Sound-based Multi-Person 3D Pose Estimation

Yusuke Oumi[1], Yuto Shibata[1], Go Irie[1,2],
Akisato Kimura[3], Yoshimitsu Aoki[1], and Mariko Isogawa[1]

[1] Keio University, Yokohama, Kanagawa 223-8522, Japan
[2] Tokyo University of Science, Katsushika, Tokyo 125-8585, Japan
[3] NTT, Inc., Keihanna Science City, Kyoto 619-0237, Japan

Abstract. Can we recover the 3D poses of multiple people using only sound? This paper presents the first attempt to estimate multi-person 3D poses solely from acoustic signals. Estimating the poses of multiple individuals using acoustic signals is inherently challenging due to the superposition of motion-dependent signal variations. Unlike single-person scenarios, the presence of multiple subjects leads to overlapping acoustic signatures, making it difficult to attribute specific signal changes to an individual's pose. Furthermore, the complexity is compounded by inter-person reflections, which introduce intricate propagation delays that obscure the temporal motion-acoustic relationship. To address these issues, we propose SoundMHPE (Sound-based Multi-person Human Pose Estimator), a novel encoder-decoder framework consisting of two key components. First, the Acoustic Multi-scale Encoder captures diverse temporal and fine-grained frequency features to isolate subtle acoustic signatures from complex, overlapping signals. Second, the Temporal Pose Decoder employs an attention mechanism to disentangle multi-person information across successive frames. By jointly accounting for temporal dynamics and inter-person dependencies, this component precisely reconstructs frame-wise individual poses. To validate our approach, we constructed the 6-hour *Acoustic Multi-person Pose (AMP)* dataset consisting of 432K synchronized frames of multi-person pose and acoustic data, and demonstrated that our SoundMHPE outperforms baseline models. Project page: https://oumi03.github.io/sound-mhpe/

1 Introduction

Understanding human poses in complex, real-world environments is a long-standing fundamental problem in computer vision [22,28]. In everyday scenarios, such as crowded city streets, workplaces, or indoor facilities, multiple people naturally coexist and interact. Therefore, obtaining multi-person pose information in a non-invasive manner is beneficial for a wide range of applications, including monitoring systems [21], sports analysis [1], and disaster relief efforts [26].

Multi-person pose estimation has primarily been studied using RGB images [2, 4, 23], but these approaches are highly susceptible to occlusion and low-light conditions [11]. In contrast, wireless-signal-based methods [8, 9, 30, 32] can

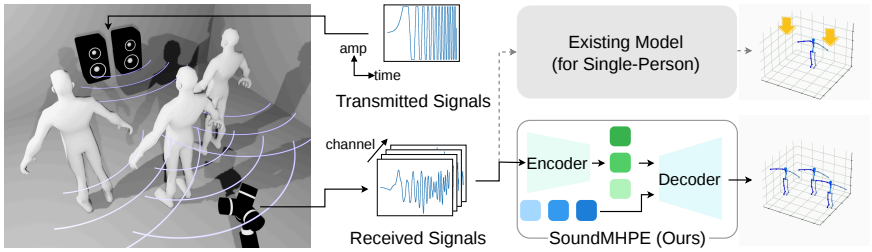

Fig. 1: We propose SoundMHPE, a sound-based multi-person 3D pose estimation method. Our system adopts an active acoustic sensing approach, where a speaker emits a transmitted signal and the received signal is used for the model input. While existing acoustic pose estimation models are limited to single-person estimation, SoundMHPE enables simultaneous estimation of multiple individuals.

be used in dark environments and can estimate the poses of occluded individuals by passing through obstacles. However, wireless signals are still susceptible to occlusion caused by water or metal [32]. To address these limitations, active acoustic sensing based pose estimation has recently gained attention [16–18, 24, 25]. Acoustic signals are unaffected by lighting conditions and, due to their relatively long wavelengths, offer the potential for estimating poses even behind obstacles such as metal. While these approaches can estimate the poses using only acoustic signals, they are limited to estimating poses for a single person.

To fill this gap, we take the first step toward estimating the poses of multiple individuals based solely on acoustic signals (Fig. 1). In contrast to single-person settings, where acoustic variations are uniquely determined by an individual's movements, the multi-person scenario introduces substantial signal ambiguity. Specifically, the observed signals represent a superposition of concurrent motion features, which is further complicated by non-trivial acoustic propagation delays caused by multi-body reflections. Together, these phenomena obscure the direct correspondence between the subjects' physical poses and the acoustic data.

To address these challenges, we propose SoundMHPE (Sound-based Multi-person Human Pose Estimator), a novel encoder-decoder framework that explicitly accounts for the complexity of acoustic signals caused by multiple subjects. To disentangle dynamic motion-dependent information from acoustically overlapped features and complex multi-body reflections, we propose an Acoustic Multi-scale Encoder. This approach employs a multi-resolution STFT with varying window sizes and effectively leverages the complementary strengths of different temporal and frequency scales. Short-window spectrograms capture high-fidelity temporal dynamics, while long-window counterparts resolve fine-grained frequency variations. This dual-representation ensures a comprehensive characterization of the signal across both the time and frequency domains. In addition, to explicitly model both inter-person interactions and intra-person temporal dynamics, we introduce the Temporal Pose Decoder. This module employs a spatio-temporal query mechanism where dedicated queries are assigned to each

Table 1: Comparison of existing human pose estimation methods and our method

Method	Modality	Restrictions	Estimation target
RGB-based [2, 4, 23]	RGB images/videos	Dark environment, occlusion	Single/Multiple persons
Wireless-signal-based [8, 9, 30]	RF/WiFi, mmWave, UWB	Occlusion, precision equipment	Single/Multiple persons
Acoustic-signal-based [18, 24]	Audio	Soundproof room	Only single person
Ours	Audio	Soundproof room	Single/Multiple persons

individual across successive frames. Such a formulation seamlessly integrates the temporal evolution of individual poses with the global context of multi-body interactions.

