UltraPoser: Pushing the Limits of IMU-based Full-Body Pose Estimation with Ultrasound Sensing on Consumer Wearables

Yadong Li*
University of Washington
Seattle, WA, USA
yadongli@uw.edu

Shuning Wang*†
Central South University
Changsha, Hunan, China
shuning.wang@csu.edu.cn

Yongjian Fu‡
Central South University
Changsha, Hunan, China
fuyongjian@csu.edu.cn

Justin Chen
University of California San Diego
La Jolla, CA, USA
juc059@ucsd.edu

Xingyu Chen
University of California San Diego
La Jolla, CA, USA
xic063@ucsd.edu

Ju Ren
Tsinghua University
Bejing, China
renju@tsinghua.edu.cn

Xinyu Zhang
University of California San Diego
La Jolla, CA, USA
xyzhang@ucsd.edu

Akshay Gadre
University of Washington
Seattle, WA, USA
gadre@uw.edu

Ke Sun
University of Michigan
Ann Arbor, MI, USA
kesuniot@umich.edu

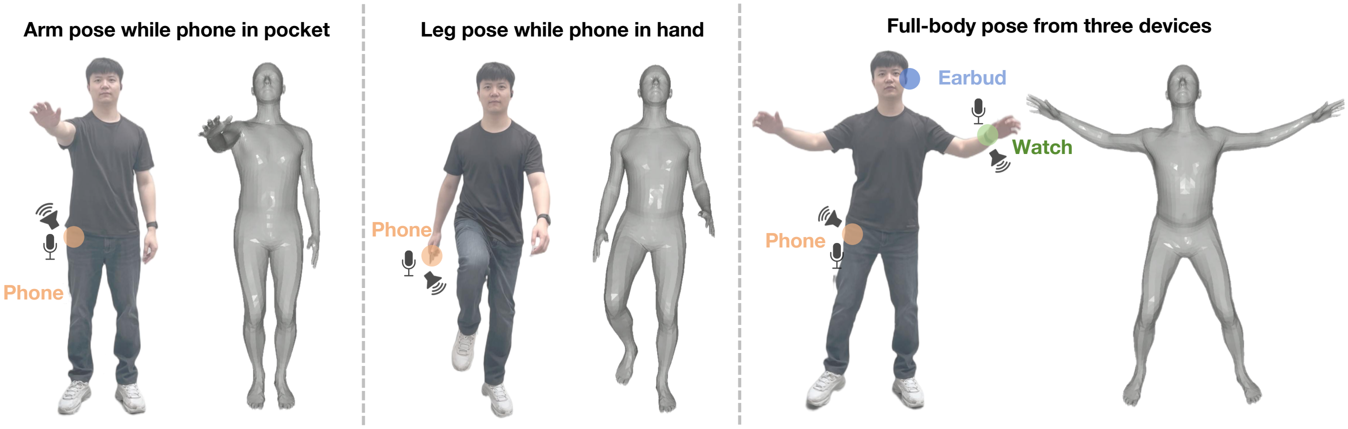

Arm pose while phone in pocket
Leg pose while phone in hand
Full-body pose from three devices
Phone
Phone
Earbud
Watch
Phone

Figure 1: ULTRAPOSER **enables ubiquitous full-body pose estimation by integrating ultrasound sensing and IMU using commodity wearable devices. In addition to measuring IMU data, a smartphone and smartwatch are used to transmit and receive ultrasound signals. The extracted ultrasound features capture motions from joints without any attached devices and offer drift-free range measurements to complement IMU data for more accurate pose estimation.**

Abstract

Full-body motion capture using IMUs embedded in consumer wearables has the potential to enable convenient, on-the-go tracking with minimal instrumentation. However, the sparse placement of these devices on the body frame presents challenges such as limited body coverage, reduced motion feature diversity, and cumulative drift errors. This paper introduces ULTRAPOSER, a multi-modal full-body motion capture system that integrates ultrasonic sensing with inertial measurements for improved fidelity, broader coverage and increased reliability. ULTRAPOSER leverages built-in microphones and speakers on commodity wearables, such as smartphones and smartwatches, to transmit and receive inaudible ultrasound signals, expanding the range of sensed body areas and providing drift-free acoustic multipath profiles. To implement ULTRAPOSER, we systematically explore ultrasound signal designs to maximize feature quality and propose a graph-based physics-aware fusion architecture to integrate heterogeneous sensing modalities. We evaluate our approach using the ULTRAPOSER Dataset, collected from 10 participants across diverse device placements and activity contexts. Compared to state-of-the-art IMU-only methods, ULTRAPOSER achieves

*Both authors contributed equally to this research.
†This work was conducted when the author was a visiting student at the University of Michigan, Ann Arbor.
‡This work was conducted when the author was a visiting student at the University of California San Diego.

UIST '25, September 28 – October 1, 2025, Busan, Republic of Korea
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2037-6/2025/09
https://doi.org/10.1145/3746059.3747714

a 28.46% improvement in overall pose estimation accuracy and up to 67.28% error reduction for specific limbs without directly attached sensors.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing**.

Keywords

Motion capture, ultrasound sensing, multi-modal fusion

ACM Reference Format:
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 *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28 – October 1, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3746059.3747714

1 Introduction

Imagine a future where a user can deal with physical injuries in the comfort of their home while the doctor is able to remotely monitor their rehabilitation plan and progress accurately. The ability to capture human pose ubiquitously can enable a wide spectrum of applications, from life logging [15] and rehabilitation [6, 22] to adaptive interfaces [1]. Today, these applications mostly employ vision-based pose estimation systems that are sensitive to lighting conditions [24, 38] and sometimes rely on active markers [32]. More importantly, these pose estimation systems only operate in line-of-sight conditions. Thus, wearable sensor-based alternatives [30, 31] were developed to mitigate these issues where the human pose is measured in-situ by a large number of inertial measurement unit (IMU) sensors – as many as 17. While these systems make the tracking process feasible by overcoming the visual sensing challenges, they typically make the pose tracking process quite unwieldy for the user. Thus, there is a growing need for developing solutions that can ubiquitously and accurately track human full-body pose using sparse off-the-shelf commodity devices.

Recent work has made significant strides in this direction. IMUPoser [29] estimates full-body pose using IMUs embedded in everyday consumer devices such as smartphones, smartwatches, and earbuds. This setup not only reduces the number of sensors a user must wear but also leverages devices they already carry or wear regularly. However, this location-sparse sensor configuration significantly limits body coverage and constrains the diversity of motion features. To mitigate this, MobilePoser [45] introduces a physics-based optimization step to enforce plausible human kinematics. While this improves accuracy to a certain degree, the system remains fundamentally under-constrained due to its limited sensing inputs. Additionally, purely IMU-based methods inherently suffer from cumulative errors due to the nature of inertial sensing. Consequently, pose estimates are prone to global drift and degradation in accuracy since they lack direct position measurements. While prior works [3, 7] have attempted to address this issue by combining IMU measurements with UWB ranging, these systems are either incompatible with commodity devices or are limited to tracking only specific body parts. Given these constraints, there remains a clear gap for ubiquitous full-body pose estimation systems that can operate using commodity devices without suffering this reduction of scope and accuracy.

This paper presents UltraPoser, a multimodal full-body pose estimation system that integrates acoustic sensing with IMUs to enrich feature space and capture spatial relationships between body joints. Our key insight is that the built-in microphones and speakers on commercial smartphones and smartwatches can be repurposed to transmit and receive inaudible ultrasound signals to expand the scope of the sensed body area. However, unlike prior work on ultrasonic ranging, UltraPoser dissects the complete acoustic multipath profile to sense additional body parts and measure distances that enhance both accuracy and robustness in pose estimation. In a nutshell, UltraPoser operates as follows: We continuously measure the IMU measurements at three common on-body mobile/wearable devices – *the watch, the phone, and the earbuds*. We also make the phone and watch transmit inaudible ultrasound signals which are captured across the other devices. We extract valuable features from these received acoustic signals such as Doppler velocity and range profiles . We then fuse the user's physical prior context (such as whether the phone was in the pocket or in hand) with the measured data to extract the full-body pose of the user. Given the above simplified operation profile of the system, there are two fundamental challenges that need to be addressed to maximize the accuracy and robustness of the system: (a) *how can we design the ultrasound sensing setup to seamlessly operate concurrently while maximizing the sensing quality?*; and (b) *how can we effectively fuse heterogeneous signals—measured from different sensors with different placements into a coherent pose estimation framework?*

Optimal Feature-Driven Ultrasound Profile Design: With the goal of maximizing the valuable information extracted from the inaudible acoustic signals transmitted by the watch and the phone, we perform a large scale measurement study (Sec. 3) to evaluate the impact of various signal designs (e.g., single-frequency vs. wideband) on the ability to extract acoustic features (e.g., Doppler vs. range) across different sensing distances. This provides valuable insights into how to optimize ultrasound configurations for different devices, thereby maximizing the quality and diversity of the extracted features. Our measurements, ablation analysis and design criterion along with the final acoustic profile design are detailed in Sec. 5.

User Context-Driven Multimodal Sensor Fusion: We leverage the key insight that different devices and sensing modalities can be attached to distinct body parts and are therefore responsible for reconstructing different regions of the body. Based on this observation, we design a multi-modal fusion framework that incorporates user context into a graph convolutional neural network (GCN). This architecture models the human body as a graph structure and integrates physical prior knowledge of spatial correlation between different modalities and devices, thus enabling accurate and robust full-body pose estimation. We provide more details in Sec. 6.

We implement UltraPoser using off-the-shelf smartphones, smartwatches, and earbuds and collect over 334,000 pose measurements across 18 action classes over 10 users. This data is collected

Table 1: Comparison of UltraPoser with prior wearable-based motion capture systems

Project	Modality	Number of Sensors	Commodity Devices	Full-body Tracking	Absolute Range
DIP-IMU [14]	IMU	6	✘	✔	✘
IMUPoser [29]	IMU	≤ 3	✔	✔	✘
Ultra Inertial Poser [3]	IMU+UWB	6	✘	✔	✔
SmartPoser [7]	IMU+UWB	2	✔	✘	✔
PoseSonic [27]	Ultrasound	2	✘	✘	✔
UltraPoser	**IMU+Ultrasound**	**3**	✔	✔	✔

from users across diverse ages, heights, and weights in multiple different environments to form a comprehensive dataset for evaluation of UltraPoser. The users perform a diverse set of actions across different device configurations (e.g., phone in hand vs. in pocket). With the baseline of considering IMUPoser [29], our experimental results demonstrate a **28.46%** improvement in overall human pose estimation and up to **67.28%** error reduction in tracking specific limbs.

