# Physics-Informed Learning for Human Whole-Body Kinematics Prediction via Sparse IMUs

Cheng Guo[1,2], Giuseppe L'Erario[1,2], Giulio Romualdi[1], Mattia Leonori[3], Marta Lorenzini[3], Arash Ajoudani[3], Daniele Pucci[1,2]

***Abstract*— Accurate and physically feasible human motion prediction is crucial for safe and seamless human-robot collaboration. While recent advancements in human motion capture enable real-time pose estimation, the practical value of many existing approaches is limited by the lack of future predictions and consideration of physical constraints. Conventional motion prediction schemes rely heavily on past poses, which are not always available in real-world scenarios. To address these limitations, we present a physics-informed learning framework that integrates domain knowledge into both training and inference to predict human motion using inertial measurements from only 5 IMUs. We propose a network that accounts for the spatial characteristics of human movements. During training, we incorporate forward and differential kinematics functions as additional loss components to regularize the learned joint predictions. At the inference stage, we refine the prediction from the previous iteration to update a joint state buffer, which is used as extra inputs to the network. Experimental results demonstrate that our approach achieves high accuracy, smooth transitions between motions, and generalizes well to unseen subjects. The source code and data are available at** $https://github.com/ami-iit/paper_guo_2025_iros_human_kinematics_prediction$**.**

## I. INTRODUCTION

The popularity of robots has surged in recent years due to their environmental interaction capabilities, particularly through the growing focus on *Human-Robot Collaboration* (HRC), which envisions robots working alongside humans to enhance productivity and efficiency [1]. A vital trait for robots to ensure safe and seamless collaboration is the ability to predict human intentions and future movements. With progress in sensing technologies and deep learning algorithms, various methods have been developed to reconstruct human motions from sensor measurements [2]. However, inferring human intentions and predicting future movements from partial observations remains a significant challenge due to the complexity and adaptability of human dynamics. This paper proposes a framework that adopts the idea of physics-informed (PI) learning to enable the prediction of physically plausible human whole-body joint kinematics from sparse inertial measurements.

Traditional *human motion capture* techniques primarily rely on *optical-based* methods, using monocular RGB cameras [3] or depth cameras [4] to reconstruct human poses. While these methods achieve high accuracy, they are constrained by narrow activity spaces and are highly sensitive to environmental factors such as lighting and occlusion. Consequently, *inertial-based* approaches have gained increasing attention. Commercial motion capture solutions [5], [6] provide precise pose estimation but require a dense placement of numerous inertial measurement units (IMUs), which could be inconvenient for industrial scenarios. Recent efforts have explored motion reconstruction with fewer IMUs [7]–[10], but challenges such as long-term drift, resolving ambiguous poses and predicting future motions remain, which are critical for improving efficiency and robustness in HRC settings.

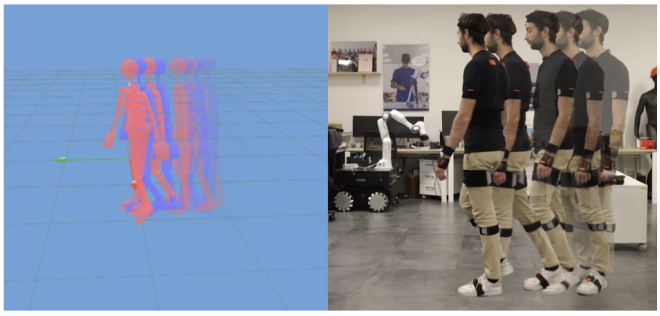


Fig. 1: The proposed approach predicts human joint kinematics via 5 IMUs attached to the pelvis, left/right forearms, and left/right lower legs. The blue model is the reconstructed human motion at the current timestamp (corresponding to the human subject on the right), and the red model is the prediction 10 steps ahead (about 167 ms).

To enable robots to predict human motions, many studies leverage deep learning techniques and frame the task as a *sequence-to-sequence* problem based on past poses [11]–[13]. However, when modeling humans as highly articulated kinematics chains, motion is typically represented as the evolution of whole-body joint positions. In practice, motion capture systems record limb poses, thus requiring an inverse kinematics solution to map these limb configurations to joint positions. Few works have pursued *End-to-End* approaches using wearable sensors [14]–[16], these methods often face restrictions such as short prediction horizons and complex sensor integration, limiting their real-world applicability.

More recently, the approach of using *Physics-Informed Neural Networks* (PINNs) has shown considerable potential in solving complex problems in various scientific and engineering fields where data is limited or noisy [17]. By embedding

*This work was supported by the Italian National Institute for Insurance against Accidents at Work (INAIL) ergoCub Project.

[1]Artificial and Mechanical Intelligence, Italian Institute of Technology, Genoa, Italy. email: firstname.lastname@iit.it

[2]Department of Computer Science, The University of Manchester, Manchester, United Kingdom.

[3]Human-Robot Interfaces and Interaction, Italian Institute of Technology, Genoa, Italy. email: firstname.lastname@iit.it

physical laws, typically expressed as partial differential equations (PDEs), directly into the neural network architecture, PINNs enable more accurate and interpretable modeling of systems with inherent physical behaviors. Inspired by this, researchers have attempted to integrate Newtonian Dynamics Equations with different deep architectures to predict muscle forces and joint angles from surface electromyogram (sEMG) signals for biomechanical analysis [18]–[20]. However, these approaches are not directly applicable for robot-related scenarios due to limitations such as signal variability and constrained movement ranges.

To this end, we present a PINN-based model that leverages neural networks alongside physics equations to predict human whole-body kinematics using the inertial data from only 5 IMUs (see Figure 1). We represent the human model using the Unified Robot Description Format (URDF), a standard framework for defining rigid-body structures in robotics, which allows a structured definition of human motion in the configuration space. The forward and differential kinematics functions are integrated into the training to guide the network in learning the intrinsic relationships between joint and limb kinematics, reducing reliance on supervision labels and improving generalization. To enable smoother transitions between motions, we introduce a *joint state buffer* that incorporates predictions from the previous iteration as extra inputs for subsequent inference, allowing for recursive updates. In conclusion, the main contributions are:

- We propose a PI learning framework that incorporates a network architecture designed to account for the temporal and spatial characteristics of human movements.
- During training, we formulate forward and differential kinematics functions as additional losses, guiding the network to learn the underlying physical relationships that are not explicitly present in the training data.
- We introduce a joint state buffer to achieve smoother transitions between motions. At inference, the previous prediction is refined to update the buffer.
- We demonstrate the performance of the proposed method in both simulation and real-world experiments.