Furthermore, since multi-person pose estimation using acoustic signals has not been previously explored, we construct our 6-hour *AMP dataset* (Acoustic Multi-person Pose), consisting of synchronized multi-person poses and acoustic signals. Extensive experiments on the *AMP dataset* show that SoundMHPE achieves superior performance over baseline methods, and detailed ablation studies further confirm the effectiveness of our proposed components.

In summary, the contributions of our paper are as follows. (1) We tackle the first approach for multi-person pose estimation using acoustic signals. (2) We introduce an Acoustic Multi-scale Encoder to capture both fine-grained temporal dynamics and subtle acoustic variations in the acoustic signals. (3) We introduce the Temporal Pose Decoder to jointly model intra-person temporal dynamics and inter-person acoustic interactions. (4) Since this task has not been previously explored, we constructed our AMP dataset and conducted experiments to evaluate our proposed SoundMHPE model.

2 Related Work

2.1 Non-invasive Multi-person Pose Estimation

Multi-person pose estimation is a traditional task in computer vision, and many effective methods using RGB images [2,13,23] and videos [20,35], radio frequency (RF)/WiFi signals [29,30,32], mmWave [5,8,10], and UWB radar signals [7,9,34] have been proposed. However, both RGB images and wireless signals (RF/WiFi, mmWave, UWB) each have scenes in which estimation using them becomes less effective (see Table 1). RGB-based methods suffer from occlusion issues and decreased estimation accuracy in low-light conditions [11]. Additionally, RGB-based approaches are prone to raising privacy concerns [6]. Wireless-signal-based methods can estimate occluded targets by penetrating certain obstacles, but they face challenges in dealing with obstruction by water or metal [32]. Furthermore, the methods using wireless signals are limited in environments where wireless

communication is restricted, such as medical facilities or aircraft. We address these challenges by using acoustic signals.

2.2 Active Acoustic Sensing for Human Pose Estimation

Active acoustic sensing estimates target states by emitting sound signals and analyzing the received acoustic signals. For non-invasive human state estimation, this technique has been applied to tasks such as action recognition [27] and mesh reconstruction combined with RGB images [12]. More recently, active acoustic sensing has also been extended to human pose estimation. In particular, prior works have explored pose estimation with chirp signals [18, 24], inaudible continuous tones [16], and even music [17, 25]. However, all these methods are limited to single-person settings and cannot handle multiple persons (Table 1).

2.3 Spatio-Temporal Modeling for Multi-Person Pose Estimation

Spatio-temporal modeling is important for multi-person pose estimation, as temporal dynamics and inter-person dependencies must be jointly captured. For time-series data such as acoustic signals and videos, models that leverage temporal information are commonly employed.

In acoustic signal-based human pose estimation, Shibata *et al.* [24] proposed a CNN-based architecture with temporal convolutions, while Oumi *et al.* [18] enhanced temporal modeling by incorporating prior information about acoustic signals. However, these methods deal with single-person settings and primarily focus on temporal modeling.

In video-based multi-person pose estimation, a range of modeling strategies for spatio-temporal representations has been explored, including temporal-CNN [19], LSTM [15], and Transformer [33]. More recently, DETR-like architectures [3] have been extended to multi-person pose estimation [20, 35], where object queries enable explicit modeling of individual instances. Notably, Snipper [35] employs frame-wise and future-frame queries to jointly perform pose estimation and forecasting for each person. Inspired by these spatio-temporal query-based designs, we construct a sequence of queries for each individual, enabling the model to jointly learn the relationship between multi-frame acoustic features and the corresponding multi-frame poses across individuals.

3 Methodology

As shown in Fig. 2, SoundMHPE estimates the 3D pose sequence of multiple subjects $\mathbf{p_j} = \{p_{j,t}\}_{t=1}^{T}$ from an acoustic signal $\mathbf{s} = \{s_t\}_{t=1}^{T \times L}$. Here, j represents subject ID, T denotes the sequence length of $\mathbf{p_j}$, and L is the length of the acoustic signal sequence corresponding to single frame pose $p_{j,t}$. The following subsections discuss the key components of SoundMHPE. We introduce our two main technical contributions, Acoustic Multi-scale Encoder (Sec. 3.1) and Temporal Pose Decoder (Sec. 3.2). In Sec. 3.3, we present detailed implementation of the self-attention mechanism used in SoundMHPE.

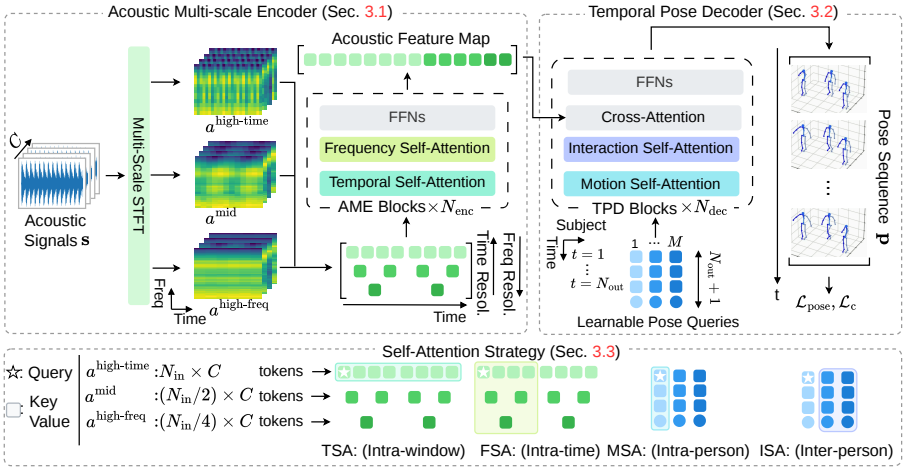

Fig. 2: Proposed framework for sound-based multi-person pose estimation. (Top left) SoundMHPE first employs an Acoustic Multi-scale Encoder to generate spectrograms with diverse time–frequency characteristics and obtain an acoustic feature map. (Top right) Subsequently, in the Temporal Pose Decoder, we assign each individual a set of learnable pose queries. (Bottom) These encoder and decoder modules leverage customized self-attention to jointly model spatio-temporal dynamics, multi-resolution features, and inter/intra-person pose relationships.

