Less is More: Multimodal Human Pose Estimation with Selective Fusion

Yutong Xu Qianyi Huang* Xu Chen
Sun Yat-sen University
xuyt85@mail2.sysu.edu.cn; {huangqy89,chenxu35}@mail.sysu.edu.cn

Abstract

Human pose estimation plays a vital role in various applications such as human-computer interaction and rehabilitation. To track human poses, multi-modal sensors, e.g., mmWave radar and LiDAR, are often employed to capture motion cues across diverse scenarios. Multi-modal learning has demonstrated great potential for robust pose estimation under challenging conditions such as occlusion and noisy backgrounds. However, existing multi-modal pose prediction methods often assume that the information from different modalities is inherently complementary. Therefore, current mainstream methods consistently tend to utilize all available modality information to accomplish the task. However, this complementarity assumption among modalities does not always hold. We observe that, in some cases, multi-modal fusion may degrade rather than improve performance.

In this paper, we introduce FlexPose, a multimodal framework for human pose prediction that adaptively determines when fusion is beneficial and when single-modality processing is preferable. Specifically, we design an adaptive modality selection module that dynamically assesses inter-modal complementarity, allowing the model to revert to single-modality learning when fusion becomes detrimental. To further enhance robustness, we develop additional modules to handle missing modalities and to exploit temporal dependencies across frames. These designs collectively yield substantial accuracy improvements for human pose prediction. When evaluated on the MM-Fi dataset, our approach achieves relative improvements of 22.63% in MPJPE and 16.47% in PA-MPJPE over the baseline. Our code is available in https://github.com/xyt-fe/FlexPose_Modality_selection.

1. Introduction

Human pose estimation (HPE) is a fundamental task in computer vision, with broad applications in human–computer interaction, sports analytics, physical therapy and elderly care. Accurate human pose estimation across diverse and dynamic scenarios is essential for enabling intelligent systems to understand and interact with humans effectively. Traditionally, HPE relies on vision-based sensors, such as RGB cameras and depth cameras [7, 28]. While effective under ideal conditions, their performance deteriorates significantly in complex environments—for instance, under severe occlusions or poor lighting [15]. Moreover, privacy concerns increasingly limit the deployment of visual sensors in public or personal spaces.

To overcome these limitations, recent research has turned to sensing modalities such as LiDAR and mmWave radar. These sensors are inherently less privacy-invasive and more robust to lighting variations. LiDAR delivers high-precision point clouds and is extremely accurate within close range. On the other hand, radar excels in capturing dynamic motion information and has a detection range that extends over large distances. Consequently, the fusion of LiDAR and radar modalities has emerged as a promising direction for achieving reliable and privacy-preserving pose estimation in real-world environments [10, 17, 21, 25, 31].

Existing multi-modal pose estimation methods typically assume that both LiDAR and radar modalities are consistently available and provide complementary information, making fusion effective for pose estimation tasks. However, in practical applications, these assumptions often do not hold. When different sensors capture highly redundant or correlated information, adding an additional modality increases the feature dimensionality without providing substantial new discriminative information. This redundancy can lead to negative transfer, where the overlap between modalities creates confusion in the model. Furthermore, incorporating low-quality data from one modality into the multimodal fusion pipeline introduces extra noise, which degrades the model's performance. Instead of improving predictions, it distracts the learner and increases the complexity of the model, making it harder to optimize.

To address this challenge in multimodal human pose estimation, we propose FlexPose, a multi-modal framework

*Corresponding author.

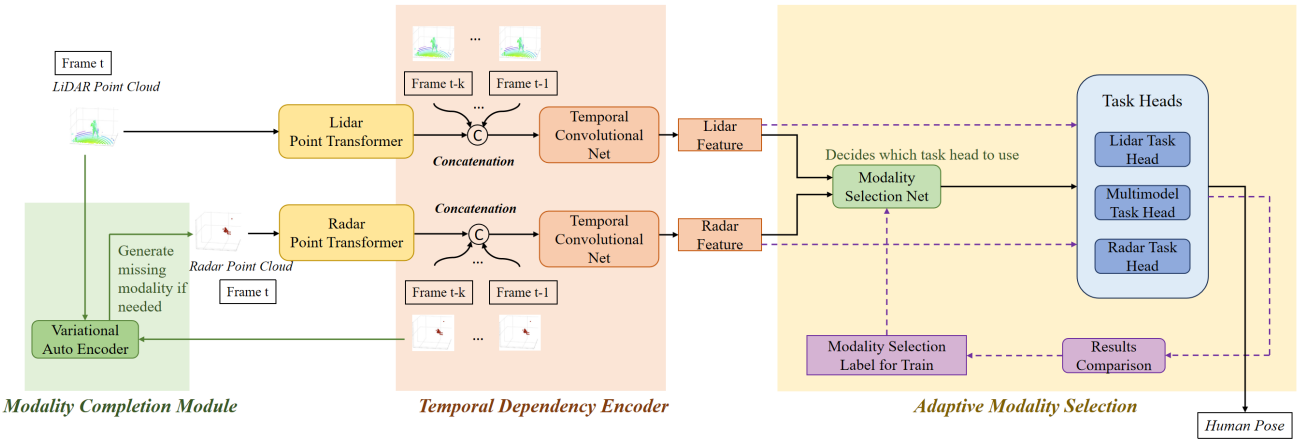


Figure 1. **The framework of FlexPose.** During training, the model first produces three preliminary predictions using the three task heads (as shown by the dotted lines). These predictions are compared to determine the optimal modality configuration for that sample, yielding a modality-selection label used to supervise the Modality Selection Network. During inference, the extracted features are fed into the Modality Selection Network, which selects the most appropriate task head for final prediction.

