3334480.3382962

[12] Eric R. Bachmann, Robert B. McGhee, Xiaoping Yun, and Michael J. Zyda. 2001. Inertial and magnetic posture tracking for inserting humans into networked virtual environments. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology* (Baniff, Alberta, Canada) *(VRST '01)*. Association for Computing Machinery, New York, NY, USA, 9–16. https://doi.org/10.1145/505008.505011

[13] Steven S. Beauchemin and John L. Barron. 1995. The computation of optical flow. *ACM computing surveys (CSUR)* 27, 3 (1995), 433–466.

[14] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[15] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR* abs/1611.08050 (2016). arXiv:1611.08050 http://arxiv.org/abs/1611.08050

[16] Wenqiang Chen, Yexin Hu, Wei Song, Yingcheng Liu, Antonio Torralba, and Wojciech Matusik. 2024. CAvatar: Real-time Human Activity Mesh Reconstruction via Tactile Carpets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 151 (Jan. 2024), 24 pages. https://doi.org/10.1145/3631424

[17] Xiaowei Chen, Xiao Jiang, Lishuang Zhan, Shihui Guo, Qunsheng Ruan, Guoliang Luo, Minghong Liao, and Yipeng Qin. 2023. Full-body human motion reconstruction with sparse joint tracking using flexible sensors. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–19.

[18] Henry M. Clever, Ariel Kapusta, Daehyung Park, Zackory Erickson, Yash Chitalia, and Charles C. Kemp. 2018. 3D Human Pose Estimation on a Configurable Bed from a Pressure Image. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid, Spain). IEEE Press, 54–61. https://doi.org/10.1109/IROS.2018.8593545

[19] Apple Developer. [n. d.]. Core ML Framework. https://developer.apple.com/documentation/coreml/. [Online; accessed 31-Mar-2025].

[20] Android Developers. 2023. Android Debug Bridge (adb). https://developer.android.com/tools/adb. [Online; accessed 31-Mar-2025].

[21] Nathan Devrio and Chris Harrison. 2022. DiscoBand: Multiview Depth-Sensing Smartwatch Strap for Hand, Body and Environment Tracking. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 56, 13 pages. https://doi.org/10.1145/3526113.3545634

[22] 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* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 79, 11 pages. https://doi.org/10.1145/3586183.3606821

[23] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[24] Di Duan, Shengzhe Lyu, Mu Yuan, Hongfei Xue, Tianxing Li, Weitao Xu, Kaishun Wu, and Guoliang Xing. 2025. *Argus: Multi-View Egocentric Human Mesh Reconstruction Based on Stripped-Down Wearable mmWave Add-on.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3715014.3722045

[25] Mahmoud El-Gohary and James McNames. 2012. Shoulder and Elbow Joint Angle Tracking With Inertial Sensors. *IEEE Transactions on Biomedical Engineering* 59, 9 (2012), 2635–2641. https://doi.org/10.1109/TBME.2012.2208750

[26] ElectroIQ. [n. d.]. Smartwatch Statistics By Vendors, Revenue And Number Of Users. https://electroiq.com/stats/smartwatch-statistics/. [Online; accessed 31-Mar-2025].

[27] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14783–14794.

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[30] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[31] Dun-Yu Hsiao, Min Sun, Christy Ballweber, Seth Cooper, and Zoran Popović. 2016. Proactive Sensing for Improving Hand Pose Estimation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2348–2352. https://doi.org/10.1145/2858036.2858587

[32] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Trans. Graph.* 37, 6, Article 185 (Dec. 2018), 15 pages. https://doi.org/10.1145/3272127.3275108

[33] Dong-Hyun Hwang, Kohei Aso, and Hideki Koike. 2019. MonoEye: Monocular Fisheye Camera-based 3D Human Pose Estimation. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 988–989. https://doi.org/10.1109/VR.2019.8798267

[34] Intel. [n. d.]. Intel RealSense. https://www.intelrealsense.com/. [Online; accessed 25-Mar-2025].

[35] Stephen S. Intille, Ling Bao, Emmanuel Munguia Tapia, and John Rondoni. 2004. Acquiring in Situ Training Data for Context-Aware Ubiquitous Computing Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/985692.985693

[36] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022. AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing. In *Proceedings of European Conference on Computer Vision*. Springer.

[37] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2020. Coherent Reconstruction of Multiple Humans from a Single Image. In *CVPR*.