3.1 Acoustic Multi-scale Encoder

In our framework, the input acoustic signal $\mathbf{s}$ is captured using active acoustic sensing. We first apply multi-scale Short Time Fourier Transform (STFT) to obtain a multi-scale log-Mel spectrogram. This spectrogram is tokenized and fed into an Acoustic Multi-scale Encoder (AME), producing an acoustic feature map. SoundMHPE simultaneously estimates N_{out} consecutive pose frames. Following [18], to appropriately capture the temporal relationships in the acoustic signals, we use the acoustic signals corresponding to $N_{\text{out}} + N_{\text{prev}}$ pose frames to predict the latter N_{out} frames. Therefore, SoundMHPE estimates the pose sequence $\{p_{j,t}\}_{t=1}^{N_{\text{out}}}$ from the acoustic signal $\{s_t\}_{t=1-N_{\text{prev}}\times L}^{N_{\text{out}}\times L}$. We define $N_{\text{in}} = N_{\text{out}} + N_{\text{prev}}$.

Active Acoustic Sensing. The acoustic signal $\mathbf{s}$ is captured using a pair of speakers and a microphone. To capture the three-dimensional spatial information, we use an ambisonics microphone that records across four channels, i.e., W as an omnidirectional channel, and X, Y, Z, which capture directional components. Following existing acoustic-sensing-based human pose estimation methods [18,24], we use a time stretched pulse (TSP) signal, which is a periodic signal whose frequency changes within each cycle, as the emitted sound source.

Log-Mel Spectrogram. We convert the acoustic signal $\mathbf{s}$ into a log-Mel spectrogram. The received signal is segmented channel-wise with a fixed interval L, and each segment is transformed into a spectrogram via STFT. Then, the spec-

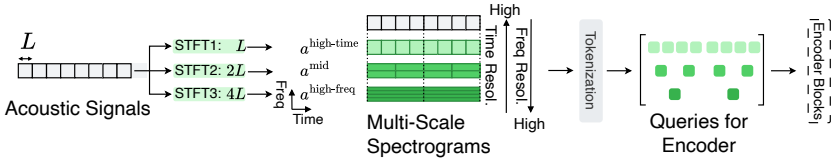

Fig. 3: By performing STFT with three different temporal window sizes, L, $2L$, and $4L$, we generate spectrograms denoted as $a^{\text{high-time}}$, a^{mid}, and $a^{\text{high-freq}}$, corresponding to high-temporal, intermediate, and high-frequency resolutions, respectively.

trogram is projected onto the Mel scale and converted to log scale as follows:

$$a_{t,c} = \log(\, H_{\text{mel}} \cdot \mathcal{F}(\{s_{t',c}\}_{t'=(t-1)\times L+1}^{t\times L})\,), \tag{1}$$

where H_{mel} denotes the Mel filter banks, $\mathcal{F}$ represents the Fourier transform operation, and c indicates the channel index. From an acoustic signal of length $T \times L$, we obtain log-Mel spectrogram $\mathbf{a} = \{a_t\}_{t=1}^{T}$, and each element $a_t \in \mathbb{R}^{(C,B)}$, where C denotes the number of microphone channels (four in this paper), and B denotes the number of Mel filter banks.

Multi-Scale STFT. Following [31], to leverage high resolution in both the temporal and frequency domains, we extract the log-Mel spectrogram using various window sizes: L, $2L$, and $4L$ paired with Mel filter banks B, $2B$, and $4B$, respectively. Given an acoustic signal of length $N_{\text{in}} \times L$, these configurations produce multi-scale spectrograms: $a^{\text{high-time}} \in \mathbb{R}^{(N_{\text{in}},C,B)}$, $a^{\text{mid}} \in \mathbb{R}^{(N_{\text{in}}/2,C,2B)}$, and $a^{\text{high-freq}} \in \mathbb{R}^{(N_{\text{in}}/4,C,4B)}$ (see Fig. 3). $a^{\text{high-time}}$ denotes a spectrogram with high temporal resolution, while $a^{\text{high-freq}}$ denotes a spectrogram with high frequency resolution. a^{mid} corresponds to an intermediate resolution between the two.

Implementation of the Encoder. The multi-scale spectrograms are concatenated along the temporal axis, after which linear layers are applied independently at each time step to unify their frequency dimensions. This yields a tensor with the size of $((7/4) \times N_{\text{in}}, C, E)$, where E denotes the embedding dimension of the model. The tensor is subsequently flattened along the time–channel dimensions to form the input queries of the encoder. Our AME consists of N_{enc} stacked AME blocks, each composed of a Temporal Self-Attention, a Frequency Self-Attention and feed-forward networks (FFNs), and outputs an acoustic feature map. The details of the two attention modules are described in Sec. 3.3.

3.2 Temporal Pose Decoder

This section introduces the Temporal Pose Decoder (TPD), which explicitly learns the temporal relationship between acoustic features and poses of each frame. We first present the preliminary concepts necessary to understand our method, followed by an explanation of the query design and loss functions.

Preliminary. DETR is a Transformer-based model designed for end-to-end object detection [3]. In their framework, the decoder prepares M queries, each corresponding to a potential object. Then, by computing cross-attention between

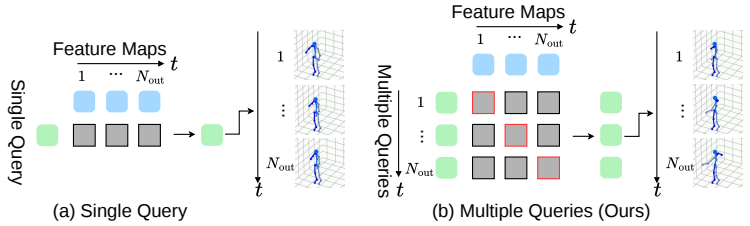

Fig. 4: Cross-attention between queries and acoustic feature maps in TPD. (a) Single Query for Multi-frame poses: The pose information from multiple frames is aggregated into a single query. (b) Multiple Queries for Multi-frame poses: Multiple queries enable the model to focus on the acoustic features of different frames for poses in each frame.

these queries and the feature maps produced by the encoder, the decoder extracts features corresponding to each object individually from the feature map. In DETR, the loss is computed by matching predicted objects with ground truth objects using the Hungarian matching algorithm. A similar approach is used in DETR-based multi-person pose estimation models [23, 30], where the decoder predicts a single-frame pose $\hat{p}_i$ and a confidence score $\hat{c}_i$ for each query. During inference, only the poses whose confidence $\hat{c}_i$ exceeds a predefined threshold are output as the final results.