comprising three modules: *Adaptive Modality Selection (AMS)*, *Modality Completion Module (MCM)* and *Temporal Dependency Encoder (TDE)*. The Adaptive Modality Selection module is the cornerstone of FlexPose, designed to dynamically decide whether multimodal or single-modality learning is more appropriate for each sample. It employs mutual information to quantify both the similarity between modalities and the signal quality of each modality, enabling the model to select the most effective modality combination. This adaptive selection improves efficiency by ensuring that multimodal fusion is used only when it provides a clear benefit, thus avoiding unnecessary complexity and overfitting. The *Modality Completion Module* tackles the instability in mmWave radar hardware by filling the gap in the sensor data using a Variational Auto encoder (VAE). It leverages information from the available modality in the current sample to provide contextual understanding, while previous samples of the missing modality inform its modal-specific attributes. This approach improves the quality of the generated sensor data. Finally, the *Temporal Dependency Encoder* exploits the continuity of human actions to capture temporal dependencies across multiple frame samples, which reduces jitter and enhances the stability of pose estimation.

Our main contributions can be summarized as follows:

- We observe that multimodal learning does not always outperform single-modality learning. We propose that the key to determining the most effective modality combination lies in assessing both the similarity between modalities and the quality of each modality.
- We propose FlexPose, a multimodal fusion framework for human pose estimation with LiDAR and mmWave radar. It includes an adaptive modality selection module that leverages mutual information to decide whether to use multi-modal learning or uni-modal learning at each instance.
- We extensively test FlexPose on the MM-Fi dataset. Compared with the baseline, the MPJPE and PA-MPJPE metrics reduced by 22.63% and 16.47%, respectively. The modality selection module is a plug-and-play component that can be readily integrated into existing state-of-the-art multimodal learning frameworks.

2. Related work

2.1. Multimodel Human Pose Estimation

In the early days of multimodal human pose estimation, vision based sensors were the primary data source. Representative work includes FuseNet [9], VoxelPose [23, 32, 35], and DualPoseNet [4, 14], which exploit RGB and RGB-D modalities [1, 2], as well as UPPET [5], which combine RGB and thermal imagery. However, with growing privacy concerns, less intrusive sensors such as LiDAR and radar are increasingly combined with visual modalities. For example, mmPose [11, 18, 19] and LidarCap [12] fuse images with point clouds.

Going one step further, recent works perform fusion using only LiDAR and radar signals, without relying on visual input. Existing fusion strategies can be broadly categorized into three levels. Point-level fusion projects radar point clouds into the LiDAR coordinate system and concatenates them into a unified point cloud [20]. While conceptually simple, this approach is highly sensitive to alignment errors due to the significant density discrepancy between LiDAR and radar points. BEV-level fusion avoids the 3D alignment challenge by projecting both modalities

onto a shared bird's-eye-view plane and merging the resulting 2D feature maps. Although effective, this projection inevitably loses vertical information, which must be compensated using additional height channels. Representative BEV-based methods include RaLiBEV [30] and BEV-Guide [16]. Feature-level fusion offers the most flexible and widely adopted solution. In this paradigm, each modality is processed by its own encoder, and the resulting modality-specific features are fused at the representation level. This strategy adapts well to diverse scenarios, supports various downstream tasks within a unified framework, and is used in methods such as RLNet [27] and DeepFusion [6]. FlexPose follows this feature-level fusion paradigm.

2.2. Adaptive Modality Fusion

In multimodal learning, we observe that fusing all modalities is not always optimal. In some cases, the information captured by different modalities is highly similar, leading to weak complementarity and high redundancy. In such instances, multimodal fusion can be counterproductive or even harmful. Previous methods often address this by weakening less important modalities. Common approaches include modality weighting [13, 22, 24, 26, 33, 34], where an adaptive mechanism allows the model to learn the importance of each modality, assigning lower weights to less relevant ones, while still fusing all modalities.

However, we argue that in scenarios where modalities are highly similar or of low quality, it is more effective to discard unnecessary modalities entirely. The traditional modality weight allocation approach only weakens low-quality modalities, but still incorporates the unfavorable modality into the fusion. Weight learning tends to focus on global preferences, pushing the network toward the dominant modality, even when the modality combination is not optimal for the current task. Moreover, weight learning is driven by the loss function and backpropagation, which typically evaluates final prediction accuracy without considering whether each modality is truly beneficial for the task. As a result, the gradient signal is indirect and may be obscured by irrelevant modality information.