The remain of this paper is organized as follows. Section II introduces the most related work. Section III provides the details of our approach. Section IV presents three sets of experiments: 1) a general comparison with other state-of-the-art network architectures for time-series forecasting, 2) an ablation study that highlights the effects of each proposed feature, and 3) an evaluation of online performance. Finally, Section V concludes the limitations and possible future work.

## II. RELATED WORK

### A. Inertial Human Motion Capture

Inertial motion capture has been widely studied due to its advantages over optical-based systems, such as robustness to occlusions and unrestricted movement spaces. Commercial solutions [5], [6] employ dense IMU placement to achieve precise pose estimation, but this setup is intrusive and less practical for real-world applications. Prior research has attempted to reduce the number of required IMUs while maintaining accuracy. For instance, Marcard et al. used 6 IMUs for SMPL-based pose estimation [21], while Huang et al. applied a learning-based method to infer poses from sparse IMU data [7]. Yi et al. further introduced physics-based optimizations to improve motion plausibility [8], [9], and Jiang et al. utilized a transformer-based model to enhance motion tracking [10]. However, these methods primarily focus on pose reconstruction rather than predicting future movements. Furthermore, they define motions in task space rather than configuration space, overlooking joint kinematics constraints that are important for physically feasible predictions.

### B. Human Motion Prediction

Human motion prediction is commonly formulated as a sequence-to-sequence learning problem, where future poses are inferred from historical poses. Martinez et al. proposed a simplified yet effective recurrent neural network (RNN) model for motion prediction, demonstrating promising performance [22]. Darvish et al. introduced a guided-mixture-of-experts approach to simultaneously recognize action and predict whole-body movements [23]. Gui et al. explored how robots can learn to anticipate human motion by analyzing past movement patterns [12]. Meanwhile, Zhang et al. incorporated physics-based priors, leveraging estimated contact forces and joint torques to enhance motion prediction accuracy [13]. Beyond motion sequence models, some works integrate external sensor data to improve human motion forecasting. Yang et al. designed a recurrent model that estimates lower-body pose and foot contact states based on past upper-body signals [15]. Zhang et al. introduced a two-stage motion forecasting network that predicts future human poses from historical IMU readings [16]. However, these approaches often require extensive past pose data or complex multi-sensor fusion, which can limit their scalability to systems with a reduced set of sensors.

### C. PINN for Human Dynamics Modeling

Physics-Informed Neural Network was first proposed in [17] for solving supervised learning tasks that involve nonlinear partial differential equations. More recently, Zhang et al. expanded on this approach, presenting a physics-informed learning framework based on a convolutional neural network to predict synchronous muscle forces and joint kinematics from surface electromyogram (sEMG) signals [18]. Similarly, Shi et al. presented a physics-informed low-shot learning method for sEMG-based estimation of muscle forces and joint kinematics [20]. They incorporated Lagrange's equations of motion and an inverse dynamics model into a generative adversarial network to enable structured feature decoding from limited sample data. In another work, Ma et al. embedded Hill muscle forward dynamics into the deep neural network to predict muscle forces without any label information during training [19]. Nevertheless, current research is largely focused on sEMG-based methods, which are less practical for application in dynamic human-robot collaborative tasks.

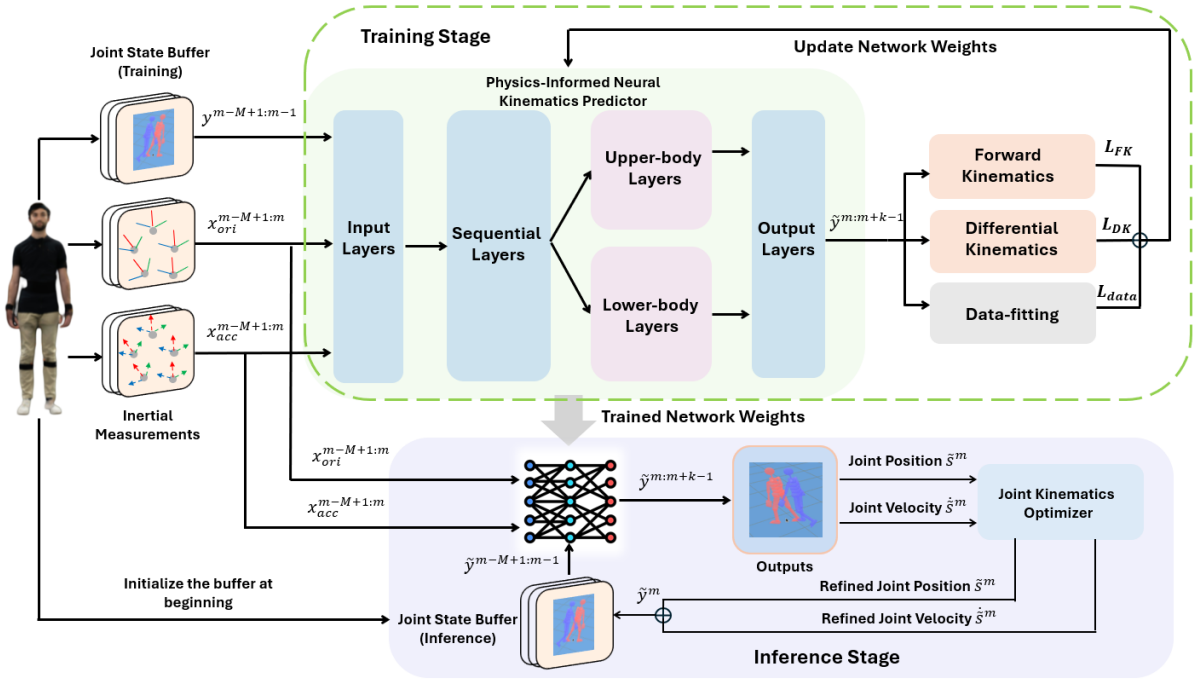



Fig. 2: Overview of the proposed Physics-Informed Neural Kinematics Predictor. The green dashed box describes the training process. The network architecture is demonstrated within the light green area, while the *physics-informed* and *data-fit* losses are presented by side. The purple area at bottom displays the inference process, where the trained network is deployed. The joint state buffer is initialized with ground truth.