Multiple-Queries for a Single Person. SoundMHPE is designed based on the aforementioned DETR-based architecture, which estimates the poses of multiple individuals by extracting information related to each person from the acoustic feature maps through cross-attention. Additionally, following existing sound-based pose estimation models [24], we estimate multi-frame poses simultaneously rather than a single-frame pose to ensure smooth transitions between adjacent frames. Therefore, in a standard DETR-style approach where M queries represent M potential individuals, each query must compress the entire temporal sequence of a person's pose into a single representation. As shown in Fig. 4 (a), this collapses the temporal dimension, making it difficult to capture fine-grained acoustic details such as inter-person reflections and their precise timing. To address this issue, we propose Temporal Pose Decoder (TPD), which uses N_{out} pose queries for each individual to estimate the multi-frame poses. By preparing N_{out} queries, each assigned to an estimated pose frame, the queries corresponding to earlier poses can attend to earlier acoustic features, while those corresponding to later poses can focus on the acoustic features of subsequent time steps (Fig. 4 (b)). To effectively disentangle individual motions across time, each query is conditioned on both a temporal positional embedding and a subject-specific embedding. In addition, since the confidence score $\hat{c}_i$ for each of the M instances is independent of a specific frame, we follow the approach of [13] and introduce M additional instance queries. The confidence score $\hat{c}_i$ is obtained by passing the decoder output associated with the instance query through a linear layer. Therefore, the number of learnable queries in TPD becomes $M \times (N_{\text{out}} + 1)$. Our TPD architecture is composed of N_{dec} stacked TPD blocks, each composed of

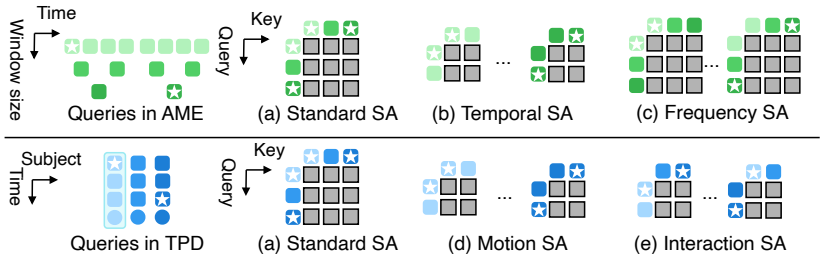

Fig. 5: Self-Attention in AME and TPD. (a) Standard Self-Attention: Compute attention across all queries. (b) Temporal Self-Attention: Models temporal dependencies within spectrograms generated using the same STFT window size. (c) Frequency Self-Attention: Captures relationships among spectrograms derived from the same acoustic sequence with different window sizes. (d) Motion Self-Attention: Learns temporal relationships of the same individual across frames. (e) Interaction Self-Attention: Models relationships between different individuals to capture inter-person interactions.

Motion Self-Attention (MSA), Interaction Self-Attention (ISA), Cross-Attention and FFNs. The details of these self-attention modules are described in Sec. 3.3.

Loss Functions. We utilize the Hungarian algorithm for matching-based loss calculation like [3]. The total loss $\mathcal{L}$ consists of two loss functions: $\mathcal{L}_{\text{pose}}$ and $\mathcal{L}_{\text{c}}$.

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + \lambda \mathcal{L}_{\text{c}}, \tag{2}$$

where $\mathcal{L}_{\text{pose}}$ is the mean squared error loss between the ground truth pose and the predicted pose matched to the GT pose. $\mathcal{L}_{\text{c}}$ is the binary cross entropy loss and is used to learn the confidence score. λ is a weight hyperparameter.

3.3 Self-Attention Strategy

The self-attention mechanisms in SoundMHPE are specifically designed to disentangle the complex spatio-temporal features inherent in multi-scale spectrograms. As demonstrated in our ablation studies (Sec. 5.2), decoupling temporal and frequency dependencies is essential for maintaining estimation precision in multi-person environments. In the following, we describe the detailed implementations of the self-attention used in our Acoustic Multi-scale Encoder (AME) and Temporal Pose Decoder (TPD), respectively.

Self-Attention in AME. Our AME performs Multi-Scale STFT using three temporal window sizes and we extract $a^{\text{high-time}} \in \mathbb{R}^{(N_{\text{in}},C,B)}$, $a^{\text{mid}} \in \mathbb{R}^{(N_{\text{in}}/2,C,2B)}$, and $a^{\text{high-freq}} \in \mathbb{R}^{(N_{\text{in}}/4,C,4B)}$. The queries generated from these spectrograms contain information from different time and frequency resolutions. As shown in Fig. 5 (a), standard self-attention merges queries from all time steps and resolutions into a single operation. This simultaneous processing hinders the model's ability to decouple temporal relationships from multi-scale frequency information, leading to less effective feature extraction. To explicitly model these

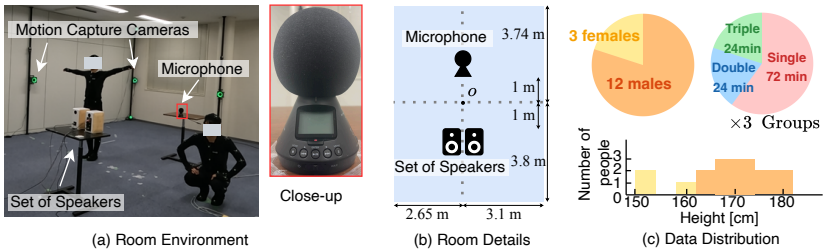

Fig. 6: Experimental setup. (a,b) Our measurement environment consists of a set of speakers and a microphone for active acoustic sensing, along with motion capture cameras to obtain ground-truth poses. (c) Our AMP dataset consists of 12 male and 3 female participants, with heights ranging from 150 cm to 181 cm. The participants were divided into three groups. For each group, we collected 72 minutes of single-person data, 24 minutes of double-person data, and 24 minutes of triple-person data.