X-Fi [3] represents the state of the art on the MM-Fi dataset [29], achieving adaptive modality fusion that enables modalities to be flexibly added or removed. Our work is orthogonal to X-Fi, and our framework can be seamlessly integrated into X-Fi to provide additional performance improvements.

3. Method

The overall framework of FlexPose is shown in Fig. 1. FlexPose mainly consists of three parts: modality completion module, temporal dependency encoder, and adaptive modality selection. We start with the adaptive modality selection, as it is the key component in FlexPose.

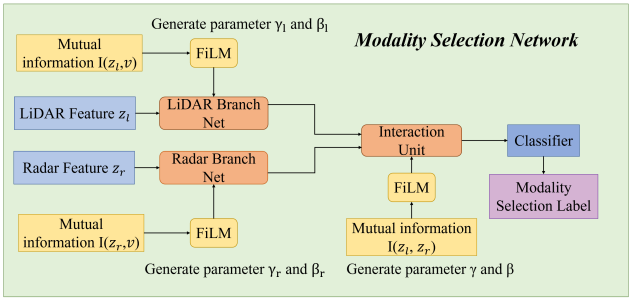


Figure 2. **The framework of Modality Selection Network.** In the Modality Selection Network, the main framework consists of two branch nets and an interaction unit. Three sets of mutual information are processed through FiLM to generate three sets of parameters, which participate in modulation at different stages. The mutual information between z_l, z_r and v are introduced in the corresponding branch nets, and the mutual information between z_l and z_r is introduced before the interaction unit.

3.1. Adaptive Modality Selection

3.1.1. Less is More: Selective Fusion

Multimodal fusion does not inherently guarantee performance gains. Its effectiveness depends on two key factors: (1) the degree of complementarity across modalities, and (2) the quality and reliability of each modality.

When the information captured by different sensors is highly redundant or even nearly identical, the diversity between modalities decreases. In such cases, introducing an additional modality increases the input dimensionality but contributes little new discriminative information beyond what a strong single modality already provides. In other words, the signals carried by the additional modalities can be fully interpreted by the existing modality. This redundancy can even cause negative transfer, where overlapping or semantically similar features interfere with each other.

A second issue arises when one modality becomes low-quality or unreliable, such as when it suffers from unstable sampling, noise, or reduced resolution. Incorporating such a modality injects extra noise into the fusion pipeline, distracts the learner, and often degrades the final performance rather than improving it.

Therefore, when modalities are either highly redundant or low in quality, blindly fusing them may be harmful. Instead, it is often more advantageous to fall back to single-modality learning. Motivated by this observation, we design an Adaptive Modality Selection module that decides, on a per-sample basis, whether multimodal fusion or single-modality learning is more suitable.

3.1.2. Modality Selection Network

We observed that the key to choosing between multimodal learning and single-modal learning lies in measuring the correlation between modalities and the quality of each

modality. We use mutual information to quantify both.

Modality correlation: Normalized Mutual information (NMI) measures how much information one variable provides about another. Computing NMI between two modality features quantifies how much information they share—higher NMI indicates stronger similarity and lower complementarity. Thus, NMI allows us to estimate how redundant two modalities are when predicting the same target.

Let m_l denote the LiDAR modality and m_r the mmWave Radar modality, with corresponding features z_l and z_r, respectively. We compute their normalized mutual information as:

$$I_s(z_l, z_r) = \iint p(z_l, z_r) \log \left(\frac{p(z_l, z_r)}{p(z_l)p(z_r)} \right) dz_l dz_r, \quad (1)$$

$$I(z_l, z_r) = \frac{I_s(z_l, z_r)}{\sqrt{H(z_l)H(z_r)}}, \quad (2)$$

which ranges from 0 to 1. Here $H(z_l)$ and $H(z_r)$ denote the entropy of z_l and z_r, respectively. Values close to 1 indicate highly similar modalities (i.e., strong redundancy), while values near 0 indicate low similarity and potentially strong complementarity.

Modality quality: Mutual information between a modality feature and the ground truth reflects how informative the modality is for the prediction task. Higher NMI means the modality contains more relevant cues and is therefore of higher quality. However, the ground truth y is only available during training and cannot be accessed during inference. To address this, we introduce a learnable vector v as a semantic proxy representation of the ground truth. The purpose of learning the vector v is to capture the relationship between modality features and the semantics of pose y as accurately as possible, rather than directly predicting y. In testing, v will be used to compute mutual information with z_l and z_r, allowing the mutual information to represent the relationship between the features and the pose. This is a common practice in mutual information theory and representation learning [8]. Thus, the measures of interest become $I(z_l, v)$ and $I(z_r, v)$.

Modality Selection Net: Fig. 2 shows the architecture of the modality selection network. The backbone of the Modality Selection Network consists of two branch nets and an interaction unit.