[38] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu Su. 2020. Towards 3D Human Pose Construction Using Wifi. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (London, United Kingdom) *(MobiCom '20)*. Association for Computing Machinery, New York, NY, USA, Article 23, 14 pages. https://doi.org/10.1145/3372224.3380900

[39] Yincheng Jin, Yang Gao, Yanjun Zhu, Wei Wang, Jiyang Li, Seokmin Choi, Zhangyu Li, Jagmohan Chauhan, Anind K. Dey, and Zhanpeng Jin. 2021. SonicASL: An Acoustic-based Sign Language Gesture Recognizer Using Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 67 (jun 2021), 30 pages. https://doi.org/10.1145/3463519

[40] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2017. End-to-end Recovery of Human Shape and Pose. *CoRR* abs/1712.06584 (2017). arXiv:1712.06584 http://arxiv.org/abs/1712.06584

[41] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. 2021. Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11510–11520.

[42] Rushil Khurana, Karan Ahuja, Zac Yu, Jennifer Mankoff, Chris Harrison, and Mayank Goel. 2018. GymCam: Detecting, Recognizing and Tracking Simultaneous Exercises in Unconstrained Scenes. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 185 (Dec. 2018), 17 pages. https://doi.org/10.1145/3287063

[43] Jiwan Kim, Chi-Jung Lee, Hohurn Jung, Tianhong Catherine Yu, Ruidong Zhang, Ian Oakley, and Cheng Zhang. 2026. WatchHand: Enabling Continuous Hand Pose Tracking On Off-the-Shelf Smartwatches. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, Article 740, 21 pages. https://doi.org/10.1145/3772318.3790932

[44] Dennis Kimm and David V Thiel. 2015. Hand speed measurements in boxing. *Procedia Engineering* 112 (2015), 502–506.

[45] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[46] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. M3Track: <U>Mm</U>Wave-Based <U>M</U>Ulti-User 3D Posture Tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services* (Portland, Oregon) *(MobiSys '22)*. Association for Computing Machinery, New York, NY, USA, 491–503. https://doi.org/10.1145/3498361.3538926

[47] Chi-Jung Lee, Jiaxin Li, Tianhong Catherine Yu, Ruidong Zhang, Vipin Gunda, François Guimbretière, and Cheng Zhang. 2025. Grab-n-Go: On-the-Go Microgesture Recognition with Objects in Hand. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 99 (Sept. 2025), 27 pages. https://doi.org/10.1145/3749469

[48] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 403, 21 pages. https://doi.org/10.1145/3613904.3642910

[49] Ke Li, Ruidong Zhang, Boao Chen, Siyuan Chen, Sicheng Yin, Saif Mahmud, Qikang Liang, Francois Guimbretiere, and Cheng Zhang. 2024. GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 497–512. https://doi.org/10.1145/3636534.3649376

[50] Ke Li, Ruidong Zhang, Siyuan Chen, Boao Chen, Mose Sakashita, François Guimbretière, and Cheng Zhang. 2024. EyeEcho: Continuous and Low-power Facial Expression Tracking on Glasses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[51] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[52] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. 2021. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. https://doi.org/10.48550/ARXIV.2111.12707

[53] Yadong Li, Shuning Wang, Yongjian Fu, Justin Chen, Xingyu Chen, Ju Ren, Xinyu
 Zhang, Akshay Gadre, and Ke Sun. 2025. UltraPoser: Pushing the Limits of IMU-based Full-Body Pose Estimation with Ultrasound Sensing on Consumer Wearables. In *ACM Symposium on User Interface Software and Technology*.

[54] Hyunchul Lim, Nam Anh Dang, Dylan Lee, Tianhong Catherine Yu, Jane Lu, Franklin Mingzhe Li, Yiqi Jin, Yan Ma, Xiaojun Bi, François Guimbretière, and Cheng Zhang. 2025. SpellRing: Recognizing Continuous Fingerspelling in American Sign Language using a Ring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 67, 17 pages. https://doi.org/10.1145/3706598.3713721

[55] Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. BodyTrak: Inferring Full-body Poses from Body Silhouettes Using a Miniature Camera on a Wristband. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 154 (sep 2022), 21 pages. https://doi.org/10.1145/3552312

[56] Jialin Liu, Dong Li, Lei Wang, Fusang Zhang, and Jie Xiong. 2022. Enabling Contact-free Acoustic Sensing under Device Motion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 128 (Sept. 2022), 27 pages. https://doi.org/10.1145/3550329

[57] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 143–152.