## III. PROPOSED METHOD

Figure 2 exhibits an overview of the proposed method. Our goal is to predict future human motion in terms of whole-body joint kinematics, using a sequence of inertial readings from 5 IMUs mounted on the pelvis, both forearms, and both lower legs. Our novel approach, *Physics-Informed Neural Kinematics Predictor* (PINKP), is trained in a supervised fashion by incorporating relevant physics knowledge alongside the data-fitting process. To further address the ambiguity arising from sparse IMU data, a joint state buffer is utilized as an additional input. The joint kinematics optimizer refines the previous first-step prediction to update the buffer, enabling a closed-loop autoregressive process.

### A. Human Kinematics Modeling

The human is modeled as a floating-base multi-rigid-body dynamic system composed of $n+1$ links and connected by $n$ joints [24], whose configuration space lies on a Lie group composition $\mathbb{Q} = \mathbb{R}^3 \times \mathrm{SO}(3) \times \mathbb{R}^n$. An element from this space is represented as $\mathbf{q} = ({}^{\mathcal{I}}\mathbf{p}_{\mathcal{B}}, {}^{\mathcal{I}}\mathbf{R}_{\mathcal{B}}, \mathbf{s})$, where ${}^{\mathcal{I}}\mathbf{H}_{\mathcal{B}} = ({}^{\mathcal{I}}\mathbf{p}_{\mathcal{B}}, {}^{\mathcal{I}}\mathbf{R}_{\mathcal{B}})$ indicates the pose of the base frame $\mathcal{B}$ w.r.t. the Inertial frame $\mathcal{I}$ and $\mathbf{s} \in \mathbb{R}^n$ being the joint positions that represent the system topology. The system velocity belongs to the set $\mathbb{V} = \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^n$, an element of $\mathbb{V}$ is given by $\boldsymbol{\nu} = ({}^{\mathcal{I}}\dot{\mathbf{p}}_{\mathcal{B}}, {}^{\mathcal{I}}\boldsymbol{\omega}_{\mathcal{B}}, \dot{\mathbf{s}})$, where $\mathbf{v}_{\mathcal{B}} = ({}^{\mathcal{I}}\dot{\mathbf{p}}_{\mathcal{B}}, {}^{\mathcal{I}}\boldsymbol{\omega}_{\mathcal{B}})$ being the twist of base frame w.r.t. $\mathcal{I}$ and $\dot{\mathbf{s}} \in \mathbb{R}^n$ being the joint velocities. The *forward kinematics* (FK) function $\boldsymbol{\mathcal{F}}_i$ for a frame attached to link $i$ describes the mapping from system configuration $\mathbf{q}$ to link pose ${}^{\mathcal{I}}\mathbf{H}_i$:

$$\boldsymbol{\mathcal{F}}_i : \mathbb{Q} \to \mathrm{SE}(3), \mathbf{q} \mapsto {}^{\mathcal{I}}\mathbf{H}_i = \boldsymbol{\mathcal{F}}_i(\mathbf{q}), \tag{1}$$

we denote the operators for retrieving the link position ${}^{\mathcal{I}}\mathbf{p}_i$ and orientation ${}^{\mathcal{I}}\mathbf{R}_i$ as $\boldsymbol{\mathcal{F}}_i^p$ and $\boldsymbol{\mathcal{F}}_i^o$, respectively.

The velocity of a frame attached to link $i$ is denoted as $\mathbf{v}_i = ({}^{\mathcal{I}}\dot{\mathbf{p}}_i, {}^{\mathcal{I}}\boldsymbol{\omega}_i)$. The relationship between the $i$-th link frame velocity $\mathbf{v}_i$ and system configuration $(\mathbf{q}, \boldsymbol{\nu})$ is described by the *differential kinematics* (DK) function $\boldsymbol{\mathcal{G}}_i$:

$$\boldsymbol{\mathcal{G}}_i : \mathbb{Q} \times \mathbb{V} \to \mathbb{R}^3 \times \mathbb{R}^3, \mathbf{q}, \boldsymbol{\nu} \mapsto \mathbf{v}_i = J_i(\mathbf{q})\boldsymbol{\nu}, \tag{2}$$

where $J_i(\mathbf{q})$ is the $i$-th link *Jacobian*.

### B. Problem Formulation

Given a sequence of past inertial measurements $X^{m-M+1:m} = [X^{m-M+1}, \ldots, X^{m-1}, X^m] \in \mathbb{R}^{M \times N \times F}$ (where $m$ is the current timestamp, $M$ is the input sequence length, $N$ is the number of IMUs, $F$ is the dimension of input features of each IMU), we want to find a mapping function that can continuously predict the future joint configurations, namely $\tilde{y}^{m:m+K-1} = f(\theta; X^{m-M+1:m}) \in \mathbb{R}^{K \times 2 \times n}$ (where $K$ is the prediction horizon, $n$ denotes the dimension of joint degree of freedoms, $\theta$ indicates the hyperparameters of the mapping function). More specifically, a single step inertial measurements $X^i = [x^i_{acc}, x^i_{ori}] \in \mathbb{R}^{5 \times (3+9)}$ contains the acceleration and flattened rotation matrix readings from all 5 IMUs, while a single step prediction $\tilde{y}^i = [\tilde{\mathbf{s}}^i, \dot{\tilde{\mathbf{s}}}^i] \in \mathbb{R}^{2 \times n}$ includes both joint position and velocity.

### C. Physics-Informed Neural Kinematics Predictor

As illustrated in Figure 2, we use a time window of length *M* to continuously collect inertial measurements of accelerations and orientations from all 5 selected links as inputs. However, a key challenge of using sparse IMUs is distinguishing between ambiguous poses when IMU readings are very similar. In this study, we assume that human movements are consistent and steady. Consequently, the prediction of future kinematics state $\tilde{y}^{m:m+K-1}$ is conditioned on both sensor measurements $X^{m-M+1:m}$ and previously estimated internal states $\tilde{y}^{m-M+1:m-1}$, following the formulation $\max P(\tilde{y}^{m:m+K-1}|\tilde{y}^{m-M+1:m-1}, X^{m-M+1:m})$. Therefore, we introduce a *joint state buffer* of length $(M-1)$ to retain historical joint kinematics, which serve as additional input to the network. During training, the joint history kinematics are obtained from the ground truth data. During inference, the initial joint configurations $y^{0:M-2}$ are provided as known values, after which the buffer is updated recursively using the previous prediction, following a *first-in-first-out* scheme.