Each branch network is dedicated to assessing a single modality, it consists of a series of convolutional layers, batch normalization layers, ReLU activation functions, and dropout layers. At the branch net stage for each modality, we inject quality-reflecting mutual information, $I(z_l, v)$ and $I(z_r, v)$, to guide the network in learning the modality's effectiveness. Specifically, $I(z_l, v)$ and $I(z_r, v)$ are fed into separate Feature-wise Linear Modulation (FiLM) layers. FiLM is a network module that uses input as conditional information to modulate features, it typically consists of convolutional and fully-connected layers, and generates parameter scaling factors γ and shifting factors β based on the input vector. Here, FiLM produce the modulation variables γ_l, β_l and γ_r, β_r, respectively, to modulate the features of the corresponding modality:

$$\left(\gamma_{l/r}, \beta_{l/r}\right) = \text{FiLM}(I(z_{l/r}, v))$$

$$\tilde{z}_{l/r} = \gamma_{l/r} \cdot \text{BranchNet}(z_{l/r}) \; + \; \beta_{l/r}.$$

The interaction unit consists of several attention layers, a learnable weight for connection, and a convolutional block for cross-modal feature fusion. It focuses on evaluating the correlation between the two modalities. The Interaction unit focuses on evaluating the correlation between the two modalities. It utilizes an attention mechanism to learn the relationships between modalities and applies weighting during feature interaction. At this stage, we inject the correlation-reflecting mutual information $I(z_l, z_r)$ into the Interaction unit. This information is processed by another FiLM layer, generating the variables γ_{IU}, β_{IU}, which modulate the features inside the Interaction unit:

$$(\gamma_{IU}, \beta_{IU}) = \text{FiLM}(I(z_l, z_r)),$$

$$\tilde{z} = \gamma_{IU} \cdot \text{InterUnit}(\tilde{z}_l + \tilde{z}_r) \; + \; \beta_{IU}.$$

This allows the unit to better learn the relationship between modalities. The output features of the Interaction unit are classified to obtain the modality selection label.

Through this feature-level modulation process, mutual information imposes conditional constraints on feature representations, ensuring they are aligned with the data's inherent requirements. This approach not only strengthens the feature representation by incorporating mutual information but also enables adaptive feature selection and refinement via the FiLM layer, allowing the model to dynamically adjust and select the most relevant features for the current task.

3.1.3. Modality-Selection Labels

To train the Modality Selection Network, we generate ground truth labels for modality selection. With two modalities, three labels are defined: (A) multimodal fusion performs best, (B) modality m_l performs best, and (C) modality m_r performs best. The features of LiDAR z_l and Radar z_r are passed through three separate paths:

Path A Multimodal learning path P_A: Both z_l and z_r are fed into this path for multimodal feature fusion and then passed to the multimodal task head to estimate human pose. We get the estimated result r_A.

Path B & C Single-model learning paths P_B and P_C for m_l and m_r, respectively: Each path has its own single-modal task head. z_l and z_r are sent to their respective paths independently for single-modal learning, and the outputs are

used to estimate human pose separately. We get the pose estimation result r_B and r_C.

Paths P_A, P_B, and P_C each employ the Mean Squared Error (MSE) loss function for loss computation, while the overall loss of the model is the sum of the losses from these three paths:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_A + \mathcal{L}_B + \mathcal{L}_C \\ &= \sum(\|\mathbf{y} - \mathbf{r}_A\|^2 + \|\mathbf{y} - \mathbf{r}_B\|^2 + \|\mathbf{y} - \mathbf{r}_C\|^2)\end{aligned} \tag{3}$$

The model trains all three subsections simultaneously. For human pose estimation task, it produces three results per sample. We rank the results by computing the Mean Per Joint Position Error (MPJPE) between the predicted result and ground truth y values, and denote as:

$$E_m = \frac{1}{N}\sum_{i=1}^{N}\left\| \mathbf{J}_{r_m} - \mathbf{J}_y \right\|_2,\ m = A, B, C \tag{4}$$

Here $\mathbf{J}_{r_m}$ is the predicted joint position, $\mathbf{J}_y$ is the true joint position, and N is the number of joints. The result with the smallest error yields the best modality or modality combination. Accordingly, we obtain the modality-selection label:

$$L = \arg\min_m \ E_m, \tag{5}$$

which are used to supervise the Modality Selection Network.

3.1.4. Working Flow

This subsection outlines the complete workflow for Adaptive Modality Selection. Prior to training, we generate modality selection labels by feeding LiDAR and Radar features into their respective single-modality task heads, as well as into a multimodal task head. Each head produces a separate prediction, and by comparing their accuracy, we determine the modality selection labels for the entire dataset. Using these labels as ground truth, we train the Modality Selection Network in a supervised manner.

During testing, the modality features are input into the trained Modality Selection Network, which decides whether to use single-modality or multimodal learning. Based on this decision, the features are routed to the appropriate task head for the final prediction.

3.2. Modality Completion Module

In real-world deployment, mmWave radar often suffers from short data dropouts due to hardware instability. These missing segments typically span 10-50 milliseconds and result in the loss of one or several consecutive frames. To ensure robust modality selection and feature fusion, FlexPose reconstructs missing radar features rather than discarding incomplete samples.

FlexPose generates these features by combining cross-modal context and temporal continuity. The available modality in the current frame (e.g., LiDAR) provides reliable environmental cues, while preceding available radar frames supply modality-specific characteristics. Integrating these two sources enables the model to infer a radar representation that is consistent with the current scene and preserves the intrinsic behavior of the radar signal over time.