[58] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–16.

[59] LowPowerLab. [n. d.]. Current Ranger. https://lowpowerlab.com/guide/currentranger/. [Online; accessed 29-Nov-2023].

[60] Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma, Tomás Palacios, Antonio Torralba, and Wojciech Matusik. 2021. Intelligent carpet: Inferring 3d human pose from tactile signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11255–11265.

[61] Kian Mahmoodi, Yudong Xie, Tan Gemicioglu, Chi-Jung Lee, Jiwan Kim, and Cheng Zhang. 2025. EchoForce: Continuous Grip Force Estimation from Skin Deformation Using Active Acoustic Sensing on a Wristband. In *Proceedings of the 2025 ACM International Symposium on Wearable Computers* (Espoo, Finland) *(ISWC '25)*. Association for Computing Machinery, New York, NY, USA, 53–60. https://doi.org/10.1145/3715071.3750405

[62] Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 111 (sep 2023), 28 pages. https://doi.org/10.1145/3610895

[63] Saif Mahmud, Kian Mahmoodi, Chi-Jung Lee, Francois Guimbretiere, and Cheng Zhang. 2024. WristSonic: Enabling Fine-grained Hand-Face Interactions on Smartwatches Using Active Acoustic Sensing. *arXiv preprint arXiv:2411.08217* (2024).

[64] Saif Mahmud, Vineet Parikh, Qikang Liang, Ke Li, Ruidong Zhang, Ashwin Ajit, Vipin Gunda, Devansh Agarwal, Francois Guimbretiere, and Cheng Zhang. 2024. ActSonic: Recognizing Everyday Activities from Inaudible Acoustic Wave Around the Body. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 183 (Nov. 2024), 32 pages. https://doi.org/10.1145/3699752

[65] Saif Mahmud, M. T. H. Tonmoy, Kishor Kumar Bhaumik, A. M. Rahman, M. A. Amin, M. Shoyaib, Muhammad Asif Hossain Khan, and A. Ali. 2020. Human Activity Recognition from Wearable Sensor Data Using Self-Attention. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain*.

[66] J.L. Marins, Xiaoping Yun, E.R. Bachmann, R.B. McGhee, and M.J. Zyda. 2001. An extended Kalman filter for quaternion-based orientation estimation using MARG sensors. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No.01CH37180)*, Vol. 4. 2003–2011 vol.4. https://doi.org/10.1109/IROS.2001.976367

[67] Meta. [n. d.]. Meta Quest. https://www.meta.com/quest/. [Online; accessed 24-Mar-2025].

[68] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. https://doi.org/10.1145/3544548.3581392

[69] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: Language-grounded Motion Sensor Translation with Multimodal Contrastive Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13246–13253. https://doi.org/10.18653/v1/2023.findings-emnlp.883

[70] Hossein Mousavi Hondori and Maryam Khademi. 2014. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of medical engineering* 2014, 1 (2014), 846514.

[71] Movella. [n. d.]. Xsens Motion Capture. https://www.movella.com/products/motion-capture. [Online; accessed 25-Mar-2025].

[72] Movella. [n. d.]. xsens MTi 1-series datasheet. https://www.mouser.com/catalog/specsheets/xsens_DatasheetMTi1series.pdf?srsltid=AfmBOopW99Euhxk-FTmRjS8h0UxVheT2zvM53usaKN44n7auA49dKNVA [Online; accessed 8-Apr-2025].

[73] Movella. [n. d.]. xsens MTi 600-series. https://www.mouser.com/datasheet/2/1484/MTi_600_series_Datasheet-3193436.pdf [Online; accessed 8-Apr-2025].

[74] Rajalakshmi Nandakumar, Vikram Iyer, Desney Tan, and Shyamnath Gollakota. 2016. FingerIO: Using Active Sonar for Fine-Grained Finger Tracking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1515–1525. https://doi.org/10.1145/2858036.2858580

[75] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. 2020. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9890–9900.

[76] OptiTrack. [n. d.]. OptiTrack Motion Capture. https://www.optitrack.com/. [Online; accessed 25-Mar-2025].