*1) Network Architecture:* The sequential inputs of inertial measurements $X^{m-M+1:m}$ and joint history states $\tilde{y}^{m-M+1:m-1}$ are initially processed through their respective fully connected layers utilizing the exponential linear unit (*elu*) activation function to extract hidden temporal features. The resulting outputs are then concatenated and passed through two additional fully connected layers, also employing the *elu* activation function. Recognizing that the dynamics of upper body and lower body motions can largely be considered independent, two separate blocks with identical structures are parallelly inserted to forecast the motion of the upper and lower bodies, respectively. Finally, the outputs from these two blocks are concatenated to facilitate the simulation of whole-body motions.

*2) Loss Function Design:* To train the network, we proposed a loss function comprising two main components, namely, *data-fit* loss and *physics-informed* loss. The *data-fit* loss $L_{\text{data}}$ is formulated as the Mean Squared Error (MSE) between the predicted and the target values:

$$\begin{aligned} L_{\text{data}} &= L_{\text{pos}} + L_{\text{vel}} \\ &= \frac{1}{2K}\sum_{t=m}^{m+K-1}(\|\tilde{\boldsymbol{s}}^t - \boldsymbol{s}^t\|_2^2 + \|\dot{\tilde{\boldsymbol{s}}}^t - \dot{\boldsymbol{s}}^t\|_2^2), \end{aligned} \tag{3}$$

where $m$ is the current timestamp, $\boldsymbol{s}$ and $\dot{\boldsymbol{s}}$ are the targets.

Moreover, the predicted joint configuration should also comply with the physical constraints imposed by Equations (1) and (2), which compose the *physics-informed* loss $L_{\text{phy}}$:

$$L_{\text{phy}} = L_{\text{FK}} + L_{\text{DK}}, \tag{4a}$$

$$L_{\text{FK}} = \frac{1}{DK}\sum_{i=1}^{D}\sum_{t=m}^{m+K-1}\Big(\|\boldsymbol{\mathcal{F}}_i^p(\tilde{\mathbf{q}}^t) - \mathbf{p}_i^t\|_2^2 + \|\boldsymbol{\mathcal{F}}_i^o(\tilde{\mathbf{q}}^t) \ominus \mathbf{R}_i^t\|_2^2\Big), \tag{4b}$$

$$L_{\text{DK}} = \frac{1}{DK}\sum_{i=1}^{D}\sum_{t=m}^{m+K-1}\|\boldsymbol{\mathcal{G}}_i(\tilde{\mathbf{q}}^t, \tilde{\boldsymbol{\nu}}^t) - \mathbf{v}_i^t\|_2^2 \tag{4c}$$

where *D* is the number of selected links, *K* is the prediction steps, *m* is the current timestamp. $\tilde{\mathbf{q}}^t$ consists of predicted joint position $\tilde{\boldsymbol{s}}^t$, and ground truth of base pose $(\mathbf{p}_{\mathcal{B}}^t, \mathbf{R}_{\mathcal{B}}^t)$, while $\tilde{\boldsymbol{\nu}}^t$ includes predicted joint velocity $\dot{\tilde{\boldsymbol{s}}}^t$, and ground truth of base twist $(\dot{\mathbf{p}}_{\mathcal{B}}^t, \boldsymbol{\omega}_{\mathcal{B}}^t)$. The values of $\tilde{\mathbf{q}}^t$ and $\tilde{\boldsymbol{\nu}}^t$ are utilized to obtain the estimated link pose and velocity at timestamp *t*. In the meanwhile, $\mathbf{p}_i^t$, $\mathbf{R}_i^t$ and $\mathbf{v}_i^t$ denote the reference position, orientation and velocity of *i*-th link frame at the same timestamp. The symbol $\ominus$ is for calculating the distances between the orientations, which are represented as flattened rotation matrix.

The total loss involved to update the network is a weighted sum of each *data-fit* and *physics-informed* individual:

$$L_{\text{total}} = \lambda_1 L_{\text{pos}} + \lambda_2 L_{\text{vel}} + \lambda_3 L_{\text{FK}} + \lambda_4 L_{\text{DK}}, \tag{5}$$

where $\lambda_i$, $i \in [1, 4]$ denote hyperparameters that are manually tuned to scale the significance of each loss component at the training stage.

### D. Joint Kinematics Optimizer

At the inference stage, the prediction errors persist and can accumulate due to the long-term drift and sensor noise. To address this, we introduce a joint kinematics optimizer that refines the network's prediction at timestamp *m*, which is then used to update the *joint state buffer*. The refinement process is carried out by solving the following optimization problem:

$$\min_{\mathbf{s}^*, \dot{\mathbf{s}}^*} \quad \|\boldsymbol{s}^* - \tilde{\boldsymbol{s}}^t\|_2^2 + \|\dot{\boldsymbol{s}}^* - \dot{\tilde{\boldsymbol{s}}}^t\|_2^2, \tag{6}$$

$$\text{s.t.} \quad \|\boldsymbol{\mathcal{G}}_i(\boldsymbol{s}^*, \mathbf{H}_{\mathcal{B}}^t, \dot{\boldsymbol{s}}^*, \mathbf{v}_{\mathcal{B}}^t) - \boldsymbol{v}_i^t\|_2^2 \leq \varepsilon, \quad i = 1, \ldots, D \tag{7}$$

where $t = m$ indicates the current timestamp. Note that we decompose the original arguments $(\mathbf{q}^t, \boldsymbol{\nu}^t)$ for function $\boldsymbol{\mathcal{G}}_i$ as $(\boldsymbol{s}^*, \mathbf{H}_{\mathcal{B}}^t, \dot{\boldsymbol{s}}^*, \mathbf{v}_{\mathcal{B}}^t)$ in the constraint. Our goal is to find a pair of $\mathbf{s}^*, \dot{\mathbf{s}}^*$ that corrects potential prediction drift by enforcing consistency with the *differential kinematics* equation for each selected link *i*, while staying close to the initial guess provided by the network at timestamp *m*. Importantly, only this single-step prediction is refined and stored in the *joint state buffer*, ensuring that future predictions from step $m+1$ to $m+K-1$ are implicitly adjusted over time. We opted not to include *forward kinematics* as a constraint because it serves a similar role to *differential kinematics* but only constrains joint position predictions, potentially introducing unnecessary computational overhead.