To implement this, we adopt a Variational Auto Encoder (VAE). The VAE encodes the non-missing modality feature from the current frame and the available radar feature from the preceding timestamps into latent representations, capturing both scene context and radar-specific temporal patterns. The combined latent vector is then decoded to reconstruct the missing radar feature, ensuring compatibility with both the current environment and the temporal dynamics of the radar modality.

3.3. Temporal Dependency Encoder

In human pose estimation, continuous pose sequences are split into frame samples, where poses across frames are physically continuous. Actions in later frames evolve from those in earlier ones—this is the temporal dependency we leverage.

Because of this temporal continuity, the data from the immediately preceding frame shares a physically continuous action relationship with the current frame's pose. Therefore, when estimating the current pose, we can incorporate information from preceding frames in addition to the current frame's data.

The challenge lies in effectively combining the preceding frame's modality data with that of the current frame. To address this, we use a Temporal Convolutional Network (TCN), a convolutional architecture designed for sequential data. TCN utilizes mechanisms like causal and dilated convolutions, which efficiently capture temporal dependencies. After initial processing with the Point Transformer, the data is passed through the TCN to obtain the final temporal features for each modality, i.e., z_l for LiDAR and z_r for mmWave radar:

$$z = \text{TCN}([\Gamma(d_{[t-k]}) : \Gamma(d_{[t-k+1]}) : \cdots : \Gamma(d_{[t]})]). \tag{6}$$

Here Γ denotes point transformer, d denotes sensor data and $[:]$ denotes concatenating operation. We will show the optimal value of k in Section 4.2.4.

4. Experiment

4.1. Experimental Setting

Dataset: We test FlexPose on MM-Fi, a multimodal, non-intrusive dataset for 4D human perception [29]. MM-Fi contains more than 320,000 synchronized frames from 40

Table 1. **Overall Results.** Both MPJPE and PA-MPJPE are measured in millimeters. The percentage denotes the relative error reduction of FlexPose over the baseline.

Metric	Model	Head	Body	Hip	Knees	Feet	Shoulder	Elbows	Hands	Average
MPJPE	MM-Fi	170.0	138.0	131.7	138.8	145.2	160.5	237.0	353.6	184.3
	FlexPose	127.4	98.4	93.3	100.0	104.7	119.2	191.8	303.1	142.6
	Percentage	25.06%	28.70%	29.16%	27.95%	27.89%	25.73%	19.07%	14.28%	22.63%
PA-MPJPE	MM-Fi	85.2	47.9	46.7	62.6	118.2	68.5	133.3	234.8	99.6
	FlexPose	62.7	37.4	34.2	55.0	102.4	54.0	115.8	194.2	83.2
	Percentage	26.41%	21.92%	26.77%	12.14%	13.37%	21.17%	13.13%	17.29%	16.47%

participants and provides multiple modalities such as LiDAR and mmWave radar point clouds. Annotations include activity labels, 2D/3D human keypoints, 3D body positions, and dense 3D pose. The dataset comprises 27 activity classes—14 daily activities and 13 rehabilitation exercises—making it suitable for smart home, health monitoring, and metaverse applications. Data were collected with a customized multi-sensor platform, with all modalities strictly time aligned and uniformly sampled at 10 Hz.

The dataset is split into training and test sets in a cross-subject manner: data from 32 subjects are used for training, and data from the remaining 8 subjects are reserved for testing. All 27 activity classes are included in the experiment. In MM-Fi, the mmWave radar modality missing rate is approximately 17%.

Metrics: Our evaluation metrics are Mean Per Joint Position Error (MPJPE) and Procrustes Aligned Mean Per Joint Position Error (PA-MPJPE), which are the widely-used error metrics in 3D human pose estimation. MPJPE computes the average Euclidean distance between predicted and ground truth joints in 3D space, directly reflecting the accuracy of joint locations and global positioning. PA-MPJPE refines MPJPE by further performing a Procrustes transformation, allowing only rotation, scaling, and translation, to optimally align the predicted skeleton with the ground truth skeleton, thereby removing systematic biases due to global orientation, size, and location. This focuses the evaluation on the relative geometric accuracy of the joints themselves. Both metrics are measured in millimeters, and lower values indicate more accurate pose estimation.

Baseline: We adopt the method proposed in MM-Fi [29] as our baseline, which assigns an adaptive weight to each modality. Both LiDAR and radar inputs are encoded using Point Transformers, with separate parameter sets tailored to the distinct characteristics of each modality. The resulting modality-specific features are then fused using the adaptive weights produced by the modality-weight module, and the fused representation is subsequently fed into a task head for human pose estimation. Note that our reproduced results of MM-Fi are consistent with those reported by X-Fi [3] under the same evaluation protocol. The results in Tab. 1 correspond to the across-subject setting, which is more challenging than the random split setting used in [3].

Implementation Details: FlexPose is trained on NVIDIA A100 GPU, using the Adam optimizer and MSE loss function. We trained the model for 30 epochs, with a batch size of 64 and a learning rate of 0.001.