Our contributions are summarized as follows:

- We present the first system that combines ultrasound and IMU sensing from sparse consumer devices for full-body pose estimation.
- We propose a graph-based, physics-aware sensor fusion framework that leverages the spatial complementarity of different sensing modalities and body regions.
- We conduct comprehensive evaluations and benchmarking experiments, demonstrating the effectiveness of each system component and design choice.[1]

Limitations: Despite the significant improvement provided by UltraPoser, there are several limitations that need to be addressed in the future (more details in Sec. 8): (1) **Hardware limitations** of commodity devices limit acoustic signal extraction capability. (2) **Device placement** of the watch and phone is required to be on different sides of the body for lateral coverage. (3) **Operability** of ultrasound frequencies needs to be evaluated for continuous transmission in the presence of humans and other animals.

2 Related Work

We categorize prior work on non-visual motion capture into three main classes based on their sensing modalities and compare UltraPoserwith the most representative related work in Table. 1.

Motion Capture with Wearable IMUs: IMU-based full-body motion capture systems are robust across various environments and have been widely adopted in state-of-the-art commercial solutions such as Xsens [30] and Noitom [31]. While highly accurate, these systems are expensive, cumbersome, and impractical for daily use, as they require wearing more than ten specialized IMUs integrated into a full-body suit. To improve usability, recent work focuses on reducing the number of required IMUs. One line of research uses specialized IMUs placed on six key joints (e.g., DIP-IMU [14], TransPose [51]), leveraging deep learning models to infer full-body pose. Other studies explore advanced learning models that enhance accuracy [18, 50, 52], generalization [54], and efficiency [44]. Another direction uses embedded IMUs in consumer devices like smartphones, smartwatches, and earbuds [29, 45], or VR hardware [8, 16, 36, 43, 56]. These systems are more affordable and better suited for everyday use. However, they face a much more under-constrained problem due to the extreme sparsity of sensors—typically fewer than 3 IMUs. Moreover, IMUs provide only relative measurements and suffer from drift, especially with consumer-grade devices.

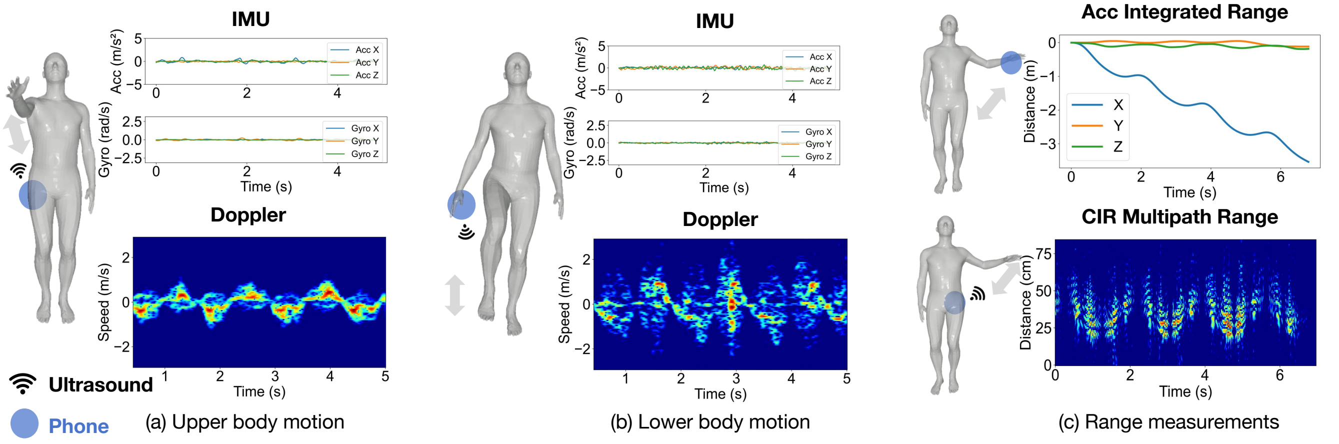
To address the challenges, UltraPoser repurposes the built-in microphones and speakers in consumer devices to perform ultrasound sensing. It enhances sensing coverage of human body joints not directly monitored by IMUs, and provides additional absolute range information, mitigating accumulation errors inherent in IMU-only systems.

Motion Capture with Wireless Signals: Wireless sensing leverages how human motion alters signal propagation to estimate pose. Most existing systems are device-free, relying on external transceivers (e.g., mmWave [5, 21, 46, 55] or Wi-Fi [17, 33, 47]) to sense human movement via multipath reflections. In the acoustic domain, Shibata et al. [34] proposed using audible chirp signals with external speakers and microphones for acoustic feature extraction and 3D pose estimation. However, these systems often require specialized hardware, operate only in constrained spaces, and degrade in complex environments due to multipath interference. Recent efforts have integrated wireless sensors into head-mounted wearable devices. For example, mmEgo [21] employs a head-mounted mmWave radar to generate 3D point clouds for posture estimation. PoseSonic [27] embeds microphones and speakers into smart glasses to enable acoustic-based upper-body pose tracking. Although these approaches improve mobility, they still rely on custom hardware and are typically limited to partial-body tracking.

UltraPoser distinguishes itself from this line of work by leveraging existing microphones, speakers, and IMUs on ubiquitous commercial devices such as smartphones and smartwatches. This design enables mobile, full-body tracking while maintaining compatibility with the devices users already carry in their daily lives.

Motion Capture with Multi-modal Sensors: To overcome the limitations of single-modality systems, prior work has explored combining IMUs with complementary modalities to improve the accuracy and robustness of pose estimation. For example, Lee et al. have combined head poses derived from head-mounted cameras with IMUs worn on smartwatches to estimate full-body motion [20]. Other strategies enhance IMU-based systems with additional sensors, such as pressure sensors (PressInPose [10]) and magnetic

[1] Dataset, code and model are available at https://github.com/leeyadong/UltraPoser



(a) Upper body motion (b) Lower body motion (c) Range measurements

Figure 2: Benefits of ultrasound sensing. Ultrasound signals capture joint motions of both the upper body (a) and lower body (b) that are not directly monitored by IMUs. (c) Ultrasound provides multipath-based absolute ranges free from the IMU drift.

sensors (MI-Poser [2]). While effective, these approaches often involve obtrusive setups or rely on custom hardware not available in commodity devices. Other ultra-wideband (UWB) radios-based systems, such as Ultra Inertial Poser [3] and SmartPoser [7], measure absolute distances between devices to mitigate IMU drift. However, these systems rely solely on device-to-device range measurements and do not leverage multipath propagation, which can offer additional motion cues—particularly valuable when the number of devices is limited.

In contrast, UltraPoser moves beyond simple device-to-device range estimation by measuring multipath range profiles and Doppler shifts, which capture both line-of-sight and multipath propagation characteristics. This approach provides detailed insights into absolute and relative path changes induced by human motion, including reflected signals from different body parts, thus enabling more accurate and robust posture estimation.

3 Motivation Study – Can Ultrasound Sensing Enhance Multi-Modal Pose Estimation?

In this section, we demonstrate how ultrasound sensing can complement IMU-based motion tracking by expanding the scope of sensing and extracting spatial features to recover human pose more reliably. To simplify the analysis, this motivation study focuses on a single-device scenario in which the user carries only a smartphone, either held in hand or placed in a trouser pocket. More complex multi-device configurations, involving smartwatches and earbuds, will be discussed in the following sections.

Ultrasound Sensing Primer: Ultrasound sensing uses high frequency acoustic signals (typically around 20 kHz when implemented on consumer wearables) to perform various sensing tasks [4, 25, 28]. A typical ultrasound sensing system consists of a speaker that transmits sound waves and a microphone that receives the reflected signals. By analyzing how these signals propagate and reflect off the human body and surrounding environment, the system can extract rich motion and spatial information. In this paper, we consider two complementary types of features from the received ultrasound signal: the Doppler spectrum and the Channel Impulse Response (CIR).

The Doppler spectrum reflects frequency shifts in the received signal caused by motion—either from the target or the sensing device itself. These shifts, known as the Doppler effect, are directly related to velocity. The CIR, in contrast, characterizes how the ultrasound signal travels through multiple propagation paths before reaching the receiver. Given the complex baseband transmitted probe frame $t[n]$ of length N and the received samples $r[n]$, the CIR $h[\tau]$ can be estimated by cross-correlation:

$$h[\tau] = \frac{1}{N} \sum_{n=0}^{N-1} t[n]\, r^{*}[n-\tau], \tag{1}$$

where τ indexes the possible propagation delays and $*$ denotes conjugated operation. Assume the baseband sampling rate is f_s, so that one sample shift corresponds to a time increment $\Delta t = \frac{1}{f_s}$. For each delay index τ_k of $h[\tau_k]$, the round-trip propagation time is $t_k = \tau_k\, \Delta t$. The corresponding one-way path length is

$$d_k = \frac{c\, t_k}{2} = \frac{c\, \tau_k}{2 f_s}, \tag{2}$$

with c is the speed of sound. Thus, we can obtain the ranges of the reflecting objects from $h[\tau]$, with its amplitude indicating the strength of that reflection.

Enhancing Feature Diversity: Our first goal is to demonstrate how ultrasound sensing enriches motion features and mitigates the limitations of sparse devices. We achieve this by evaluating two scenarios where the smartphone (and hence the IMU) is not located on the actively moving joint: *(1)* the smartphone is placed in a trouser pocket while the user performs upper body motions (i.e., arm push), and *(2)* the smartphone is held in hand during lower-body motions (i.e., leg kick). As shown in Fig. 2a and Fig. 2b, in both scenarios, the IMU alone cannot accurately capture motion at joints, it is not directly attached to (the peak values of acceleration and angular velocity are only 0.73 m/s^2 and 0.18 rad/s, respectively).