## IV. EXPERIMENTS AND RESULTS

### A. Experimental Setup

**Dataset.** To the best of our knowledge, there is no publicly available dataset that includes both inertial measurements and whole-body human joint kinematics. To address this, we leverage our CLIK-based framework [1], which processes data from an Xsens [5] wearable sensing system equipped with 17 wireless IMUs distributed across the entire body.

[1] $https://github.com/ami_iit/bipedal_locomotion_framework$

TABLE I: Description of human motion dataset.

| **Motion Type** | Forward Walking | Forward Walking | Backward Walking | Side Stepping | Forward Walking Lifting Arms | Forward Walking Waving Arms | Forward Walking Clapping Hands |
|---|---|---|---|---|---|---|---|
| **Speed** | Normal | Fast | Normal | Normal | Normal | Normal | Normal |
| **Duration [min]** | 8.0 | 7.9 | 8.1 | 7.9 | 7.9 | 7.9 | 7.9 |
| **Frames** | 28798 | 28332 | 29117 | 28440 | 29401 | 28508 | 28505 |

Our self-collected dataset captures a diverse range of whole-body motions performed at varying speeds on flat terrain, with continuously changing facing direction. Each recorded sequence includes stops and restarts. As summarized in Table I, each motion lasts approximately 8 min, resulting in a final dataset of about 1 hour, sampled at 60 Hz.

**Metrics.** We define two metrics for the general evaluation of models' predictions, namely, *Mean Absolute Joint Position Error* (pMAE) and *Mean Absolute Joint Velocity Error* (vMAE) which measure the average absolute error in the predicted joint positions/velocities for selected Degrees of Freedom (DoFs). To further distinguish the effects of the physics-informed component, we utilize three more metrics: *Root Mean Squared Link Position Error* (pRMSE), *Root Mean Squared Link Orientation Error* (oRMSE) and *Root Mean Squared Link Velocity Error* (vRMSE) that measure the root mean squared error of retrieved link position/orientation/velocity using the joint position predictions.

**Implementation Details.** The training and testing are conducted on a laptop with an Intel(R) Core(TM) i7-10750H CPU and an NVIDIA GeForce GTX 1650 Ti GPU. We use PyTorch with CUDA 12.2 to implement the kinematics prediction network and leverage the Automatic Differentiation Rigid-Body-Dynamics Algorithms Library [2] to enable the differentiation of inserted physics-informed quantities during backpropagation. The network is trained for 30 epochs with a batch size of 256 using the Adam optimizer. The learning rate is initialized as 1e-3 and is updated by the StepLR scheduler every 5 epochs. The inertial sequence length that is fed to the network is equal to 10 steps, while the prediction horizon is set as 60 steps.

### B. Comparison Studies

We conduct a comparison of our method with several leading sequence forecasting networks: LSTM [25] (a canonical recurrent architecture), TCN [26] (a temporal convolutional network for sequence modeling), and TIP [10] (a Transformer-based approach for human motion reconstruction). This comparison is performed across all tasks outlined in Table I. For comparative analysis, the inputs and outputs of the baseline networks have been adjusted to match our model, as depicted in Figure 2.

**Quantitative.** In Table II, we present a quantitative comparison of our method against baselines across various activities, each evaluated at different prediction timestamps (i.e., t=1, 30, 60). For the *locomotion tasks* (first four activities), metrics are computed as the average errors of lower-body joints, while for the *whole-body tasks* (last three activities), the analysis considers full-body joints. Our method consistently achieves lower mean absolute errors in predicting both joint positions and velocities across most activities and prediction timestamps compared to LSTM, TCN, and TIP, indicating superior predictive accuracy on joint position and velocity. Additionally, as shown in Table III, our method requires on average less than 1 ms for each iteration inference, making it well-suited for real-time applications.

[2] $https://github.com/ami_iit/adam$

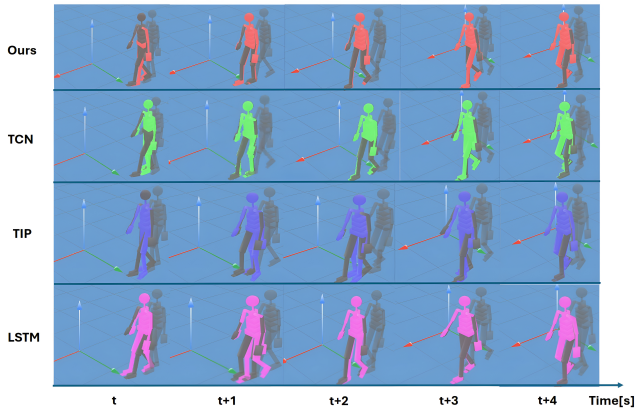



Fig. 3: Qualitative results of different methods on the task *forward walking with normal speed*. Screenshots are captured at 1 second intervals.

We attribute this superior performance to the integration of physical priors at multiple levels. Therefore the proposed approach enhance the accuracy of joint position and velocity prediction along with smoothness of transitions between motions. The physical constraints applied during training guide the prediction of joint states, which results in a less noisy initial guess for the joint state buffer. Then, during inference, the joint kinematics optimizer leverages inertial measurements from the selected links to further reduce the potential accumulation of prediction errors. The combination of these techniques yields more accurate predictions and smoother motion transitions.

**Qualitative.** We further illustrate the qualitative comparison results in Figure 3. The presented frames are captured from a sequence where the subject walks forward at an average pace. In each subplot, the semi-transparent gray avatar denotes the ground truth at the present timestamp. Meanwhile, the solid gray and colored avatars represent the ground truth and the prediction, projected 30 steps ahead, equivalent to 0.5 seconds into the future. As shown in the first column, our method can predict the subject's transition from a *standing* pose to a *walking* pose with mitigated motion delay thanks to the joint state buffer updated with refined prediction. In contrast, the baselines have difficulty precisely reconstructing the motion transition. Moreover, in the subsequent frames, our method demonstrates enhanced overlap with the reference, indicating higher predictive accuracy.