[77] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin P. Murphy. 2017. Towards Accurate Multi-person Pose Estimation in the Wild. *CoRR* abs/1701.01779 (2017). arXiv:1701.01779 http://arxiv.org/abs/1701.01779

[78] Vineet Parikh, Saif Mahmud, Devansh Agarwal, Ke Li, François Guimbretière, and Cheng Zhang. 2024. EchoGuide: Active Acoustic Guidance for LLM-Based Eating Event Analysis from Egocentric Videos. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers* (Melbourne VIC, Australia) *(ISWC '24)*. Association for Computing Machinery, New York, NY, USA, 40–47. https://doi.org/10.1145/3675095.3676611

[79] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[80] Jihua Peng, Yanghong Zhou, and PY Mok. 2024. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1123–1132.

[81] Yili Ren, Zi Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. Winect: 3D Human Pose Tracking for Free-Form Activity Using Commodity WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 176 (dec 2022), 29 pages. https://doi.org/10.1145/3494973

[82] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang. 2022. GoPose: 3D Human Pose Estimation Using WiFi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 69 (jul 2022), 25 pages. https://doi.org/10.1145/3534605

[83] Rokoko. [n. d.]. Smartsuit Pro II. https://www.rokoko.com/products/smartsuit-pro. [Online; accessed 25-Mar-2025].

[84] Samsung. 2023. Samsung Galaxy Watch6. https://www.samsung.com/us/watches/galaxy-watch6/buy/galaxy-watch6-classic-47mm-black-lte-sm-r965uzkaxaa/. [Online; accessed 24-Mar-2025].

[85] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[86] Md Mobashir Hasan Shandhi, Karnika Singh, Natasha Janson, Perisa Ashar, Geetika Singh, Baiying Lu, D Sunshine Hillygus, Jennifer M Maddocks, and Jessilyn P Dunn. 2024. Assessment of ownership of smart devices and the acceptability of digital health data sharing. *NPJ digital medicine* 7, 1 (2024), 44.

[87] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. 2018. Closing the Gaps in Inertial Motion Tracking. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking* (New Delhi, India) *(MobiCom '18)*. Association for Computing Machinery, New York, NY, USA, 429–444. https://doi.org/10.1145/3241539.3241582

[88] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a Smartwatch and I can Track my User's Arm. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (Singapore, Singapore) *(MobiSys '16)*. Association for Computing Machinery, New York, NY, USA, 85–96. https://doi.org/10.1145/2906388.2906407

[89] Yuto Shibata, Yutaka Kawashima, Mariko Isogawa, Go Irie, Akisato Kimura, and Yoshimitsu Aoki. 2023. Listening human behavior: 3d human pose estimation with acoustic signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13323–13332.

[90] SteroLabs. [n. d.]. ZED 2i Stereo Camera. https://www.stereolabs.com/store/products/zed-2i. [Online; accessed 25-Mar-2025].

[91] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. 2019. xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera. In *Proceedings of the IEEE International Conference on Computer Vision*. 7728–7738.

[92] M. Tanjid Hasan Tonmoy, Saif Mahmud, A. K. M. Mahbubur Rahman, M. Ashraful Amin, and Amin Ahsan Ali. 2021. Hierarchical Self Attention Based Autoencoder for Open-Set Human Activity Recognition. In *Advances in Knowledge Discovery and Data Mining*, Kamal Karlapalem, Hong Cheng, Naren Ramakrishnan, R. K. Agrawal, P. Krishna Reddy, Jaideep Srivastava, and Tanmoy Chakraborty (Eds.). Springer International Publishing, Cham, 351–363.

[93] ultraleap. 2021. leap motion controller. https://www.ultraleap.com/product/leap-motion-controller/

[94] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. 2023. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5785–5795.

[95] Vicon. [n. d.]. Vicon Motion Capture. https://www.vicon.com/. [Online; accessed 25-Mar-2025].

[96] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. 2017. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Comput. Graph. Forum* 36, 2 (May 2017), 349–360. https://doi.org/10.1111/cgf.13131

[97] Bastian Wandt, James J. Little, and Helge Rhodin. 2021. ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses. https://doi.org/10.48550/ARXIV.2112.07088

[98] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15909–15920.

[99] Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. 2023. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13031–13040.

[100] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. 1, 4, Article 170 (jan 2018), 20 pages. https://doi.org/10.1145/3161188

[101] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[102] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6184–6193.

[103] Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. 2024. MobilePoser: Real-Time Full-Body Pose Estimation and 3D Human Translation from IMUs in Mobile Consumer Devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 70, 11 pages. https://doi.org/10.1145/3654777.3676461

[104] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, P. Fua, Hans-Peter Seidel, and Christian Theobalt. 2018. Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics* 25 (2018), 2093–2101.