4.2. Results

In this section, we first present the overall results for FlexPose, followed by a detailed analysis for the modality selection module. Finally, we give the results for abalation study.

4.2.1. Overall Result

In the dataset, there are 17 joints in total. We consolidate some of these joints for representation. We categorize the total 17 joints into eight classes: Head (nose, neck), Body (clavicle, torso, pelvis), Hip (left hip, right hip), Knees (left knee, right knee), Feet (left foot, right foot), Shoulder (left shoulder, right shoulder), Elbows (left elbow, right elbow), and Hands (left hand, right hand). The results for each body part and the average of the total 17 joints are presented in Tab. 1.

When analyzing the results by body part, we observe that both the baseline and FlexPose achieve the lowest errors on the body and hips. These joints typically undergo smaller positional changes across actions, and the corresponding point-cloud signals are denser and more stable, making them easier to predict. In contrast, hands and elbows consistently exhibit the highest errors because they experience larger motion amplitudes, and the point-cloud data for these distal joints tend to be much sparser, increasing prediction difficulty. In general, the farther a joint is from the torso, the larger its estimation error tends to be; for example, feet are less accurate than knees, and hands are less accurate than elbows.

Despite these inherent challenges, FlexPose outperforms the baseline across all body parts, reducing the error values consistently. Averaged over all 17 joints, FlexPose reduces MPJPE and PA-MPJPE from 184.3 mm and 99.6 mm (base-

line) to 142.6 mm and 83.2 mm, representing improvements of 22.63% and 16.47%, respectively. These results demonstrate that our framework delivers consistent gains across individual joints and overall, effectively lowering estimation error and validating the effectiveness of FlexPose for multimodal human pose estimation.

Digging deeper into the joint-wise improvement percentages, we observe that FlexPose achieves the largest gains on the body and hip regions, whereas improvements on elbows and hands are noticeably smaller. In general, joints that are inherently easier to estimate tend to benefit more, while highly dynamic or distal joints gain less. This pattern suggests that our enhancement mechanism is most effective in regions where the point cloud is dense and joint motion is relatively limited, and less effective in areas with sparse data and frequent motion. Consequently, the model's improvements are concentrated on body parts with more stable structure. Additionally, joints such as the hands already exhibit large absolute errors; therefore, even a meaningful reduction in error may correspond to only a modest relative improvement.

4.2.2. Analysis of Adaptive Modality Selection

In the Adaptive Modality Selection module, we use a Modality Selection Network to determine, on a per-sample basis, whether the model should adopt multimodal learning or rely on a single modality. The network formulates this decision as a classification task by predicting modality-selection labels. During testing, it achieves an accuracy of 80.94% on label prediction, demonstrating that the proposed architecture can effectively distinguish when multimodal fusion is beneficial and when single-modality learning is preferable.

To estimate the upper bound of our modality selection strategy, we conduct an oracle experiment in which all modality features are fed into the three task heads and their outputs are directly compared. The best-performing result among the three is selected as the final pose estimation outcome. This simulates an ideal scenario where the Modality Selection Network achieves 100% decision accuracy. Under this oracle setting, the model attains MPJPE and PA-MPJPE scores of 140.4 mm and 82.8 mm, representing 15.27% and 10.29% error reductions over the baseline (184.3 mm and 99.6 mm).

Interestingly, this oracle performance is only marginally better than the actual performance of FlexPose. This is because, in most cases where the Modality Selection Network makes an incorrect choice, the performance gap between single-modality and multimodal learning is inherently small, and the preferred modality is ambiguous. As a result, even perfect selection would not yield substantial gains.

In our results, the total number of samples is 64,152, among which 30,748 samples achieve better performance with multimodel learning, accounting for 47.93%, while 33,404 samples perform better with single-modal learning, accounting for 52.07%. However, within the single-modality cases, LiDAR consistently outperforms radar. This indicates that when single-modality learning is sufficient, LiDAR alone typically provides the strongest signal, and adding radar in these scenarios often introduces noise that can degrade performance.

4.2.3. Comparison with Adaptive Weight Mechanism

We compare the modality selection strategy in FlexPose and the adaptive weighting mechanism in dataset MM-Fi. We summarize the 27 activities in the dataset and analyze the label/weight distribution. The results are shown in Fig. 3.

Fig. 3(a) and Fig. 3(c) display the label counts for two different activities, i.e., "limb extension" and "lunge". As shown in the figures, the sample counts for optimal multi-modal learning and optimal single-modality learning differ across activities. This indicates that FlexPose successfully learns when multi-modal or single-modal learning is more appropriate.

Fig. 3(b) and Fig. 3(d) illustrate the weight distribution statistics for LiDAR and mmWave radar, respectively. From these figures, we observe that, regardless of the activity, the model consistently assigns stable weights—approximately 0.8 to LiDAR and 0.2 to mmWave radar. This suggests that the model does not learn to adjust the roles of the modalities based on varying activities.

4.2.4. Analysis of Temporal Dependency Encoder

As mentioned in Sec. 3.3, to exploit the temporal continuity of human poses, FlexPose incorporates not only the current frame but also the preceding k frames. We sequentially set $k = 1, 2, 3, 4$ and identify the value that yields the best performance on the MM-Fi dataset. The result is shown is in Tab. 3.