characteristics, we introduce two types of self-attention mechanisms: Temporal Self-Attention (TSA) and Frequency Self-Attention (FSA). TSA performs self-attention within the features generated from each specific STFT window resolution (Fig. 5 (b)), focusing on the temporal relationships within each spectrogram. Specifically, self-attention is calculated over $N_{\text{in}} \times C$, $(N_{\text{in}}/2) \times C$, and $(N_{\text{in}}/4) \times C$ queries corresponding to $a^{\text{high-time}}$, a^{mid}, and $a^{\text{high-freq}}$, respectively. FSA computes attention across spectrograms generated with different temporal window sizes, but originating from the same acoustic signal sequence (Fig. 5 (c)). Specifically, since temporal windows of L, $2L$, and $4L$ are employed, four queries from $a^{\text{high-time}}$, two queries from a^{mid}, and one query from $a^{\text{high-freq}}$ correspond to the same acoustic sequence of length $4L$. Therefore, self-attention is computed over these $7 \times C$ queries.

Self-Attention in TPD. To efficiently extract pose features in multi-person scenarios, our TPD decouples the self-attention process into two distinct axes: (i) intra-person temporal dynamics and (ii) inter-person interactions. While a standard DETR-based decoder employs a single query per person to model inter-instance relationships, our TPD assigns $N_{\text{out}} + 1$ queries to each individual to capture fine-grained temporal changes. Applying global self-attention across all $M \times (N_{\text{out}} + 1)$ queries would indiscriminately mix temporal cues with inter-person reflective dependencies, hindering the model's ability to focus on either. To resolve this, we introduce the Motion Self-Attention (MSA) and Interaction Self-Attention (ISA). The MSA restricts its attention to the $N_{\text{out}} + 1$ queries belonging to the same individual. By focusing solely on an individual's own queries, the model effectively extracts consistent intra-person temporal dynamics (Fig. 5 (d)). In contrast, the ISA computes attention between the i-th subject and the remaining $(M-1) \times (N_{\text{out}} + 1)$ queries of other subjects. This allows the model to explicitly capture inter-person dependencies and mutual acoustic influences (Fig. 5 (e)).

4 Experimental Settings

4.1 Multi-Person Acoustic Pose (AMP) Dataset

Since we are tackling a novel task for which no existing dataset is available, we constructed the 6-hour Acoustic Multi-person Pose (AMP) dataset. This dataset consists of synchronized acoustic data acquired through active acoustic sensing and 3D coordinate data of multiple individuals.

The measurement environment is shown in Fig. 6 (a,b). We employed a pair of loudspeakers (Edifier ED-S880DB) and an ambisonics microphone (Zoom H3-VR) for acoustic data collection in an indoor room where background noise and reverberation were present. A Motive motion capture system (OptiTrack) equipped with 16 cameras was utilized for obtaining ground-truth pose data.

The AMP dataset statistics are shown in Fig. 6 (c). It consists of 15 subjects (12 male, 3 female) with heights ranging from 150 to 181 cm. To prepare a multi-person dataset, these subjects were divided into three groups. Within each group, subject pairings and positions were randomized during data collection. Participants performed a variety of poses, including walking, twisting, and raising both hands, in a random order and at random speeds. We used a skeleton consisting of 21 joints. Each joint is represented by the head, neck, shoulders, arms, forearms, hands, waist, thighs, shins, feets, toes, hip, and spine. For each group, we collected 72 minutes of single-person data, and 24 minutes each of double-person and triple-person data. Ground-truth poses were recorded at 20 fps, resulting in a total of approximately 432K frames.

4.2 Baseline Methods

Since no prior work exists for acoustic multi-person pose estimation, we adapted two closely related models as baselines: (1) Adapted Shibata *et al.* [24]: Although originally for single-person estimation, this work achieved precise active pose estimation with the TSP signal. To enable a fair comparison, we extended its feature extractor with a multi-person regression head. (2) Repurposed Yan *et al.* [30]: We adapted this multi-person WiFi-based model because WiFi CSI and acoustic log-Mel spectrograms are functionally analogous; both capture environment-induced signal perturbations in the time–frequency domain. We modified the input stem to accommodate an acoustic spectrogram while preserving the original transformer architecture. Both baselines were retrained from scratch on our dataset to ensure fair comparison.

4.3 Evaluation Metrics

We employ three evaluation metrics: mean per joint position error (MPJPE), Procrustes-aligned mean per joint position error (PA-MPJPE), and percentage of correct keypoints (PCK). MPJPE is computed as the mean Euclidean distance between the predicted and ground-truth joint positions. PA-MPJPE first aligns the predicted pose to the ground-truth pose using Procrustes analysis, which

Table 2: Comparison against baselines

Method	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)
Shibata *et al.* [24]	121.7	71.5	0.36
Yan *et al.* [30]	119.9	69.7	0.36
Ours	**106.5**	**65.0**	**0.43**

Table 3: Ablation study

Method	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)
Ours w/o AME	115.2	67.5	0.38
Ours w/o TPD	116.5	69.0	0.38
Ours	**106.5**	**65.0**	**0.43**