### C. Ablation Studies

**Physics-Informed Component.** In this section, we want to understand the impact of *forward kinematics* (FK) and *differential kinematics* (DK) on prediction performance. To

TABLE II: Quantitative comparison with human kinematics prediction baselines.

| Prediction Timestamp | Method | Forward Walking Normal | | Forward Walking Fast | | Backward Walking Normal | | Side-stepping Normal | | Forward Walking Lifting Arms | | Forward Walking Waving Arms | | Forward Walking Clapping Hands | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pMAE (deg) | vMAE (deg/s) | pMAE (deg) | vMAE (deg/s) | pMAE (deg) | vMAE (deg/s) | pMAE (deg) | vMAE (deg/s) | pMAE (deg) | vMAE (deg/s) | pMAE (deg) | vMAE (deg/s) | pMAE (deg) | vMAE (deg/s) |
| t=1 | LSTM | 6.25 | 45.28 | 5.14 | 41.40 | 3.49 | 23.49 | 3.99 | 25.47 | 6.01 | 38.25 | 6.49 | 37.26 | 4.20 | 29.41 |
| | TIP | 5.38 | 41.85 | 6.47 | 54.90 | 3.99 | 30.91 | 3.24 | 28.25 | 7.26 | 53.04 | 7.22 | 54.39 | 9.07 | 73.87 |
| | TCN | 3.76 | 27.19 | **2.72** | 18.77 | 3.02 | 22.38 | 3.46 | 25.82 | 6.34 | 34.34 | 6.48 | 29.02 | 4.51 | 26.74 |
| | Ours | **2.11** | **12.58** | 3.80 | **15.18** | **2.20** | **14.09** | **2.41** | **19.93** | **5.24** | **28.35** | **2.73** | **14.83** | **3.53** | **17.93** |
| t=30 | LSTM | 6.22 | 45.52 | 4.77 | 36.87 | 3.41 | 22.56 | 4.01 | 25.32 | 6.56 | 35.89 | 6.89 | 37.73 | 4.57 | 31.75 |
| | TIP | 5.46 | 39.25 | 6.25 | 52.69 | 4.08 | 29.22 | 3.33 | 28.57 | 8.33 | 47.10 | 8.44 | 43.34 | 8.43 | 79.77 |
| | TCN | 3.84 | 27.42 | 2.77 | 19.29 | 3.11 | 21.42 | 3.57 | 27.05 | 6.47 | 36.45 | 6.56 | 29.91 | 4.76 | 30.19 |
| | Ours | **2.27** | **14.89** | **2.33** | **15.49** | **2.34** | **15.21** | **2.89** | **22.18** | **6.30** | **29.48** | **2.96** | **16.93** | **3.77** | **22.25** |
| t=60 | LSTM | 6.05 | 45.60 | 4.98 | 35.68 | 3.40 | 22.74 | 4.03 | 25.13 | 7.48 | 28.38 | 7.88 | 31.14 | 5.04 | 33.95 |
| | TIP | 5.26 | 38.95 | 6.12 | 46.38 | 4.07 | 28.43 | 3.52 | 28.80 | 9.16 | 38.70 | 9.35 | 44.43 | 9.47 | 61.57 |
| | TCN | 3.91 | 28.27 | 3.07 | 21.79 | 3.16 | 20.74 | 3.59 | 26.77 | **7.01** | 37.31 | 6.59 | 26.22 | 5.45 | 34.17 |
| | Ours | **2.63** | **18.47** | **2.71** | **18.68** | **2.60** | **17.84** | **3.17** | **24.16** | 7.22 | **27.34** | **3.60** | **19.44** | **4.29** | **28.05** |

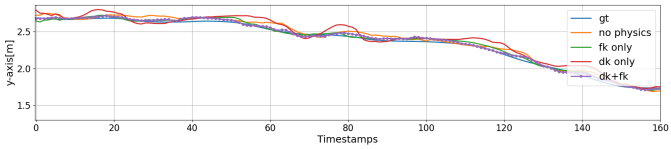
TABLE III: Inference time (ms) per iteration of different methods on various tasks.

| Method | Forward Walking Normal | Forward Walking Fast | Backward Walking Normal | Side stepping Normal | Forward Walking Lifting Arms | Forward Walking Waving Arms | Forward Walking Clapping Hands |
|---|---|---|---|---|---|---|---|
| LSTM | **0.84** | **0.89** | **0.84** | **0.88** | **0.86** | **0.83** | **0.92** |
| TIP | 7.24 | 7.34 | 7.23 | 7.48 | 7.50 | 7.44 | 7.69 |
| TCN | 2.16 | 2.28 | 2.16 | 2.23 | 2.18 | 2.11 | 2.35 |
| Ours | **0.86** | **0.95** | **0.89** | **0.94** | **1.11** | **1.12** | **1.22** |

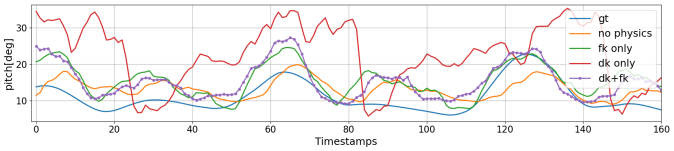
isolate their effects, we disabled the joint state buffer, eliminating potential confounding influences from other modules. We trained four models, each with different physical loss components enabled: (1) no kinematics, (2) forward kinematics only, (3) differential kinematics only, and (4) both forward and differential kinematics.

The four models were evaluated on a sequence of the *Side Stepping* task. By leveraging the Equations (1) and (2), we retrieved the poses and velocities of the desired link (e.g., left lower leg) using the first step of the sequences predicted by the ablation models. The link position along the y-axis and the pitch values are illustrated in Figure 4, while Table IV presents more detailed numerical results. We can see that different formulations of physical constraints influence the model's predictive performance in distinct ways. For instance, the model with only DK loss activated demonstrated strong accuracy in predicting link twist but performed worse in retrieving link poses compared to other models. In contrast, the model utilizing both FK and DK losses achieved the highest accuracy in terms of link positions and velocities but exhibited slightly reduced performance in link orientation compared to the baseline model that doesn't include the kinematics information. We attribute this to the influence of the injected PI components, which incorporate physical principles that are not explicitly represented in the training data. By embedding the fundamental laws into the network, the PI components help the model generalize better, especially in scenarios with limited data.