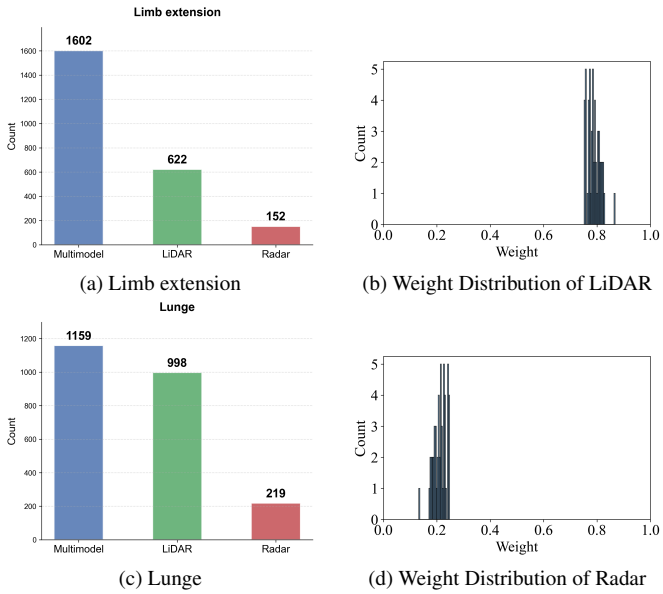
By comparing the results of different k values, it can be observed that the best performance is achieved when $k = 2$, where the MPJPE and PA-MPJPE metrics are 142.6 mm and 83.2 mm, respectively. When $k = 1$, the performance degrades because the valid information from two preceding frames is not utilized, resulting in fewer informative inputs and insufficient exploitation of temporal cues. When $k = 3$ or $k = 4$, the performance also declines compared to $k = 2$ as frames that are too far from the current one may have weak physical correlations due to longer temporal gaps. Therefore, when selecting the value of k, it is important to make full use of the most relevant preceding frames while avoiding those with weak temporal correlation.

4.2.5. Abalation study

We conduct ablation studies on the three core components of FlexPose– Modality Completion Module (MCM), Temporal Dependency Encoder (TDE) and Adaptive Modality

Table 2. **The results of ablation study**. The increase percentage indicates how much the model's MPJPE and PA-MPJPE metrics increase compared with the complete FlexPose when the corresponding module is removed.

Module	MPJPE	Increase Percentage	PA-MPJPE	Increase Percentage
FlexPose	142.6 mm	—	83.2 mm	—
Without MCM	154.3 mm	8.20%	87.9 mm	5.65%
Without TDE	162.8 mm	14.17%	91.2 mm	9.62%
Without AMS	165.7 mm	16.20%	92.3 mm	10.94%



(a) Limb extension (b) Weight Distribution of LiDAR

(c) Lunge (d) Weight Distribution of Radar

Figure 3. **Comparison between modality selection and adaptive weight mechanism.** The left two figures show the label distribution for different modality combinations in FlexPose (Top: limb extension; Bottom: lunge); The right two figures show the weight distribution assigned to LiDAR and Radar by the adaptive weight strategy in MM-Fi for the 27 activities.

Table 3. **The results with different values of** k. By comparison, it can be seen that the optimal experimental result is achieved when $k = 2$.

K	MPJPE	PA-MPJPE
1	144.2 mm	83.9 mm
2	**142.6** mm	**83.2** mm
3	144.9 mm	83.7 mm
4	145.6 mm	84.2 mm

Selection (AMS) – to evaluate the contribution of each module. The ablation settings are as follows:

- Full model (MCM + AMS + TDE) vs. without MCM
- Full model vs. without TDE
- Full model vs. without AMS

The results are presented in Tab. 2. As shown, removing any of the three modules increases human pose estimation error and degrades overall performance, confirming the necessity of each component.

The TDE module leverages temporal continuity by incorporating information from preceding frames. Without TDE, the model relies solely on the current frame and loses access to historical temporal cues, resulting in a clear drop in prediction accuracy.

The MCM module reconstructs missing-modality features. When MCM is removed, missing sensor data can disrupt feature fusion and mislead AMS during modality selection, ultimately harming model performance.

The AMS module enables the model to identify samples where single-modality learning outperforms multimodal fusion. By selecting the appropriate learning strategy per sample, AMS improves estimation accuracy and boosts overall performance. Without AMS, the model is forced to fuse modalities for all samples, even when fusion is suboptimal, leading to noticeable degradation.

Through all the experiments above, we convincingly demonstrate the positive impact of FlexPose on human pose estimation and verify that each of the three proposed modules is indispensable.