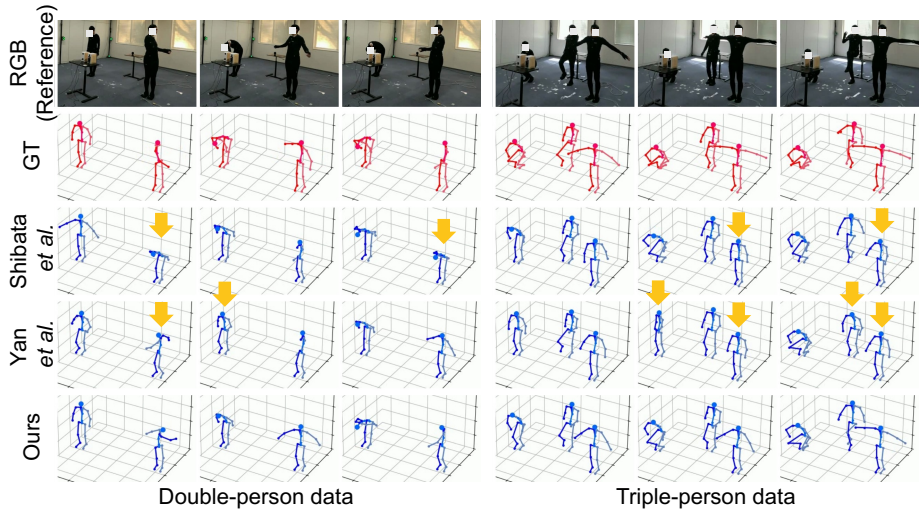

Fig. 7: Qualitative results. Frames were sampled at one-second intervals for both the double-person data and the triple-person data. The yellow arrow indicates a skeleton for which pose estimation failed.

removes differences in global translation, rotation, and scale. MPJPE is then computed on the aligned poses. PCK measures the percentage of joints whose euclidean distance to the ground truth is within a predefined threshold. We used PCKh@0.5, where the threshold is half the distance between the head and neck.

4.4 Implementation Details

The values of N_{out} and N_{prev} were set to 8 and 16, respectively. The number of Mel filter banks B for the log-Mel spectrograms was set to 128 and embedding dimension E was 256. We set the number of AME blocks and TPD blocks to $N_{\text{enc}} = 3$ and $N_{\text{dec}} = 2$, respectively. The number of queries in the TPD is set to 135 ($M = 15, N_{\text{out}} = 8$). For loss calculation, the weight was set to $\lambda = 0.5$. All experiments used AdamW [14] as the optimizer, with a weight decay of 1×10^{-4}. The number of training epochs was 500, and the learning rate was 5×10^{-5}.

5 Experimental Results

We conduct seven experiments to validate the effectiveness of SoundMHPE: (1) a comparison against baseline methods; (2) an ablation study to evaluate the effectiveness of Acoustic Multi-scale Encoder and Temporal Pose Decoder; (3) analysis of the performance gap between single-person and multi-person settings; (4) evaluation of the proposed attention methods in AME and TPD; (5) investigation of the impact of temporal window selection in multi-scale STFT; (6) generalization to unseen environments; and (7) a cross-modal evaluation demonstrating that SoundMHPE can be successfully applied to WiFi signals with clear performance gains.

5.1 Comparison with Baseline Methods

To demonstrate the effectiveness of SoundMHPE, we conduct comparisons with baseline models. We trained these models using data from two of the three groups defined in Sec. 4.1 (consisting of ten subjects), and conducted cross-subject (group) evaluation with the remaining unseen group (consisting of five subjects). The test group was rotated in a three-fold cross-validation setup, and the final evaluation results were obtained by averaging. Table 2 presents a quantitative comparison between SoundMHPE and the baseline models. SoundMHPE outperforms all baselines across all evaluation metrics.

Fig. 7 represents a qualitative comparison with baseline methods. In the double-person scenario, the baseline models struggle to track dynamic "twisting" motions, whereas SoundMHPE reconstructs them with high fidelity. Similarly, for the triple-person data, SoundMHPE successfully estimates "raising both arms", a motion characterized by a small sound reflection area, even when performed alongside a walking individual. We hypothesize that for motions involving subtle acoustic perturbations, such as twisting or arm raising, the integration of fine-grained frequency features via the AME and the explicit modeling of inter- and intra-person dependencies in the TPD are particularly effective.

5.2 Ablation Study

To demonstrate the effectiveness of our main technical contributions, we compared two ablation settings: (1) excluding Acoustic Multi-scale Encoder (w/o AME), and (2) excluding Temporal Pose Decoder (w/o TPD). In setting (1), the spectrogram $a^{\text{high-time}}$ extracted using a single temporal window is used as the model input. In setting (2), a total of M queries, each assigned to each individual, are fed into the decoder. Each query predicts both the multi-frame pose sequence and its associated confidence scores. In both settings, self-attention is implemented using standard self-attention.

Tab. 3 presents a quantitative comparison by evaluating the model's performance when specific components are removed. The results demonstrate that our complete method, which incorporates all proposed components, achieves the highest estimation accuracy across all evaluation metrics. In addition, the results

Table 4: Comparison on single and triple person

Method	Single-person			Triple-person		
	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)
Shibata *et al.* [24]	111.3	65.8	0.39	124.5	73.3	0.39
Yan *et al.* [30]	108.7	64.1	0.40	122.4	71.2	0.38
Ours	**95.0**	**58.9**	**0.47**	**111.2**	**68.1**	**0.44**

Table 5: Comparison on self-attention methods

Encoder	Decoder	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)
Standard	MSA+ISA	111.7	65.1	0.40
TSA+FSA	Standard	114.4	67.9	0.39
TSA+FSA	MSA+ISA	**106.5**	**65.0**	**0.43**

indicate that TPD contributes most significantly to the performance improvement, highlighting the importance of assigning a separate query to each predicted pose frame and extracting temporal information through cross-attention.

5.3 Comparison by the Number of Subjects

To analyze the performance gap between single-person and multi-person pose estimation, we compare the easiest setting (single-person) and the most challenging setting (three-person). Table 4 shows that SoundMHPE maintains competitive accuracy even in the multi-person setting, outperforming the baseline without significant degradation in performance.

5.4 Effect of self-attention

To investigate the effectiveness of the proposed Temporal Self-Attention (TSA) and Frequency Self-Attention (FSA) in the encoder, as well as Motion Self-Attention (MSA) and Interaction Self-Attention (ISA) in the decoder, we conduct ablation studies by replacing either the encoder or the decoder attention modules with standard self-attention computed over all queries. Table 5 presents the experimental results, showing that in both the encoder and decoder settings, the models incorporating our proposed attention mechanisms consistently achieve higher performance than those using standard self-attention.