Next, we configure the smartphone to perform ultrasound-based motion tracking by transmitting a 17 kHz single-tone sine wave.

The received signal is first passed through a notch filter to remove the carrier frequency, and then processed using a Short-Time Fourier Transform (STFT) to extract the Doppler spectrum, which captures frequency shifts induced by body movement. As shown in Fig. 2a and Fig. 2b, the ultrasound Doppler features exhibit rich distinct patterns associated with different motion speeds. This clearly demonstrates that ***ultrasound acoustic response can capture complementary information that reflects movements of body parts beyond the IMU's direct measurement scope.***

Robustness to IMU Cumulative Drift: As we know, IMUs inherently suffer from drift and provide only relative measurements, i.e., acceleration and angular velocity, at each attached joint. In contrast, we aim to show that ultrasound sensing can provide absolute ranging information by measuring the time-of-flight (ToF) of sound reflected off different body parts. We demonstrate it by configuring the smartphone to transmit a wide-band ultrasound signal spanning 4 kHz (from 18 kHz to 22 kHz) and extract the CIR to estimate the propagation distance of each reflection path (details in Sec. 5). This setup yields a range resolution of approximately $\Delta R = c/2B = 4.3$ cm, where $c = 343$ m/s is the speed of sound and $B = 4$ kHz is the signal bandwidth. The smartphone is placed in the user's trouser pocket while the user performs upper-body motion (i.e., arm lifting).

The CIR shown in Fig. 2c captures the absolute range variations in path length as the user moves their arms. For comparison, we repeat the same motion with the smartphone held in hand to collect acceleration data from the IMU, and then apply double integration to estimate range variation. The ***IMU-derived range measurements exhibit significant deviation over time due to cumulative error, while the ultrasound-based measurements remain accurate and stable***. This demonstrates the potential of leveraging ultrasound sensing to provide drift-free, absolute multipath range profiles for accurate full-body pose estimation.

4 System Overview

UltraPoser achieves ubiquitous full-body pose estimation via the following stages: First, it continuously records IMU data from all three devices (i.e., a smartwatch, a smartphone, and a pair of wireless earbuds). Second, both the smartwatch and the smartphone transmit ultrasound signals while simultaneously recording audio. This setup enables multiple acoustic links between these smart devices. Although ultrasound sensing is theoretically feasible on any device equipped with both a speaker and a microphone, hardware limitations (Sec. 8) restrict our current implementation to a subset of links, including phone-to-phone, watch-to-watch, and watch-to-phone. Next, UltraPoser extracts motion-related ultrasound sensing features from the recorded acoustic signals, including Doppler velocity via STFT and CIR via cross-correlation [40]. These multi-modal features, along with the IMU data, are then fused using a machine learning-based pose estimator to generate accurate full-body pose representations. Fig. 3 describes this pipeline.

The rest of this paper introduces our approach to solving two fundamental challenges for maximizing the quality and robustness of pose estimation by extracting valuable features from the available resources. (1) **Designing optimal ultrasound profile:** Sec. 5 describes how we systematically design ultrasound signal profiles to enhance sensing quality and extract optimal motion features. (2) **Graph-based multi-modal sensor fusion:** Sec. 6 introduces our multi-modal fusion framework, which exploits the skeletal graph structure of the human body through graph-based modeling while incorporating user-contextual physical knowledge to achieve accurate pose estimation.

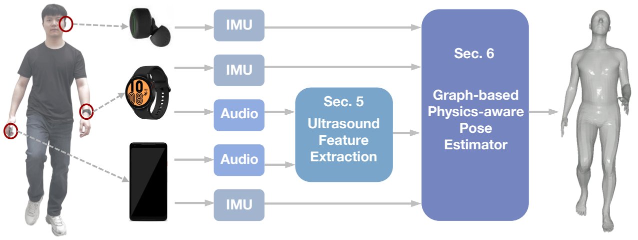

IMU
IMU
Audio
Audio
IMU
Sec. 5
Ultrasound Feature Extraction
Sec. 6
Graph-based Physics-aware Pose Estimator

Figure 3: UltraPoser Overview.

5 Ultrasound Profile Design

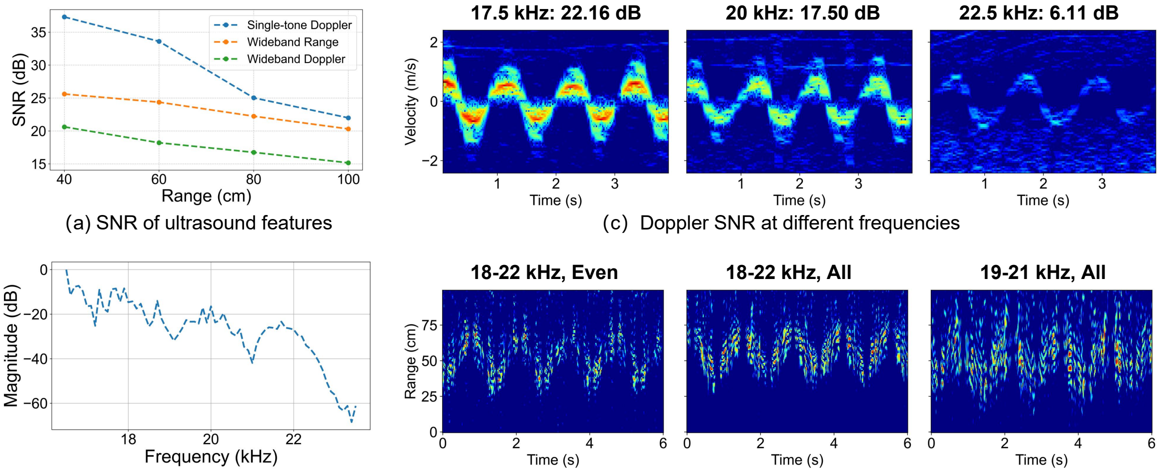
While ultrasound sensing offers significant advantages over IMUs, as discussed in Sec. 3, fully leveraging its potential for full-body posture estimation requires meeting the following key requirements. *(1) High-SNR Doppler and CIR measurements with long sensing range*: Accurate pose estimation depends on high-SNR Doppler shifts to capture additional motion cues, and CIR to compensate for IMU drift. Moreover, a sensing range of at least 1 meter is needed to cover distant limbs like feet or hands. *(2) Concurrent ultrasound sensing with multiple devices*: Both the phone and watch must be capable of simultaneously extracting ultrasound features from different body parts without causing mutual interference.

However, achieving these goals on commodity wearables is challenging due to hardware limitations. The audio system in commodity smartphones and smartwatches is quite heterogeneous and primarily optimized for audible sound playback rather than ultrasound sensing. As a result, usable frequency bands and output power are limited, constraining sensing range and degrading SNR. Thus, a carefully designed ultrasound signal profile (which device transmits what signal) is required to match device capabilities while maximizing sensing performance.

To overcome the above challenges and meet the system requirements, we conduct a series of microbenchmarks and design our ultrasound signal profile based on two key insights: *(1) Single-tone signals provide higher-SNR Doppler measurements than wideband signals. (2) Proper frequency division enables simultaneous ultrasound transmission without interference or loss of resolution*, as detailed below.

5.1 How to best capture Doppler and CIR measurements using ultrasound?

Prior ultrasound-based sensing systems typically use either single-tone signals [9, 35] for Doppler-based velocity estimation, or wideband waveforms [41, 42] for extracting both Doppler and ToF-based CIR. While wideband signals offer richer features, they suffer from lower SNR due to energy spread across frequencies—particularly



(a) SNR of ultrasound features

(c) Doppler SNR at different frequencies

(b) Frequency reponse

(d) CIR with different frequency bands and subcarriers

Figure 4: Ultrasound feature extraction insights: (a) Doppler features extracted from single-tone signals are more robust to noise than those derived from wideband signals. (b) The smartphone's frequency response shows significant degradation in received signal quality at higher frequency bands. (c) Lower-frequency tones demonstrate higher SNR in capturing Doppler profiles. (d) CIR range measurement using a subset of subcarriers achieves a similar range resolution to using all subcarriers.

degrading Doppler quality. In contrast, single-tone signals concentrate energy at a specific frequency, yielding stronger echoes and higher SNR at longer ranges.

We empirically explore the trade-offs for capturing the Doppler between signal types. We configure a smartphone (Pixel 2 XL) to transmit either (1) a 20 kHz single-tone or (2) a wideband signal spanning 18-22 kHz. During the experiment, the phone remains stationary while the user performs push gestures from distances of 40, 60, 80, and 100 cm. We compute SNRs by measuring the peak signal power relative to background noise for both Doppler (from the single-tone) and Doppler/CIR (from the wideband signal). As shown in Fig. 4 (a), the single-tone consistently achieves higher Doppler SNR, especially at longer ranges. Given this insight, we further explore the frequency at which the commodity hardware enables the best performance. As shown in Fig. 4 (b), the smartphone (Pixel 2 XL) exhibits significantly reduced gain above 22 kHz. The impact of this gain is further quantified in Fig. 4 (c), where we transmit single-tone signals at 17.5, 20, and 22.5 kHz and see a significant SNR drop at higher frequencies. This leads us to the first key insight – **Key Insight 1: Narrowband tone can measure Doppler more robustly than wideband tone and it can do it better at lower frequencies.**

Next, we aim to evaluate the efficacy and ability of a wideband acoustic signal to capture the CIR of the environment. The first parameter that captures that limit is the bandwidth. As is well known in the literature, the higher the bandwidth, the better the resolution. We choose a 4 kHz bandwidth between 18-22 kHz that enables a 4.3 cm range resolution. This choice is made to optimize for three factors: avoid audible bands, enable maximum resolution, compatible with commodity hardware gains. The next factor however is the coexistence of these devices. Specifically, we expect the watch and the phone (and perhaps even the earbuds in the future) to capture the CIR simultaneously. This means the planned bandwidth of 4 kHz needs to be shared across time or frequency. The first approach could be time-division multiplexing (TDM) where each device takes turns transmitting the wideband signal. However, given these devices are separated and the length of our designed chirps will be several μs, the lack of synchronization will lead to signal leakage across the devices. Hence, we take a frequency division multiplexing (FDM) where they each transmit across different frequencies. However, there are two ways, the devices can share frequencies - (1) a half-and-half approach where bandwidth is split down the middle or (2) interspersed signals at alternate subfrequencies.