5. Conclusion

In this paper, we present FlexPose, a framework for multimodal human pose estimation. Our study reveals an important insight: in certain scenarios when the similarity between modalities is excessively high or when one modality is of low quality, single-modality learning can outperform multimodal learning. Motivated by this finding, we design the Adaptive Modality Selection module that determines, on a per-sample basis, whether multimodal or single-modality learning is more beneficial. In addition, we introduce two complementary modules: the Modality Completion Module and the Temporal Dependency Encoder, to further enhance the model. Experimental results demonstrate that our proposed framework achieves superior performance on human pose estimation tasks, reducing the MPJPE and PA-MPJPE by 22.63% and 16.47% on a public dataset.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China under

Grant 62472452; in part by Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515010262, 2023B1515120058)

References

[1] Martin Brenner, Napoleon H Reyes, Teo Susnjak, and Andre LC Barczak. Rgb-d and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:82410–82442, 2023. 2

[2] Martin Brenner, Napoleon H Reyes, Teo Susnjak, and Andre LC Barczak. Rgb-d and thermal sensor fusion: A systematic literature review. *IEEE Access*, 11:82410–82442, 2023. 2

[3] Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. *arXiv preprint arXiv:2410.10167*, 2024. 3, 6

[4] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 2

[5] Mickael Cormier, Andreas Specker, and Jürgen Beyerer. Uppet: Unified pedestrian pose estimation in thermal imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4551–4560, 2025. 2

[6] Florian Drews, Di Feng, Florian Faion, Lars Rosenbaum, Michael Ulrich, and Claudius Gläser. Deepfusion: A robust and modular 3d object detector for lidars, cameras and radars. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 560–567. IEEE, 2022. 3

[7] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. 1

[8] Gokul Gowri, Xiao-Kang Lun, Allon M Klein, and Peng Yin. Approximating mutual information of high-dimensional variables using learned representations. *Advances in Neural Information Processing Systems*, 37:132843–132875, 2024. 4

[9] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016. 2

[10] Xun Huang, Ziyu Xu, Hai Wu, Jinlong Wang, Qiming Xia, Yan Xia, Jonathan Li, Kyle Gao, Chenglu Wen, and Cheng Wang. L4dr: Lidar-4dradar fusion for weather-robust 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3806–3814, 2025. 1

[11] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 2

[12] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20502–20512, 2022. 2

[13] Ruimin Li, Jiajun Xiang, Feixiang Sun, Ye Yuan, Longwu Yuan, and Shuiping Gou. Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 26:852–863, 2023. 3

[14] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3560–3569, 2021. 2

[15] Michael Lötscher, Nicolas Baumann, Edoardo Ghignone, Andrea Ronco, and Michele Magno. Assessing the robustness of lidar, radar and depth cameras against ill-reflecting surfaces in autonomous vehicles: An experimental study. In *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, pages 1–6. IEEE, 2023. 1

[16] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21960–21969, 2023. 3

[17] Xiangyuan Peng, Yu Wang, Miao Tang, Bierzynski Kay, Lorenzo Servadei, and Robert Wille. Moral: Motion-aware multi-frame 4d radar and lidar fusion for robust 3d object detection. *arXiv preprint arXiv:2505.09422*, 2025. 1

[18] Arindam Sengupta and Siyang Cao. mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8418–8429, 2022. 2

[19] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE sensors journal*, 20 (17):10032–10044, 2020. 2

[20] Jingyu Song, Lingjun Zhao, and Katherine A Skinner. Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18250–18257. IEEE, 2024. 2

[21] Jingyu Song, Lingjun Zhao, and Katherine A Skinner. Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18250–18257. IEEE, 2024. 1

[22] Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 1823–1833, 2020. 3

[23] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European conference on computer vision*, pages 197–212. Springer, 2020. 2

[24] Hu Wang, Salma Hassan, Yuyuan Liu, Congbo Ma, Yuanhong Chen, Qing Li, Jiahui Geng, Bingjie Wang, Yu Tian, Yutong Xie, et al. Meta-learned modality-weighted knowledge distillation for robust multi-modal learning with missing data. *arXiv preprint arXiv:2405.07155*, 2024. 3

[25] Yingjie Wang, Jiajun Deng, Yao Li, Jinshui Hu, Cong Liu, Yu Zhang, Jianmin Ji, Wanli Ouyang, and Yanyong Zhang. Bi-lrfusion: Bi-directional lidar-radar fusion for 3d dynamic object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13394–13403, 2023. 1

[26] Yifeng Wang, Jiahao He, Di Wang, Quan Wang, Bo Wan, and Xuemei Luo. Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis. *Neurocomputing*, 572:127181, 2024. 3

[27] Ruoyu Xu and Zhiyu Xiang. Rlnet: Adaptive fusion of 4d radar and lidar for 3d object detection. In *European Conference on Computer Vision*, pages 181–194. Springer, 2024. 3

[28] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023. 1

[29] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36:18756–18768, 2023. 3, 5, 6

[30] Yanlong Yang, Jianan Liu, Tao Huang, Qing-Long Han, Gang Ma, and Bing Zhu. Ralibev: Radar and lidar bev fusion learning for anchor box free object detection systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[31] Yanlong Yang, Jianan Liu, Tao Huang, Qing-Long Han, Gang Ma, and Bing Zhu. Ralibev: Radar and lidar bev fusion learning for anchor box free object detection systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1

[32] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision*, pages 142–159. Springer, 2022. 2

[33] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*, 2023. 3

[34] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023. 3

[35] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2613–2626, 2022. 2