5.5 Effect of Temporal Window Selection

By employing window sizes of $L, 2L$, and $4L$, our AME extracts multi-resolution acoustic features. This configuration captures high-fidelity temporal dynamics aligned with the target pose frame rate (L) while simultaneously leveraging extended windows to extract finer frequency information. To validate the rationale behind our window size selection, we compare our proposed $(L, 2L, 4L)$ configuration against three alternative variants: **Shifted Resolution Range** $(L/2, L, 2L)$: This setting balances the scales around the pose-aligned resolution L by including both narrower and wider windows. **High Temporal Resolution Focus** $(L/4, L/2, L)$: This configuration prioritizes fine-grained temporal cues at the expense of spectral depth, using only windows equal to or narrower

Table 6: Comparison on STFT window selection

Window	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)
$(L/4), (L/2), L$	118.5	69.6	0.37
$(L/2), L, 2L$	117.0	68.9	0.37
$L, 2L$	114.5	67.9	0.38
$L, 2L, 4L$	**106.5**	**65.0**	**0.43**

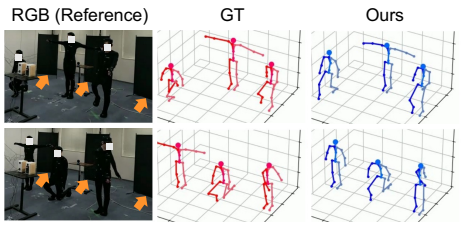

Fig. 8: Experiments in an unseen acoustically reflective environment created by black partitions (orange arrows).

than L. **Reduced Multi-scale Variety** $(L, 2L)$: This variant evaluates the impact of limited scale diversity by employing only two window sizes. As shown in Table 6, our $(L, 2L, 4L)$ configuration achieves the highest accuracy, confirming that this specific combination of temporal and frequency resolutions is most effective. Furthermore, we observe that estimation accuracy decreases when using finer temporal windows at the expense of frequency resolution. This suggests that high frequency resolution is critical to resolving the subtle acoustic variations induced by the fine-grained movements of multiple individuals.

5.6 Generalization to unseen environments

Generalization to environments with unseen reflection characteristics remains a challenge for sound-based pose estimation [18, 24]. To evaluate the performance of SoundMHPE in unseen environments, we placed multiple partitions in the room to create different reflection properties. As shown in Fig. 8, SoundMHPE can still estimate coarse poses even under these different reflection conditions, indicating adaptability to unseen environments.

5.7 Cross-Modal Applicability

To evaluate the generality of SoundMHPE beyond acoustic sensing, we apply its modality-agnostic components to Person-in-WiFi 3D (PiW) [30], a WiFi-based multi-person 3D pose estimation benchmark. We compare the original single-frame PiW model, denoted as PiW (Single), with PiW (Multi), a diagnostic variant extended to input and predict a 0.27-s sequence (T=4), and our method, which also uses T=4. Because WiFi channel state information (CSI) is not a waveform, we omit the waveform-specific Multi-Scale STFT and apply AME self-attention and the full TPD directly to the CSI features. All methods are trained and evaluated on the PiW dataset under the same protocol. As shown in Table 7, our method outperforms both PiW (Single) and PiW (Multi). The performance degradation of PiW (Multi) indicates that simply extending

Table 7: Quantitative comparison on the Person-in-WiFi 3D (PiW) dataset.

Method	MPJPE [mm] (↓)	PA-MPJPE [mm] (↓)	PCKh @0.5 (↑)
PiW (Single)	127.4	71.5	0.13
PiW (Multi)	161.0	82.7	0.06
Ours	**122.6**	**69.1**	**0.31**

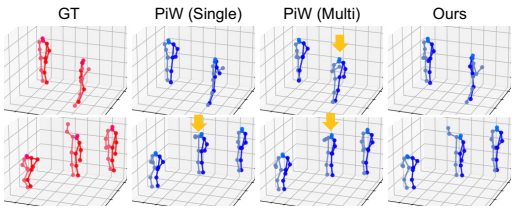

Fig. 9: Visualization of predictions on PiW dataset. Yellow arrows indicate failure cases.

a single-frame model to sequence prediction is insufficient for modeling frame-specific signal-to-pose correspondences and temporal dependencies. In contrast, the proposed self-attention mechanism and TPD-based temporal modeling remain effective in the WiFi domain, demonstrating the cross-modal applicability of our architecture. The qualitative results in Fig. 9 further show that our method estimates dynamic poses, such as walking and hand waving, more accurately than both PiW (Single) and PiW (Multi).

6 Conclusion

In this paper, we presented the first approach for multi-person pose estimation using active acoustic sensing with a pair of speakers and a microphone. To address the complexity of acoustic signals arising from multiple people, we introduced the Acoustic Multi-scale Encoder, which leverages spectrograms with diverse temporal and frequency characteristics. Furthermore, we employ the Temporal Pose Decoder to capture the relationship between a pose in each frame and the corresponding acoustic signals. These contributions improved the robustness of SoundMHPE, allowing it to outperform baselines across all evaluation metrics.

Because sound-based human pose estimation is still at a very early stage of research, one of the general challenges in this task is its real-world deployment. We consider that one of the reasons for this is that large-scale, comprehensive datasets have not yet been constructed. We believe that our work in constructing an original dataset and developing methods for multi-person pose estimation will contribute to pioneering this field. For future research, we plan to further advance sound-based pose estimation by evaluating our SoundMHPE across multiple experimental environments and improving the robustness and generalization capability of the model.

Acknowledgments.

This work was partially supported by JSPS Grant-in-Aid for Challenging Research (Exploratory) 24K22296, JST FOREST Program JPMJFR242I, and JST BOOST JPMJBS2409.