[105] Jackie (Junrui) Yang, Tuochao Chen, Fang Qin, Monica S. Lam, and James A. Landay. 2022. HybridTrak: Adding Full-Body Tracking to VR Using an Off-the-Shelf Webcam. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 348, 13 pages. https://doi.org/10.1145/3491102.3502045

[106] Xiaoying Yang, Xue Wang, Gaofeng Dong, Zihan Yan, Mani Srivastava, Eiji Hayashi, and Yang Zhang. 2023. Headar: Sensing Head Gestures for Confirmation Dialogs on Smartwatches with Wearable Millimeter-Wave Radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 138 (Sept. 2023), 28 pages. https://doi.org/10.1145/3610900

[107] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[108] Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. TransPose: real-time 3D human translation and pose estimation with six inertial sensors. *ACM Trans. Graph.* 40, 4, Article 86 (July 2021), 13 pages. https://doi.org/10.1145/3450626.3459786

[109] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. 2023. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8818–8829.

[110] Tianhong Catherine Yu, Guilin Hu, Ruidong Zhang, Hyunchul Lim, Saif Mahmud, Chi-Jung Lee, Ke Li, Devansh Agarwal, Shuyang Nie, Jinseok Oh, François Guimbretière, and Cheng Zhang. 2024. Ring-a-Pose: A Ring for Continuous Hand Pose Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 189 (Nov. 2024), 30 pages. https://doi.org/10.1145/3699741

[111] Tianhong Catherine Yu, Manru Mary Zhang, Peter He, Chi-Jung Lee, Cassidy Cheesman, Saif Mahmud, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2024. SeamPose: Repurposing Seams as Capacitive Sensors in a Shirt for Upper-Body Pose Tracking. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 72, 13 pages. https://doi.org/10.1145/3654777.3676341

[112] Tianhong Catherine Yu, Manru Mary Zhang, Luis Miguel Malenab, Chi-Jung Lee, Jacky Hao Jiang, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2025. SeamFit: Towards Practical Smart Clothing for Automatic Exercise Logging. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 24 (March 2025), 22 pages. https://doi.org/10.1145/3712287

[113] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (Niagara Falls, New York, USA) *(MobiSys '17)*. Association for Computing Machinery, New York, NY, USA, 15–28. https://doi.org/10.1145/3081333.3081356

[114] Yu Zhan, Fenghai Li, Renliang Weng, and Wongun Choi. 2022. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization. https://doi.org/10.48550/ARXIV.2203.11471

[115] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-Obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 852, 18 pages. https://doi.org/10.1145/3544548.3580801

[116] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 37 (March 2020), 26 pages. https://doi.org/10.1145/3381008

[117] Yang Zhang, Chouchang (Jack) Yang, Scott E. Hudson, Chris Harrison, and Alanson Sample. 2018. Wall++: Room-Scale Interactive and Context-Aware Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3173574.3173847

[118] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3425–3435.

[119] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-Wall Human Pose Estimation Using Radio Signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7356–7365. https://doi.org/10.1109/CVPR.2018.00768

[120] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. 2018. RF-Based 3D Skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication* (Budapest, Hungary) *(SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 267–281. https://doi.org/10.1145/3230543.3230579

[121] Weixi Zhao, Yunjie Tian, Qixiang Ye, Jianbin Jiao, and Weiqiang Wang. 2021. GraFormer: Graph Convolution Transformer for 3D Pose Estimation. https://doi.org/10.48550/ARXIV.2109.08364

[122] Bo Zhou, Daniel Geissler, Marc Faulhaber, Clara Elisabeth Gleiss, Esther Friederike Zahn, Lala Shakti Swarup Ray, David Gamarra, Vitor Fortes Rey, Sungho Suh, Sizhen Bian, Gesche Joost, and Paul Lukowicz. 2023. MoCaPose: Motion Capturing with Textile-integrated Capacitive Sensors in Loose-fitting Smart Garments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 41 (March 2023), 40 pages. https://doi.org/10.1145/3580883

[123] 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 '25)*. Association for Computing Machinery, New York, NY, USA, Article 91, 16 pages. https://doi.org/10.1145/3746059.3747695

[124] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5745–5753.

[125] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15085–15099.

[126] Chengxu Zuo, Yiming Wang, Lishuang Zhan, Shihui Guo, Xinyu Yi, Feng Xu, and Yipeng Qin. 2024. Loose inertial poser: Motion capture with IMU-attached loose-wear jacket. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2209–2219.