We conduct an experiment to validate this design, by configuring the smartphone to transmit three types of signals: (1) all subcarriers across 18–22 kHz, (2) only even subcarriers across 18–22 kHz, (3) all subcarriers across a narrower 19–21 kHz band. During the experiment, the smartphone remains stationary while a user performs a push gesture at a distance of 50 cm. As shown in Fig. 4 (d), the interspersed-subcarrier configuration (2) achieves range resolution comparable to the full-bandwidth case (1), and significantly outperforms the narrower-bandwidth baseline (3). Furthermore, the CIR correlation coefficient between the even subcarriers (2) and the full 18–22 kHz range (1) is 0.80, compared to 0.75 for the 19–21 kHz band (3), demonstrating superior signal similarity achieved by the even-odd subcarrier allocation. The only loss due to interspersed measurements is primarily the reduction in the maximum range of sensing the CIR which can be overcome by sending a longer duration symbol. **Key Insight 2: Interspersed frequency-division multiplexing can achieve spectrally efficient CIR estimates with reasonable resolution for tracking pose.**

5.2 Concurrent Ultrasound Sensing Profile Design and Feature Extraction

In this section, we describe the design of our ultrasound acoustic signals being transmitted from the two devices – the phone and the watch. Our primary goal is to maximize the ability to capture both the Doppler and CIR at long ranges. Both of these measurements are essential to UltraPoser design – Doppler for enhancing motion features and CIR for IMU drift correction.

Based on the above insights described in Sec. 5.1, we adopt a hybrid transmission strategy that combines the benefits of both signal types – the single frequency tone and a wideband symbol. We choose the length of the symbol to compensate for the loss of range due to the interspersing of the symbols across devices. We first make the watch and the phone transmit a single tone at 17 kHz and 17.5 kHz respectively, where both devices demonstrate the maximum transmitter-receiver gains at high frequencies. Next, both devices emit OFDM-modulated wideband signals generated from a Zadoff–Chu sequence with a 16 ms frame duration spanning 18–22 kHz, corresponding to a range resolution of 4.3 cm. An interleaved subcarrier allocation scheme is used: the smartphone occupies even-numbered subcarriers, while the smartwatch uses odd-numbered ones. This design enables simultaneous extraction of high-SNR Doppler and accurate CIR measurements for both devices, which meet the requirement for full-body pose estimation.

Ultrasound Feature Extraction: The ultrasound signal is processed through two parallel pipelines to extract Doppler and CIR features. For Doppler processing, we first apply a notch filter to remove the carrier frequency. Then we apply an STFT with an FFT window length of 8192 samples and an overlap of 7680 samples at a 48 kHz sampling rate to obtain the time-varying Doppler spectrum. The resulting Doppler amplitude spectrograms from three links (phone-to-phone, watch-to-watch, watch-to-phone) are used as input to the pose estimation model. Note that the inter-device Doppler channel captures relative motion between different body parts, providing complementary features beyond intra-device channels to enhance pose estimation accuracy.

For CIR processing, the received signal is cross-correlated with the transmitted template to estimate the CIR, where each correlation peak corresponds to the propagation delay and thus the distance of individual multipath components. To extract motion-related information, we apply first-order temporal differentiation to the CIR over time and use the amplitude of the resulting dynamic CIR from two links: phone-to-phone and watch-to-watch. Note that the phone-to-phone CIR has a much better SNR and a significant overlap in the sensed information with that of the watch-to-watch CIR which is extremely noisy due to the watch's limited speaker volume. Thus, our model primarily relies on the phone-to-phone CIR to expand the spatial scope of sensing pose behavior. We anticipate that our ultrasound profile can be readily adapted to multiple new acoustic link pairs as commodity wearable hardware improves.

6 Graph-based Multi-modal Fusion

Overview and Design Motivation. After extracting ultrasound features, the next step is to effectively fuse multi-modal features (IMU, Doppler, CIR) collected from different wearables (Phone, Watch, Earbud) into a unified pose estimation framework. This task presents two key challenges: (1) IMU captures local orientation and acceleration, whereas ultrasound provides complementary velocity and range information. *These heterogeneous signals reflect distinct physical quantities and have different temporal and spatial properties.* (2) *Each device is located on a different body part and has a biased spatial coverage.* For instance, the phone may be more sensitive to movement of the right-side joints when placed in the right hand/pocket.

To address these challenges, we design a graph-based, physics-aware pose estimation network, as illustrated in Fig. 5. *Our key insight is to leverage the inherent skeletal graph structure of the human body and integrate physics-aware priors based on the locations and sensing coverage of each device. This design explicitly encodes which modalities from which devices should contribute more significantly to reconstructing specific body parts, allowing the network to perform more targeted and accurate pose refinement.*

Model Input and Output. UltraPoser takes IMU data, Doppler spectrograms, and CIR profiles as input features, denoted as (X^I, X^D, X^C). Specifically, $X^I \in \mathbb{R}^{T\times C}$ represents IMU signals from the smartphone, smartwatch, and earbud, where $T = 30$ corresponds to a 1-second time window and $C = 36$ is the total number of IMU channels. Following prior work [29, 45], each device provides 3-axis acceleration and a 3×3 rotation matrix, contributing 12 IMU channels per device. $X^D \in \mathbb{R}^{T\times(F\times N_l)}$ denotes Doppler spectrograms extracted from three acoustic links—phone-to-phone, watch-to-watch, and watch-to-phone—where $F = 100$ is the number of Doppler frequency bins and $N_l = 3$ is the number of links. $X^C \in \mathbb{R}^{T\times D}$ represents CIR profiles from the phone-to-phone link, with $D = 200$ range bins capturing multipath range information.

The model outputs 144 SMPL [26] pose parameters, representing 24 human body joints using 6D rotations (i.e., 6D unconstrained vectors that are mapped to rotation matrices via Gram-Schmidt), which offer continuous representations that are more suitable for learning [57].

Cross-modal Feature Embedding: As shown in Fig. 5, we employ specialized encoders to effectively extract spatial and temporal features from each modality. For IMU data, we use a BiLSTM-based IMU encoder, which is well-suited for modeling local pattern variations within short-term temporal signals [14]. In contrast, ultrasound data (i.e., Doppler and CIR) exhibit high-dimensional time–frequency and time–distance structures. To capture their complex contextual dependencies, we adopt two Transformer-based encoders (i.e., Doppler encoder and CIR encoder), respectively [39]. All encoded features are compressed to a common embedding dimension for more effective multi-modal fusion. This results in feature embeddings $F^I, F^D, F^C \in \mathbb{R}^{T\times d}$ for IMU, Doppler and CIR inputs, respectively, where $d = 256$ is the feature dimension.

Traditional concatenation fusion solutions overlook intrinsic inter-correlations among modalities jointly describing the same underlying motion. In contrast, UltraPoser adapt the cross-attention mechanism to enable cross-modal feature fusion, allowing information from different modalities to interact and reinforce one another [12, 23]. Given the modality consistency and computational simplicity, we first fuse the Doppler and CIR features to obtain the ultrasound feature $F^U = F^D \odot F^C$, where $\odot$ denotes element-wise multiplication. Then, we develop two cross-attention modules: the IMU Cross-Attention module uses IMU features F^I as queries and the ultrasound feature F^U as keys and values to enhance F^I, while the Ultra Cross-Attention module reverses the roles, refining F^I

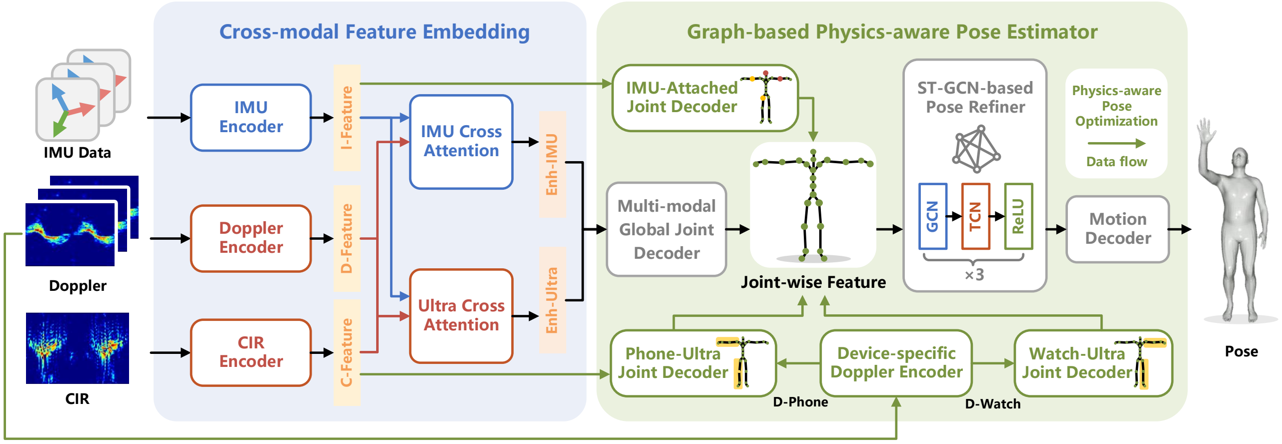

Cross-modal Feature Embedding
Graph-based Physics-aware Pose Estimator
IMU Data
Doppler
CIR
IMU Encoder
Doppler Encoder
CIR Encoder
I-Feature
D-Feature
C-Feature
IMU Cross Attention
Ultra Cross Attention
Enh-IMU
Enh-Ultra
IMU-Attached Joint Decoder
Multi-modal Global Joint Decoder
Joint-wise Feature
ST-GCN-based Pose Refiner
GCN
TCN
ReLU
×3
Physics-aware Pose Optimization
Data flow
Motion Decoder
Pose
Phone-Ultra Joint Decoder
Device-specific Doppler Encoder
Watch-Ultra Joint Decoder
D-Phone
D-Watch

Figure 5: Overview of our multimodal fusion framework, which comprises two key modules: (1) *Cross-modal Feature Embedding*, which extracts and fuses features from IMU and ultrasound data with cross-attention; (2) *Graph-based Physical-aware Pose Estimator*, which leverages the human skeletal graph structure to capture spatial complementarity across sensing modalities and body regions.

with F^U. The resulting enhanced features, denoted as F_e^U and F_e^I, are passed to the subsequent modules.