**Joint State Buffer.** In this section, we examine the impact of the joint state buffer. We deployed three models,




(a) The retrieved link position along the y-axis.



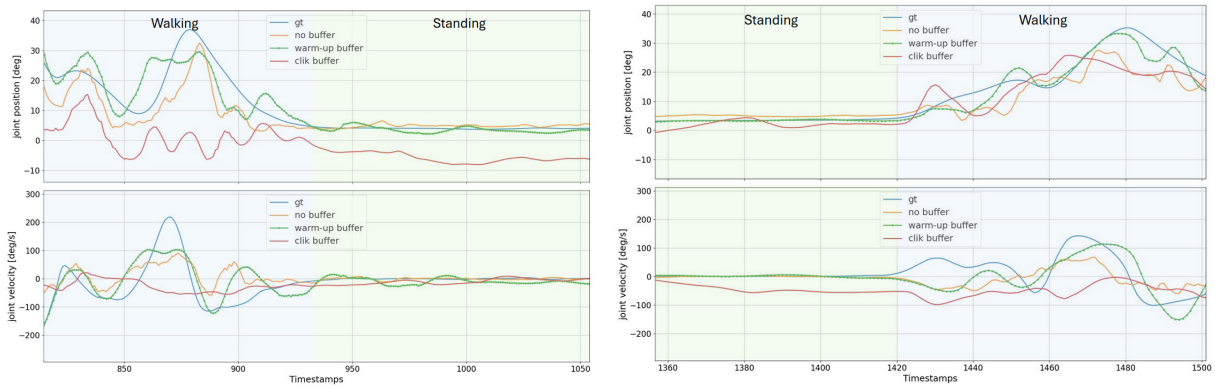

(b) The retrieved pitch values (flexion/extension) of link orientation.

Fig. 4: A sequence of *side-stepping* trajectory. The poses of the left lower leg are retrieved using first-step predictions from different models. The base is assumed to be given.

TABLE IV: Ablation study on the impact of different physics-informed formulations.

| Ablation Models | pRMSE [m] | oRMSE [deg] | vRMSE | |
|---|---|---|---|---|
| | | | Linear [m/s] | Angular [deg/s] |
| no kinematics | **0.0217** | **4.34** | 0.195 | 38.469 |
| forward kinematics only | 0.0252 | 4.80 | 0.168 | 36.936 |
| differential kinematics only | 0.0236 | 8.86 | **0.122** | **31.363** |
| both forward and differential kinematics | **0.0195** | **4.50** | **0.118** | **34.637** |

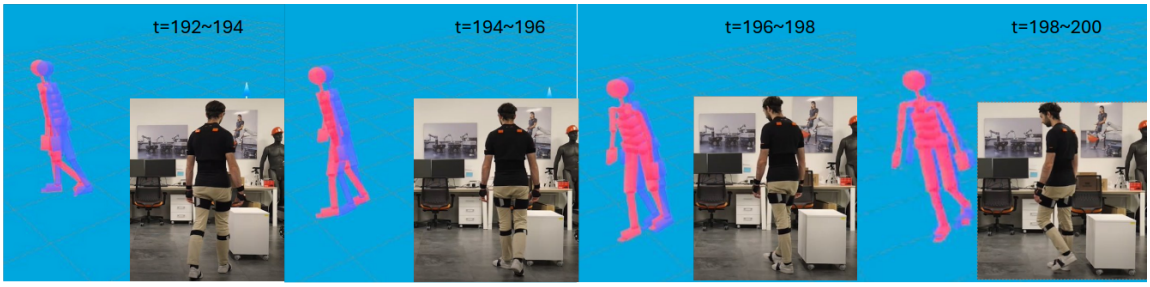
namely, (1) *no buffer*: using only inertial inputs, (2) *CLIK buffer*: incorporating both inertial inputs and historical states, with the state buffer updated via a CLIK-based IK module, that is, given inertial measurements of 5 IMUs, obtain the one-step joint configuration via a closed-loop iterative algorithm, and (3) *warm-up buffer*: incorporating both inertial inputs and historical states, with the state buffer updated using the joint kinematics optimizer. We tested the three ablation models on a sequence of *Side Stepping* trajectory, which includes transitions between *standing* and *walking* modes. Figure 5 illustrates the positions and velocities of the *Left Knee* joint around the y-axis, capturing its flexion-extension movement. The third model, which utilizes the



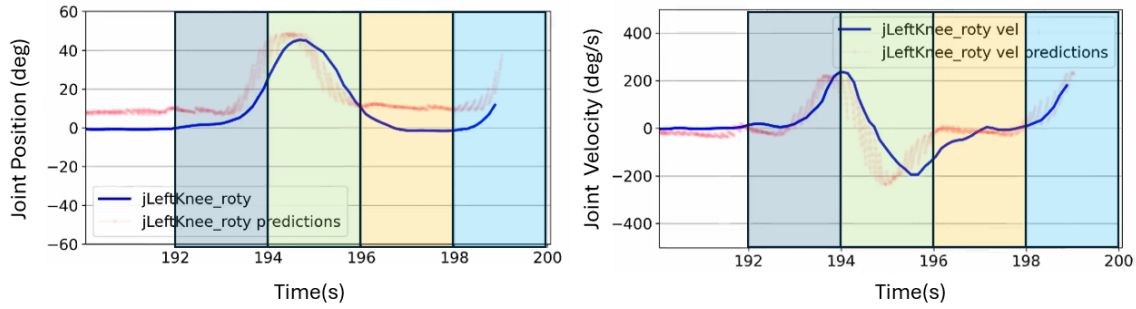

(a) The transition from *walking* to *standing*.

(b) The transition from *standing* to *walking*.

Fig. 5: A sequence of *side-stepping* trajectory. The position and velocity of joint *left knee* around the y-axis are compared since it represents the flexion and extension movement during walking. The blue line is the ground truth. The orange, red, and green lines indicate the first-step predictions of models 1 to 3, respectively.