References

1. Badiola-Bengoa, A., Mendez-Zorrilla, A.: A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. Sensors **21**(18), 5996 (2021)
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE TPAMI **43**(1), 172–186 (2021)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
4. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV. pp. 2334–2343 (2017)
5. Feng, Y., Zhang, Q., Wang, Z., Hui, X., Zhou, Y.: Multi-person pose estimation using velocity-dependent enhanced mmwave radar point clouds. IEEE IoTJ (2025)
6. Hinojosa, C., Niebles, J.C., Arguello, H.: Learning privacy-preserving optics for human pose estimation. In: ICCV. pp. 2573–2582 (2021)
7. Huang, G., Hu, J., Liu, H., Lin, J., Xie, Z.: Joint loss-optimized end-to-end network for through-wall multi-person pose estimation with portable uwb radar. IEEE Sensors J. (2025)
8. Kato, S., Yataka, R., Wang, P.P., Miraldo, P., Fujihashi, T., Boufounos, P.: Raptr: Radar-based 3d pose estimation using transformer. In: NeurIPS. vol. 38 (2025)
9. Kim, G.W., Lee, S.W., Son, H.Y., Choi, K.W.: A study on 3d human pose estimation using through-wall ir-uwb radar and transformer. IEEE Access **11**, 15082–15095 (2023)
10. Kong, H., Xu, X., Yu, J., Chen, Q., Ma, C., Chen, Y., Chen, Y.C., Kong, L.: m3track: mmwave-based multi-user 3d posture tracking. In: ACM MobiSys. pp. 491–503 (2022)
11. Lee, S., Rim, J., Jeong, B., Kim, G., Woo, B., Lee, H., Cho, S., Kwak, S.: Human pose estimation in extremely low-light conditions. In: CVPR. pp. 704–714 (2023)
12. Liang, X., Zhang, W., Zhou, H., Wei, Z., Zhu, S., Li, Y., Yin, R., Yuan, J., Gummeson, J.: Sonicmesh: Enhancing 3d human mesh reconstruction in vision-impaired environments with acoustic signals. arXiv preprint arXiv:2412.11325 (2024)
13. Liu, H., Chen, Q., Tan, Z., Liu, J.J., Wang, J., Su, X., Li, X., Yao, K., Han, J., Ding, E., et al.: Group pose: A simple baseline for end-to-end multi-person pose estimation. In: CVPR. pp. 15029–15038 (2023)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. ICLR (2017)
15. Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., Lin, L.: Lstm pose machines. In: CVPR. pp. 5207–5215 (2018)
16. Mosuily, M., Chauhan, J.: Echomotion: Enhancing exercise analysis with acoustic sensing. In: EUSIPCO. pp. 1772–1776 (2025)
17. Nakamura, R., Hirao, Y., Perusquía-Hernández, M., Uchiyama, H., Kiyokawa, K.: Generalizing listening human behavior: 3d human pose estimation using music. In: IEEE GCCE. pp. 1277–1278 (2024)
18. Oumi, Y., Shibata, Y., Irie, G., Kimura, A., Aoki, Y., Isogawa, M.: Acoustic-based 3d human pose estimation robust to human position. In: BMVC (2024)
19. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR. pp. 7753–7762 (2019)
20. Qiu, Z., Yang, Q., Wang, J., Feng, H., Han, J., Ding, E., Xu, C., Fu, D., Wang, J.: Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In: CVPR. pp. 21254–21263 (2023)

21. Raza, A., Yousaf, M.H., Ahmad, W., Velastin, S.A., Viriri, S.: Human fall detection using pose estimation: From traditional machine learning to vision transformers. EAAI **143**, 109809 (2025)
22. dos Reis, E.S., Seewald, L.A., Antunes, R.S., Rodrigues, V.F., da Rosa Righi, R., da Costa, C.A., da Silveira Jr, L.G., Eskofier, B., Maier, A., Horz, T., et al.: Monocular multi-person pose estimation: A survey. PR **118**, 108046 (2021)
23. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: CVPR. pp. 11069–11078 (2022)
24. Shibata, Y., Kawashima, Y., Isogawa, M., Irie, G., Kimura, A., Aoki, Y.: Listening human behavior: 3d human pose estimation with acoustic signals. In: CVPR. pp. 13323–13332 (2023)
25. Shibata, Y., Oumi, Y., Irie, G., Kimura, A., Aoki, Y., Isogawa, M.: Bgm2pose: Active 3d human pose estimation with non-stationary sounds. In: CVPRW (2025)
26. Song, Y., Jin, T., Dai, Y., Song, Y., Zhou, X.: Through-wall human pose reconstruction via uwb mimo radar and 3d cnn. Remote Sensing **13**(2), 241 (2021)
27. Tanigawa, R., Ishii, Y.: hear-your-action: human action recognition by ultrasound active sensing. In: ICASSP. pp. 7260–7264 (2024)
28. Wang, C., Zhang, F., Ge, S.S.: A comprehensive survey on 2d multi-person pose estimation methods. EAAI **102**, 104260 (2021)
29. Wang, F., Zhou, S., Panev, S., Han, J., Huang, D.: Person-in-wifi: Fine-grained person perception using wifi. In: CVPR. pp. 5452–5461 (2019)
30. Yan, K., Wang, F., Qian, B., Ding, H., Han, J., Wei, X.: Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In: CVPR. pp. 969–978 (2024)
31. Yao, S., Piao, A., Jiang, W., Zhao, Y., Shao, H., Liu, S., Liu, D., Li, J., Wang, T., Hu, S., et al.: Stfnets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In: The Web Conf. pp. 2192–2202 (2019)
32. Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. In: CVPR. pp. 7356–7365 (2018)
33. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: ICCV. pp. 11656–11665 (2021)
34. Zheng, Z., Zhang, D., Liang, X., Liu, X., Fang, G.: Radarformer: End-to-end human perception with through-wall radar and transformers. IEEE TNNLS (2023)
35. Zou, S., Xu, Y., Li, C., Ma, L., Cheng, L., Vo, M.: Snipper: A spatiotemporal transformer for simultaneous multi-person 3d pose estimation tracking and forecasting on a video snippet. IEEE TCSVT **33**(9), 4921–4933 (2023)