Graph-based Physics-aware Full-body Pose Estimator. Our key insight is that the human skeleton naturally forms a graph structure, with intrinsic connectivity between the various joints. To leverage this prior knowledge for better fusion, we represent the 24 body joints as nodes in a graph (see Fig. 5) and develop a Spatial-Temporal-GCN (ST-GCN) network [48] as the backbone of our full-body pose estimator. This approach enables structured and semantically meaningful propagation of motion features across the skeleton [49]. To align the enhanced features (i.e., F_e^U and F_e^I) from Sec. 5 with the skeleton graph-based representation, we first design a Multi-Layer Perceptron (MLP)-based [37] multimodal global joint decoder (Fig. 5). We begin by fusing the enhanced IMU and ultrasound features using element-wise dot product $F_e = F_e^U \odot F_e^I$, where $F_e \in \mathbb{R}^{T\times d}$ and d denotes the feature dimension. We then apply the multimodal global joint decoder to F_e into global joint embeddings $F_p \in \mathbb{R}^{T\times 24\times d}$, corresponding to the 24 skeletal joints.

Another important insight is that each wearable device and its corresponding sensing modalities have distinct advantages in sensing the nearby joint movements. IMU sensors offer high-precision readings for joints to which they are directly attached, while the ultrasound features are more sensitive to motion in joints located near the device. To exploit this, we introduce a physics-aware pose optimization scheme that leverages the physical relationships between device locations, body joints, and sensing capabilities. Specifically, we implement physics-aware modality-specific decoders tailored to each device, as shown in Fig. 5. The IMU-attached joint decoder maps IMU features to the joints nearest to the device location. Doppler spectra from the phone and watch are reused but processed independently by a device-specific Doppler encoder to extract motion-related features corresponding to nearby joints. Then, the phone ultrasound joint decoder maps the phone-specific ultrasound features to the joints on the same side of the body as the phone, while the watch ultrasound joint decoder focuses on joints on the watch-wearing side. The output of these physics-aware joint decoders is fused with the global joint embeddings F_p with element-wise multiplication, aiming to guide the network with structural priors about the spatial constraints between wearable devices and body joints.

Next, the fused joint embedding is fed into an ST-GCN-based pose refiner module, which captures both the spatial skeletal topology and temporal motion dynamics to further enhance the joint representations. This module consists of a three-layer architecture, with each layer integrating a GCN [53], a Temporal Convolutional Network (TCN) [13] and a ReLU activation function, as shown in Fig. 5. By progressively reducing the feature dimension and temporal kernel size across layers, the network hierarchically abstracts motion patterns from increasingly localized spatiotemporal contexts, resulting in more precise pose refinement. Additionally, we incorporate a learnable importance weighting mechanism on the graph edges, which allows the model to dynamically assign varying importance to different skeletal connections—emphasizing joints in proximity to device-equipped regions.

Finally, followed by an MLP-based motion decoder, the model outputs final full-body joint prediction results, i.e., 6D rotations of 24 body joint points $R \in \mathbb{R}^{T\times(24\times 6)}$.

Loss Function Design. The loss function of our model consists of two main components: a joint 6D rotation estimation loss $\mathcal{L}_{\text{rot}}$ for ensuring global coherence of the skeleton, and a joint position estimation loss $\mathcal{L}_{\text{pos}}$ for providing local structural constraints. Among them, the joint 6D rotation estimation loss is computed as follows:

$$\mathcal{L}_{\text{rot}} = \|R - R_{GT}\|_2^2 \tag{3}$$

where R_{GT} refers to the ground truth of joint 6D rotations. To obtain a full-body joint position estimate P, we employ a standard forward kinematics module [26] that computes the output joint rotations propagated through the human skeleton hierarchy to determine the global positions of joints and mesh vertices. The joint position

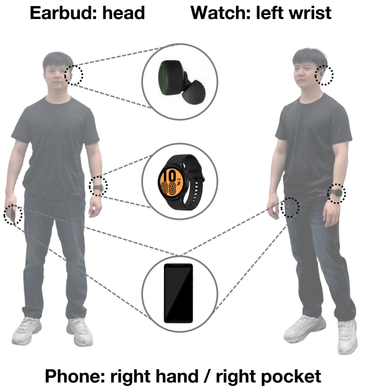


Figure 6: UltraPoser system setup. UltraPoser collects data from users wearing a smartphone, smartwatch, and earbud. The smartphone is either held in the hand or placed near the trouser pocket during data collection.

estimation loss is as follows:

$$\mathcal{L}_{\text{pos}} = \|P - P_{GT}\|_2^2 \tag{4}$$

where P_{GT} refers to the ground truth of joint positions. Therefore, the final loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{pos}} \tag{5}$$

7 Evaluation

This section describes our data collection setup and neural network training settings. We then evaluate the effectiveness of UltraPoser by comparing its performance with the SoTA IMU-only solution, i.e., IMUPoser [29], across various modalities, joints, and motion types. Finally, we conduct ablation studies to assess our designed model components and the impact of each device.

7.1 Dataset Collection

Data Collection Setup: Since UltraPoser is the first system that combines IMU and ultrasound sensing using commodity devices for full-body pose estimation, we collected our own dataset for both training and evaluation. As shown in Fig. 6, our data collection setup includes a smartwatch (Samsung Galaxy Watch 4) worn on the left wrist, a smartphone (Pixel 2XL), and a pair of wireless earbuds (ESense). While our evaluation is based on this specific combination of devices, the underlying framework is designed to be generalizable to other smartphones and wearable platforms. We collected data and evaluated UltraPoser in two real-world usage scenarios: (1) "Phone in Hand": the user holds the smartphone in their right hand, and (2) "Phone in Pocket": the user places the smartphone in their right trouser pocket. Each participant performs the same set of motion sequences once in each scenario. In the "Phone in Pocket" setup, the smartphone's position may shift during motion, which can alter the IMU coordinate frame and introduce acoustic noise due to friction. To mitigate this, we use an attachable pocket clip to secure the phone near the user's trouser pocket. The ESense earbuds are used on the left side only to collect IMU data, as this configuration provides more stable measurements for this particular device.

We set the IMU sampling rates to 50 Hz for both the smartphone and the earbud, and 100 Hz for the smartwatch. We utilize all available speakers (two on the smartphone and one on the smartwatch) and microphones (two on the smartphone and one on the smartwatch) for acoustic transmission and recording. Among the two recorded audio channels for the smartphone, we calculate the SNR and select the one with higher quality for ultrasound sensing. For the ground truth pose, we use a ZED 2i stereo camera to record RGB images at 30 Hz. These images are processed using 4DHumans [11], a SoTA human mesh recovery network, to generate full-body poses in the SMPL [26] format. We then synchronize all interpolated IMU and ultrasound data from the various devices with the ground truth poses to ensure temporal alignment across different sensing modalities. Following prior work, such as IMUPoser [29], we perform calibration on the devices' IMU data before data collection. We first align all devices to a common reference frame before recording data then ask participants to perform a template pose (i.e., T-pose) at the beginning of each recording session. This calibration step accounts for individual differences in body dimensions (e.g., height and limb lengths), ensuring accurate mapping of sensor data to full-body pose representations.

Dataset Details We recruited 10 volunteers for data collection, with weights ranging from 52 kg to 83 kg, ages from 19 to 27 years, and heights between 162 cm and 188 cm[2]. Participants were asked to wear the devices in a manner that felt most comfortable to them. To better reflect real-world usage, we did not control for clothing differences or smartwatch placement preferences. Following the motion classes defined in the DIP-IMU [14] dataset, we incorporated the following set of motion actions:

- Upper Body: Right arm raise, left arm raise, both arms raise, right arm push, left arm push, both arms push, right/left arm circle.
- Lower Body: Right leg kick, left leg kick, right knee raise, left knee raise, squats, lunges with the left/right leg.
- Full Body: Raising both arms while stepping side to side, spreading arms while performing forward lunges, and marching in place.
- Walking: Walking continuously back and forth.

Participants followed instructional videos during the motions but were not required to replicate the motions exactly. Each motion category – upper body, lower body, full body, and walking – lasted approximately 240 seconds, 180 seconds, 90 seconds, and 60 seconds, respectively. Each participant completed all motion sequences twice: once with "Phone in Pocket" and once with "Phone in Hand". On average, each participant contributed approximately 20 minutes of data, resulting in a total dataset of around 185 minutes and 334, 000 data frames.[3]

Neural Network Training. The pose estimation model is implemented in PyTorch and trained using an NVIDIA RTX 4080 Ti Super

[2] While the current dataset primarily includes male subjects, we have open-sourced it to support future research and encourage its expansion with more diverse participants for broader generalization.

[3] All data collection procedures were approved by the Institutional Review Board (IRB) of our institution.

Table 2: Performance comparison of different input modalities under two phone placement scenarios.

Input Modal	**Phone in Pocket**			**Phone in Hand**		
	MPJRE (°) ↓	MPJPE (cm) ↓	MPJVE (cm) ↓	MPJRE (°) ↓	MPJPE (cm) ↓	MPJVE (cm) ↓
IMU (IMUPoser [29])	10.01	4.77	5.85	8.97	4.33	4.81
IMU + Doppler	8.33	3.71	4.41	8.80	3.84	4.37
IMU + CIR	7.83	3.65	4.43	8.17	3.69	4.23
IMU + Doppler + CIR (UltraPoser)	7.12	3.21	3.84	7.62	3.33	3.80
↑ **Error Reduction (%)**	**28.9%**	**32.7%**	**34.4%**	**15.1%**	**23.1%**	**21.0%**

GPU. Each input sample consists of IMU, Doppler, and CIR data over a 30-frame window (1 second) with no overlap. We use the Adam optimizer with a learning rate of 3×10^{-4} and a batch size of 128. A dropout rate of 0.2 is applied to prevent overfitting. Early stopping is employed with a patience of 5 epochs based on the validation loss. The training, validation, and testing datasets are split in a ratio of 7:1:2, with samples from different phone placements combined. The total training time was approximately 8 minutes.