(a) Recording of the volunteer walking around. The time displayer at the top (e.g., t=192 $\sim$ 194 sec) represents the interval during which the motion occurs. The analysis starts at 190 sec because the subject remains standing for a while at the beginning of each experiment. The blue avatar is the ground truth at the current timestamp, and the red avatar is the prediction at 10 timestamps ahead, sampled at 60 Hz.




(b) (Left) Joint position of left knee rotation around y-axis. (Right) Joint velocity of left knee rotation around y-axis. The blue line denotes the ground truth, while the red dotted line denotes the 10 timestamps prediction in the future.

Fig. 6: Online experimental results of predicting human joint kinematics via the inertial readings of a sparse set of IMUs.

joint kinematics optimizer to update the buffer, showcased a much smoother transition between *standing* and *walking* poses compared to the other models. Additionally, it also provided more accurate predictions for both joint position and velocity when the subject remains stationary. The results of the second model, which leverages a CLIK-based module to update the buffer, suggests that updating the buffer solely through iterative IK solutions is suboptimal. Given that only five IMUs provide inertial measurements, the computed IK-based updates can be less reliable than directly refining the first-step prediction from the network.

### D. Online Performance Evaluation

We evaluate our method using 17 Xsens IMUs for calibration and video recording, while processing only inertial data from five IMUs on the pelvis, forearms, and lower legs. As depicted in Figure 6, the subject begins with the right foot in contact, lifting the left foot at t = 193 s. The left knee's velocity (y-axis) peaked first, followed by its maximum position. At t = 196.5 s, the right foot regained contact, and the left knee's motion returned to zero. A new walking cycle

started at t = 198 s. As illustrated in the third row, our method accurately captures the walking patterns, enabling precise predictions. Additionally, our method demonstrates its ability to generalize to the kinematics of unseen subjects.

### *E. Limitation Discussion*

Despite the promising performance in demonstrations, our work has several limitations. First, while our dataset includes diverse activities, subject diversity could be further improved. Second, our method assumes smooth and consistent motions, limiting its ability to handle discontinuous movements or rapid external perturbations. Lastly, relying on a predefined floating base (i.e., pose and velocity) constrains broader real-world applications.

## V. CONCLUSIONS

In this work, we present a physics-informed learning architecture to predict human kinematics using only 5 wearable IMUs. First, we represent the human motions in a constrained joint configuration space rather than the task space (i.e., limb poses). Then, we propose a network architecture that accounts for the temporal and spatial characteristics of human motion by separately modeling the upper and lower bodies within a unified framework. By incorporating forward and differential kinematics as additional components alongside data-fit loss during training, the network is guided to learn the underlying relations between joints and limbs, which are not explicitly present in the training data. At last, a joint kinematics optimizer is deployed at the inference stage to further enhance motion smoothness by updating a state buffer with a refined prediction.

In contrast to prior PINN-based approaches in biomechanics, our approach enables accurate, efficient full-body kinematics prediction with minimal sensors, eliminating the need for complex setups and fusion. Concerning the current limitations, our future work will first focus on tracking the pose and velocity of the floating base. Expanding the dataset to include a more diverse range of subjects will also enhance our model generalization.

## REFERENCES


[1] A. Ajoudani, A. Zanchettin, S. Ivaldi, G. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of human–robot collaboration," *Autonomous Robots*, vol. 42, pp. 957–975, 2018.

[2] Z. Niu, K. Lu, J. Xue, X. Qin, J. Wang, and L. Shao, "From method to application: A review of deep 3d human motion capture," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[3] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, and C. Theobalt, "Deepcap: Monocular human performance capture using weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5052–5063.

[4] Y. Zheng, R. Shao, Y. Zhang, T. Yu, Z. Zheng, Q. Dai, and Y. Liu, "Deepmulticap: Performance capture of multiple characters using sparse multiview cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6239–6249.

[5] Xsens. Xsens 3d motion tracking. [Online]. Available: https://www.movella.com/products/xsens

[6] Noitom. Perception neuron series. [Online]. Available: https://www.noitom.com/

[7] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018.

[8] X. Yi, Y. Zhou, and F. Xu, "Transpose: Real-time 3d human translation and pose estimation with six inertial sensors," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[9] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu, "Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] Y. Jiang, Y. Ye, D. Gopinath, J. Won, A. W. Winkler, and C. K. Liu, "Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation," in *SIGGRAPH Asia 2022 Conference Papers*, 2022.

[11] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2015.

[12] L.-Y. Gui, K. Zhang, Y.-X. Wang, X. Liang, J. M. Moura, and M. Veloso, "Teaching robots to predict human motion," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 562–567.

[13] Z. Zhang, Y. Zhu, R. Rai, and D. Doermann, "Pimnet: Physics-infused neural network for human motion prediction," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8949–8955, 2022.

[14] Y. Katsuhara and H. Kaji, "Towards multi-person motion forecasting: Imu based motion capture approach," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 2019.

[15] D. Yang, D. Kim, and S.-H. Lee, "Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals," *Computer Graphics Forum*, vol. 40, no. 2, 2021.

[16] X. Zhang, H. Zhang, J. Hu, J. Deng, and Y. Wang, "Motion forecasting network (mofcnet): Imu-based human motion forecasting for hip assistive exoskeleton," *IEEE Robotics and Automation Letters*, 2023.

[17] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[18] J. Zhang, Y. Zhao, F. Shone, Z. Li, A. F. Frangi, S. Xie, and Z. Zhang, "Physics-informed deep learning for musculoskeletal modeling: Predicting muscle forces and joint kinematics from surface emg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 484–493, 2022.

[19] S. Ma, J. Zhang, C. Shi, P. Di, I. D. Robertson, and Z. Zhang, "Physics-informed deep learning for muscle force prediction with unlabeled semg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024.

[20] Y. Shi, S. Ma, Y. Zhao, and Z. Zhang, "A physics-informed low-shot adversarial learning for semg-based estimation of muscle force and joint kinematics," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[21] T. Von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus," in *Computer graphics forum*, vol. 36, no. 2. Wiley Online Library, 2017, pp. 349–360.

[22] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2891–2900.

[23] K. Darvish, S. Ivaldi, and D. Pucci, "Simultaneous action recognition and human whole-body motion and dynamics prediction from wearable sensors," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 488–495.

[24] S. Traversaro, "Modelling, estimation and identification of humanoid robots dynamics," *Ph. D. dissertation*, 2017.

[25] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[26] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.