7.2 Evaluation Metrics

We use the IMU-only SoTA solution, i.e., IMUPoser [29], as our baseline for comparison. We do not compare with MobilePoser [45] as its physics-based post-processing optimizer is complementary and can be applied to our method without modifying our system. For evaluation, we randomly split the dataset into training, validation, and testing sets with a 7:1:2 ratio. When assessing generalization across users, we conduct five-fold cross-validation, where data from 8 users is used for training and the remaining 2 users for testing in each fold. Following prior work [14], we adopt the following metrics to assess the quality of the model's predictions:

- Mean Per Joint Rotation Error (MPJRE): the average angular error across all joints in degrees (°).
- Mean Per Joint Position Error (MPJPE): the mean Euclidean distance between the predicted and ground truth joint positions in centimeters (cm), with the root joint aligned.
- Mean Per Joint Vertex Error (MPJVE): the average Euclidean distance error across all mesh vertices of the estimated SMPL model in centimeters (cm), with the root joint aligned.

7.3 Micro Benchmark

This section evaluates the effectiveness of UltraPoser and compares it with an IMU-only solution [29] to demonstrate the superiority of our system. We analyze the contribution of each input modality, provide qualitative comparisons, and assess performance across different body regions, motion types, and users.

Performance across Input Modalities: We evaluate the effectiveness of UltraPoser by comparing four input modality configurations: IMU-only (i.e., the baseline IMUPoser [29]), IMU + Doppler, IMU + CIR, and the full multi-modal input of IMU + Doppler + CIR (i.e., UltraPoser). Table 2 summarizes the performance across two usage scenarios, i.e., "Phone in Pocket" and "Phone in Hand".

Augmenting the IMU with either Doppler or CIR significantly reduces errors across all evaluation metrics. The combination of all three modalities in UltraPoser *consistently yields the best performance.* This demonstrates the complementary benefits of ultrasound sensing, as discussed in Sec. 3. Compared to the IMU-only baseline, UltraPoser achieves 28.9%, 32.7%, and 34.4% reductions in MPJRE, MPJPE, and MPJVE, respectively, under the "Phone-in-Pocket" scenario. For the "Phone-in-Hand" scenario, it also achieves 15.1%, 23.1%, and 21.0% reductions. On average across both scenarios, UltraPoser reduces MPJVE by 28.46%. We also perform paired t-tests comparing IMUPoser and UltraPoser across MPJPE, MPJRE, and MPJVE, yielding p-values below 0.0001 and confirming statistically significant improvements. The benefits are more pronounced for absolute measurements of joint position and vertex errors due to the improved understanding of absolute locations with the additional information from the ultrasound sensing profile.

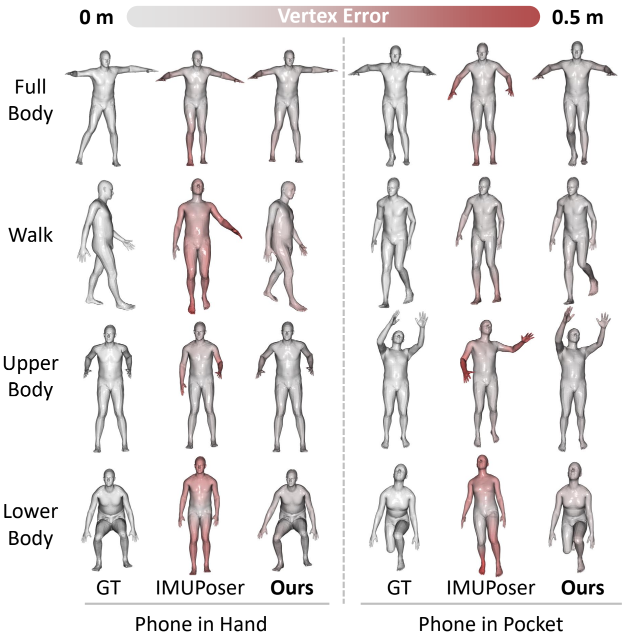


Figure 7: Qualitative comparison across different types of motions and device placements. A more intense red color indicates a higher vertex error.

Qualitative Comparison with Prior Work. Fig. 7 presents a qualitative comparison of mesh predictions generated by UltraPoser and IMUPoser across various postures and phone placements.

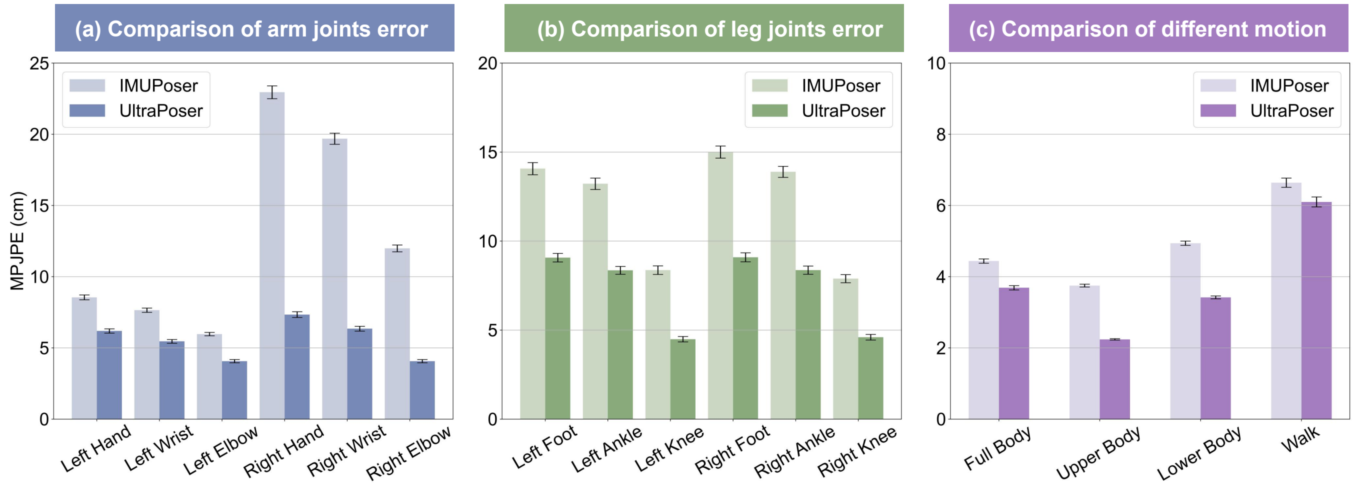


Figure 8: Comparison of MPJPE across different joints and motion types. (a) Arm joint errors during upper-body motion in the phone-in-pocket scenario. (b) Leg joint errors during lower-body motion in the phone-in-hand scenario. UltraPoser demonstrates significant improvement for body parts without directly attached sensors. (c) Comparison of pose estimation accuracy across various motion types. Error bars represent the 99% confidence interval.

As expected, IMUPoser exhibits the most significant errors when the phone is held in hand during lower body motion and when placed in the pocket during upper body motion. In both cases, the IMU fails to capture the motion of actively moving body parts to which it is not attached. In contrast, ultrasound sensing provides complementary motion cues by capturing reflected acoustic signals, effectively compensating for missing IMU data and offering valuable drift-free range measurements. Hence, UltraPoser can accurately reconstruct different body postures under different device placements. *These results highlight that integrating ultrasound sensing significantly improves full-body pose estimation accuracy, particularly in scenarios where IMU-only methods struggle.*

Performance across Body Regions: We further quantify the *significant improvements of* UltraPoser *for body joints that are not directly associated with IMUs.* We evaluate MPJPE in two representative scenarios: (1) arm joints under the "Phone-in-Pocket" scenario during upper-body movement, and (2) leg joints under the "Phone-in-Hand" scenario during lower-body movement. As shown in Fig. 8 (a) and (b), UltraPoser consistently outperforms the IMU-only baseline across all evaluated joints. For arm joints errors shown in Fig. 8 (a), UltraPoser substantially reduces pose errors for the right arm—including the hand, wrist, and elbow, where no IMU is present, achieving a 67.28% error reduction compared to IMUPoser. For leg joints shown in Fig. 8 (b), even when both the phone and smartwatch are located on the upper body, UltraPoser can still provide consistent improvements, reducing errors across all lower-limb joints by an average of 39.92%. These results highlight that *ultrasound sensing significantly expands the spatial coverage of motion cues without requiring additional worn devices.*

Performance across Motion Types: We assess the generalizability of our approach across different types of body motion by comparing the performance of UltraPoser and IMUPoser on full-body, upper-body, lower-body, and walking-specific motions separately.

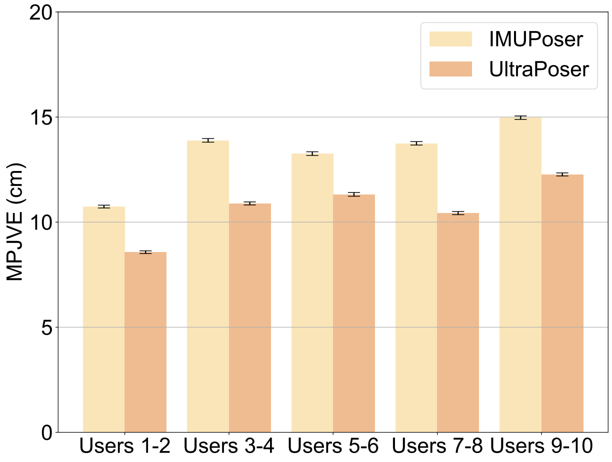


Figure 9: Comparison of MPJVE on different user groups with 5-fold cross-validation. Error bars represent the 99% confidence interval.

As shown in Fig. 8 (c), UltraPoser consistently outperforms the IMU-only baseline across all categories. The most notable gains are observed in upper-body motions. This is likely because upper-body movements are more variable and less constrained, making them harder to estimate without ultrasound sensing. Among all categories, walking exhibits the highest absolute error for both models. This is because walking is a highly dynamic activity that involves nearly all major body joints moving, making it particularly challenging to estimate accurately. Despite this, UltraPoser still maintains a performance advantage over IMUPoser, confirming its robustness for complex motions.

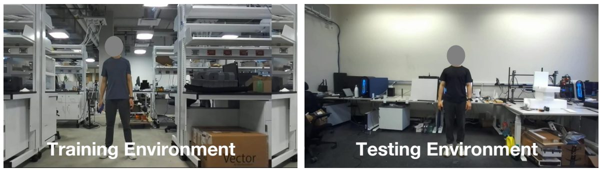


Figure 10: Different environments for training and testing.

Performance across Users: We conduct a five-fold cross-validation study to evaluate generalization across different users, as discussed in Sec. 7.2. This setting poses a greater challenge than within-user evaluation, as different users naturally exhibit variations in motion style, scale, speed, and joint coordination, even while performing the same actions. As shown in Fig. 9, although the absolute MPJVE increases for the cross-user setting, UltraPoser outperforms IMUPoser across all user groups. On average, UltraPoser achieves a 19.61% reduction in MPJVE compared to IMUPoser. Regarding MPJRE and MPJPE, IMUPoser achieves 26.78° MPJRE and 10.60 cm MPJPE, while UltraPoser achieves 21.55° MPJRE and 8.64 cm MPJPE, demonstrating consistent improvements across all metrics. Furthermore, cross-user paired t-tests conducted between IMUPoser and UltraPoser for MPJPE, MPJRE, and MPJVE yield p-values under 0.0001. These results highlight the potential of ultrasound sensing in generalizing across different body shapes and motion patterns. As part of future work, we plan to investigate user-adaptive modeling techniques and leverage simulated ultrasound data to better capture inter-user variability and further improve cross-user generalization.

Table 3: Comparison in unseen environments.

System	MPJRE (°)	MPJPE (cm)	MPJVE (cm)
IMUPoser	18.94	8.38	10.49
UltraPoser	16.41	7.19	8.69
Error Reduction (%)	13.4%	14.2%	17.2%

Environmental Sensitivity and Robustness: Different environments may have different furniture and layouts, which in turn affect ultrasound propagation and the resulting multipath patterns. To evaluate UltraPoser 's robustness and its ability to generalize to new environments, we train and validate our model using 8 users' data collected in two laboratory settings, then test it on 2 users' data from an unseen laboratory with a different size, layout, furnishings, and surroundings. Fig. 10 shows one of the training environments with 7 users' data and the testing environment.

As shown in Table 3, compared to IMUPoser, which is purely based on IMU and not affected by environmental variations, UltraPoser still consistently demonstrates improved performance across all metrics in unseen environments. As shown in Fig. 4 (a), the effective device-free sensing range of our system is approximately 1 meter. When the objects are beyond this range, UltraPoser will be largely unaffected. While extremely close objects may introduce variations in the measurements, such situations are relatively uncommon in applications like motion tracking for AR/VR interactions. Future research could also investigate methods to mitigate the impact of nearby moving objects, which may cause more significant distortions in the ultrasound spectrum. One promising direction is to leverage the co-located IMU sensors to distinguish user-induced motion from environmental interference, thereby enhancing system robustness.

Table 4: Ablation study on each network module. Both proposed modules contribute to system performance, with the graph-based physics-aware module providing the most significant improvement.

Module	MPJRE (°)	MPJPE (cm)	MPJVE (cm)	**MPJVE Rise**
wo/ Cross-Modal Attention	7.60	3.45	4.04	**5.76%**
wo/ Graph-based Optimization	9.34	4.18	4.89	**28.01%**
w/ both Modules	7.37	3.27	3.82	-

Table 5: Ablation study on each device. All three devices contribute to the overall system performance, with the phone having the greatest impact, followed by the watch and then the earbud.

Device	**MPJRE** (°)	**MPJPE** (cm)	**MPJVE** (cm)
wo/ earbud	7.86	3.47	4.07
wo/ watch	9.21	4.20	5.03
wo/ phone	14.43	5.71	6.90
w/ all devices	**7.37**	**3.27**	**3.82**

7.4 Ablation Study

This section presents an ablation study to evaluate the impact of each module design in our multimodal pose estimation network. We also analyze the contribution of each device to the overall system performance.

Contribution of each Module: We conduct an ablation study to evaluate the contributions of UltraPoser's key components: the cross-modal attention mechanism and the graph-based optimization module. First, we replace the cross-attention module with element-wise multiplication to evaluate the impact of explicit cross-modal interaction. Next, we remove the graph-based pose refinement and instead use an MLP-based pose estimator as used in prior work [29] to directly predict the final pose. Table 4 reports the model performance when each component is individually removed. Removing the cross-modal attention leads to moderate performance degradation across all metrics, increasing MPJRE and MPJPE to 7.60° and 3.45 cm, respectively, and causing a 5.76% increase in MPJVE. This indicates that leveraging the correlations between different sensing modalities helps refine temporal dynamics and spatial coherence.

In contrast, removing the graph-based physics-aware optimization results in more substantial performance drops, with MPJRE rising to 9.34°, MPJPE to 4.18 cm, and MPJVE to 4.89 cm, corresponding to a 28.01% increase in MPJVE. This shows the effectiveness of body-structure-aware reasoning in improving global consistency. The full model with both modules achieves the best performance across all metrics, confirming the complementary strengths of cross-modal fusion and graph-based physics-aware pose estimation.

Contribution of each Device: We assess the individual contribution of each device in our system and perform ablation experiments by selectively removing one device's data at a time—earbuds, watch, or phone—while retaining the data from the remaining two. When a device's data is removed, we adjust the network architecture accordingly to accommodate the modified input. As shown in Table 5, all three devices contribute to overall pose estimation performance. We observe that removing the phone leads to the most severe performance degradation across all metrics, likely due to its superior ultrasound sensing capability that provides richer motion and range information. The watch also plays an important role, and its removal causes a moderate performance decline. In contrast, removing the earbud leads to the smallest degradation, reflecting its more limited spatial coverage despite its contribution to head motion. Given that acoustic signals provide much of the benefits in UltraPoser and the earbuds are unable to measure or transmit these signals (Sec. 8), the least reduction in accuracy after removing them is not surprising. Besides, these results demonstrate that UltraPoser can still function effectively in reduced-device configurations, which is important for real-world applications where users may not consistently carry all three devices.

8 Limitations and Discussions

While UltraPoser demonstrates the feasibility of combining commodity IMUs and ultrasound sensing for full-body pose estimation, there remain several design trade-offs and practical limitations.

Hardware Limitations: Today's commodity hardware limitations rein UltraPoser's ability to make an even larger impact on pose estimation. First, most commercial earbuds reduce sampling for energy efficiency, making them unable to transmit and receive ultrasound signals. This prevents their integration into the acoustic sensing pipeline. However, recent advances in hearables [19] have the potential to change this limitation very soon. Second, the phone-to-watch acoustic link is excluded due to the limited receiving ability of the smartwatch's microphone, which limits UltraPoser's ability to leverage two-way Doppler for robustness.

Device Placement: While UltraPoser can achieve vertical spatial diversity by leveraging additional ultrasound features, our implementation requires users to wear the smartwatch and phone on opposite sides of the body laterally. However, in practice, users may wear devices differently—both on the same side, in a jacket pocket, or even in a backpack. Such variations can lead to overlapping sensing regions or degraded signal quality. Future iterations of UltraPoser could explore adaptive calibration strategies or device-aware modeling to support more flexible, user-specific placement scenarios. Additionally, to address the challenge of changes in device position or orientation during user movement, future work could explore auto-recalibration techniques that estimate the device's orientation using IMU data.

Our current implementation uses an attachable pocket clip to secure the phone near the user's trouser pocket to mitigate friction-induced acoustic noise. This noise arises from random, rapid vibrations between clothing and devices. While such friction primarily generates low-frequency components, it can also produce sharp, impulsive events with broadband spectral content that extends into the ultrasound band, causing interference with signal measurements. To address this issue, future work could explore solutions such as advanced filtering techniques or machine learning-based denoising to minimize contact-induced noise and ensure more robust system performance.

Operability in the Real World: UltraPoser operates in the 17–22 kHz frequency range. While the majority of this range (18-22 kHz) is generally inaudible to users, some individuals and animals—particularly when close to the speaker—may still perceive faint high-frequency sounds. This is primarily due to the lower end of the band (17–18 kHz) approaching the upper limit of human and animal hearing. This limitation stems from the restricted high-frequency performance of typical commodity speakers. As wearable audio hardware continues to improve, future systems may be able to operate at higher, fully inaudible frequencies with better SNR. Further, decreasing the bandwidth for less auditory overlap with human and animal auditory range only reduces the benefit provided by CIR. UltraPoser will still beat the state-of-the-art solutions on all metrics despite a reduced bandwidth.

Power Consumption: Another practical consideration is the increased energy consumption of the audio system, which may impact battery life in longer sessions. Our current implementation keeps the audio stream active throughout the sensing process to ensure robustness. However, future adaptive sensor fusion strategies can selectively activate it to balance accuracy and energy efficiency, such as activating the ultrasound module only when motion is detected or improved accuracy is necessary.

Privacy of the Sensed Pose: A good side of our proposed solution is that it can run on a phone, keeping the data secure and private on the phone. The size of the model is roughly 61 MB, which does not take significant storage resources to store on a phone and can allow modern devices to store all the detected poses locally. In fact, UltraPoser can generate pose embeddings that obfuscate all the other acoustic information and present only a human mesh for rehabilitation and high-endurance training applications maintaining the privacy of the user.

Application-relevant benefits of UltraPoser: (1) Rehabilitation requires an accurate understanding of the joint rotation and how much the affected limb or part of the body was raised. In some scenarios, this can enable effective intervention via doctors prescribing new exercises. Further, accurate pose estimation will make future AI-based rehabilitation systems even more robust. (2) Posture detection is a critical problem in early-middle-aged humans where a bad posture can lead to severe problems in the lower back. Using ubiquitous sensing solutions, like UltraPoser, has the potential to periodically warn users about these issues early and provide an improvement plan. (3) Adaptive gesture-based interfaces in applications such as AR/XR, smart fabrics, or even hardware-augmented gestures can leverage the pose as an additional input market to develop richer gestures for interaction using UltraPoser.

9 Conclusion and Future Work

This paper presents UltraPoser , the first system that integrates ultrasound sensing and IMUs for full-body pose estimation using off-the-shelf wearable devices. Our motivation study demonstrates that ultrasound sensing can enhance IMU-only solutions by capturing complementary motion features and providing multipath range profiles that compensate for IMU drift. To implement UltraPoser, we conduct systematical benchmark studies to design an optimal ultrasound profile that maximizes sensing quality. We further develop a graph-based, physics-aware multi-modal pose estimation framework that fuses the spatial complementarity of different wearable devices and sensing modalities. Extensive evaluations show that UltraPoser effectively overcomes the performance bottlenecks of existing IMU-based methods and achieves significant improvements, particularly for joint motions that IMUs alone fail to capture. While this work targets a specific device setup and application, it holds promise for future extensions to more general, distributed ultrasound or multimodal sensing systems built on commercial hardware.

We believe that UltraPoser's advancement of practical and low-cost pose tracking using commodity devices has the potential to enable applications in rehabilitation and sports training by reducing costs in daily life. Furthermore, other interactive applications, such as gesture-based smart device control, immersive interaction in XR environments, and motion capture for media development, can adopt UltraPoser as a foundational component to complement and enhance existing solutions.

Acknowledgments

This research was supported in part by the University of Michigan EECS Startup Fund, the National Natural Science Foundation of China under Grant No. 62432004, and a grant from the Guoqiang Institute at Tsinghua University. Xingyu Chen and Xinyu Zhang were supported by the Center for Wireless Communications at UC San Diego. Shuning Wang contributed to this project during her time as a visiting student at the University of Michigan, Ann Arbor. Yongjian Fu participated in the project while he was a visiting student at the University of California, San Diego.

References

[1] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. 2020. Socially and Contextually Aware Human Motion and Pose Forecasting. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6033–6040. https://doi.org/10.1109/LRA.2020.3010742

[2] 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

[3] 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

[4] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous Acoustic Sensing on Commodity IoT Devices: A Survey. *IEEE Communications Surveys Tutorials* 24, 1 (2022), 432–454. https://doi.org/10.1109/COMST.2022.3145856

[5] Xingyu Chen and Xinyu Zhang. 2024. RF Genesis: Zero-Shot Generalization of mmWave Sensing through Simulation-Based Data Synthesis and Generative Diffusion Models. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems* (Istanbul, Turkiye) *(SenSys '23)*. Association for Computing Machinery, New York, NY, USA, 28–42. https://doi.org/10.1145/3625687.3625798

[6] Kendra M. Cherry-Allen, Margaret A. French, Jan Stenum, Jing Xu, and Ryan T. Roemmich. 2023. Opportunities for Improving Motor Assessment and Rehabilitation After Stroke by Leveraging Video-Based Pose Estimation. *American Journal of Physical Medicine & Rehabilitation* 102, 2S (2023). https://journals.lww.com/ajpmr/fulltext/2023/02001/opportunities_for_improving_motor_assessment_and.13.aspx

[7] 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

[8] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. 2023. Avatars Grow Legs: Generating Smooth Human Motion From Sparse Tracking Inputs With Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 481–490.

[9] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2022. Svoice: Enabling voice communication in silence via acoustic sensing on commodity devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 622–636.

[10] Yang Gao, Wenbo Zhang, Junbin Ren, Ruihao Zheng, Yingcheng Jin, Di Wu, Lin Shu, Xiangmin Xu, and Zhanpeng Jin. 2024. PressInPose: Integrating Pressure and Inertial Sensors for Full-Body Pose Estimation in Activities. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 197 (Nov. 2024), 28 pages. https://doi.org/10.1145/3699773

[11] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *ICCV*.

[12] Sun Haoran, Wang Yang, Liu Haipeng, and Qian Biao. 2023. Fine-grained cross-modal fusion based refinement for text-to-image synthesis. *Chinese Journal of Electronics* 32, 6 (2023), 1329–1340.

[13] Pradeep Hewage, Ardhendu Behera, Marcello Trovati, Ella Pereira, Morteza Ghahremani, Francesco Palmieri, and Yonghuai Liu. 2020. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft Computing* 24 (2020), 16453–16482.

[14] 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

[15] Ahmad Jalal and S. Kamal. 2014. Real-time life logging via a depth silhouette-based human activity recognition system for smart home services. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 74–80. https://doi.org/10.1109/AVSS.2014.6918647

[16] 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 *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 443–460.

[17] 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

[18] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. 2022. Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 9 pages. https://doi.org/10.1145/3550469.3555428

[19] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[20] Jiye Lee and Hanbyul Joo. 2024. Mocap Everyone Everywhere: Lightweight Motion Capture With Smartwatches and a Head-Mounted Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1091–1100.

[21] Wenwei Li, Ruofeng Liu, Shuai Wang, Dongjiang Cao, and Wenchao Jiang. 2024. Egocentric Human Pose Estimation using Head-mounted mmWave Radar. In

Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems (Istanbul, Turkiye) *(SenSys '23)*. Association for Computing Machinery, New York, NY, USA, 431–444. https://doi.org/10.1145/3625687.3625799

[22] Ying Li, Chenxi Wang, Yu Cao, Benyuan Liu, Joanna Tan, and Yan Luo. 2020. Human pose estimation based in-home lower body rehabilitation system. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207296

[23] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. 2022. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17182–17191.

[24] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. End-to-End Human Pose and Mesh Reconstruction with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1954–1963.

[25] Kang Ling, Haipeng Dai, Yuntang Liu, Alex X. Liu, Wei Wang, and Qing Gu. 2022. UltraGesture: Fine-Grained Gesture Sensing and Recognition. *IEEE Transactions on Mobile Computing* 21, 7 (2022), 2620–2636. https://doi.org/10.1109/TMC.2020.3037241

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[27] 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 (Sept. 2023), 28 pages. https://doi.org/10.1145/3610895

[28] Jess McIntosh, Asier Marzo, Mike Fraser, and Carol Phillips. 2017. EchoFlex: Hand Gesture Recognition using Ultrasound Imaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1923–1934. https://doi.org/10.1145/3025453.3025807

[29] 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

[30] Movella. 2025. Xsens Motion Capture Systems. https://www.movella.com/products/xsens. Accessed: 2025-03-23.

[31] Noitom Ltd. 2025. Noitom Motion Capture Solutions. https://www.noitom.com/. Accessed: 2025-03-23.

[32] OptiTrack. 2025. OptiTrack Motion Capture Systems. https://www.optitrack.com/. Accessed: 2025-03-26.

[33] 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 (July 2022), 25 pages. https://doi.org/10.1145/3534605

[34] 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 (CVPR)*. 13323–13332.

[35] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th annual international conference on mobile computing and networking*. 160–173.

[36] Jiangnan Tang, Jingya Wang, Kaiyang Ji, Lan Xu, Jingyi Yu, and Ye Shi. 2024. A Unified Diffusion Framework for Scene-aware Human Motion Estimation from Sparse Signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21251–21262.

[37] Hind Taud and Jean-Franccois Mas. 2017. Multilayer perceptron (MLP). In *Geomatic approaches for modeling land change scenarios*. Springer, 451–455.

[38] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2023. Recovering 3D Human Mesh From Monocular Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 12 (2023), 15406–15425. https://doi.org/10.1109/TPAMI.2023.3298850

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[40] Lei Wang, Haoran Wan, Ting Zhao, Ke Sun, Shuyu Shi, Haipeng Dai, Guihai Chen, Haodong Liu, and Wei Wang. 2024. SCALAR: Self-Calibrated Acoustic Ranging for Distributed Mobile Devices. *IEEE Transactions on Mobile Computing* 23, 2 (2024), 1701–1716. https://doi.org/10.1109/TMC.2023.3241304

[41] Shuning Wang, Linghui Zhong, Yongjian Fu, Lili Chen, Ju Ren, and Yaoxue Zhang. 2024. UFace: Your Smartphone Can" Hear" Your Facial Expression! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–27.

[42] Wei Wang, Alex X. Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking* (New York City, New York) *(MobiCom '16)*. Association for Computing Machinery, New York, NY, USA, 82–94. https://doi.org/10.1145/2973750.2973764

[43] Alexander Winkler, Jungdam Won, and Yuting Ye. 2022. QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/3550469.3555411

[44] Xuan Xiao, Jianjian Wang, Pingfa Feng, Ao Gong, Xiangyu Zhang, and Jianfu Zhang. 2024. Fast Human Motion reconstruction from sparse inertial measurement units considering the human shape. *Nature Communications* 15, 1 (18 Mar 2024), 2423. https://doi.org/10.1038/s41467-024-46662-5

[45] 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

[46] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. 2023. Towards Generalized mmWave-based Human Pose Estimation through Signal Augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking* (Madrid, Spain) *(ACM MobiCom '23)*. Association for Computing Machinery, New York, NY, USA, Article 88, 15 pages. https://doi.org/10.1145/3570361.3613302

[47] Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. 2024. Person-in-WiFi 3D: End-to-End Multi-Person 3D Pose Estimation with Wi-Fi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 969–978.

[48] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[49] Honghong Yang, Hongxi Liu, Yumei Zhang, and Xiaojun Wu. 2024. FMR-GNet: Forward Mix-Hop Spatial-Temporal Residual Graph Network for 3D Pose Estimation. *Chinese Journal of Electronics* 33, 6 (2024), 1346–1359.

[50] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13167–13178.

[51] 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

[52] Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2024. Physical Non-inertial Poser (PNP): Modeling Non-inertial Effects in Sparse-inertial Human Motion Capture. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 50, 11 pages. https://doi.org/10.1145/3641519.3657436

[53] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6, 1 (2019), 1–23.

[54] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. 2024. Dynamic Inertial Poser (DynaIP): Part-Based Motion Dynamics Learning for Enhanced Human Pose Estimation with Sparse Inertial Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1889–1899.

[55] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[56] Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, and Xiaojie Jin. 2023. Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14678–14688.

[57] 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 (CVPR)*.