Incorporating Physics Principles for Precise Human Motion Prediction

Yufei Zhang[1], Jeffrey O. Kephart[2], Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]IBM Research
{zhangy76, jiq}@rpi.edu, kephart@us.ibm.com

Abstract

A variety of real-world applications rely on accurate predictions of 3D human motion from their past observations. While existing methods have made notable progress, their predictions over subsecond horizons can still be off by many centimeters. In this paper, we argue that achieving precise human motion prediction requires characterizing the fundamental physics principles governing body movements. We introduce ***PhysMoP****, a novel framework that incorporates* ***Phys****ics for human* ***Mo****tion* ***P****rediction. PhysMoP estimates the body configuration of the next frame by solving the Euler-Lagrange equations, a set of Ordinary Different Equations describing the physical motion rules. To limit the inherent problem of error accumulation over time, PhysMoP leverages a data-driven model and iteratively guides the physics-based prediction via a fusion model. Through extensive experiments, we demonstrate that PhysMoP significantly outperforms existing approaches at subsecond prediction horizons. For example, at a prediction horizon of 80 msec, PhysMoP outperforms traditional data-driven approaches by a factor of 10 or more.*

1. Introduction

A wide range of real-world applications such as autonomous driving [13], intelligent robotics [28], and animation [50], rely on accurate prediction of the 3D position and configuration of a human body based on observations of its past motion. Traditional approaches to human motion prediction employ statistical methods that include Gaussian process models [31, 59, 62] and Restricted Boltzmann Machines (RBMs) [7, 57]. These approaches incorporate specific assumptions about the distribution of motion data. Although these assumptions can represent simple body movements, accurately modeling complex motion patterns requires a more sophisticated characterization.

Over the last decade, as the availability of publicly accessible motion capture data has increased [21, 25, 43], researchers have made promising progress on existing benchmarks [12, 22, 41, 45, 46, 72] by applying deep learning to human motion prediction. Various deep models have been proposed to model motion sequences, such as those based on Recurrent Neural Networks (RNN) [10, 16, 19, 47], Convolutional Neural Networks (CNN) [4, 9, 32, 34, 56], and Transformers [3, 39, 48]. Additionally, variants of Generative Adversarial Networks (GAN) [6, 20, 26] and Variational Auto-encoder (VAE) [66] have been used to model motion data distributions [30, 53, 65, 69]. While these deep models have shown the capacity to capture complex motion patterns, they are purely data-driven, overlooking a crucial fact: human bodies are physical bodies and their movement adheres to physical motion rules. The central thesis of this paper is that, *by incorporating fundamental principles of physical motion into predictive models of human motion, the prediction accuracy can be improved significantly, especially over subsecond time horizons.*

The human body is an intricate physical system, where multiple interconnected body parts work together to facilitate intricate movements. To effectively model and analyze its dynamic behavior, one can employ the Euler-Lagrange equations [14], which capture the same basic physical principles as Newton's laws of motion but provide a more useful description of 3D body motion in a generalized coordinate system. Specifically, the generalized coordinate system can be chosen as needed, such as body joint angles; the Euler-Lagrange equations describe body movements through a set of second order Ordinary Differential Equations (ODEs) of the generalized position over time. For motion prediction, exploiting the Euler-Lagrange equations can enable the inference of additional physical information for a prediction model to improve its performance [71]. The Euler-Lagrange equations can also be utilized to derive physical constraints, the imposition of which can enhance the quality of human motion synthesis outputs [42, 63, 67] as well as monocular 3D human reconstruction results [17, 24, 33, 54]. Particularly, the physical artifacts often presented in data-driven estimates, such as unrealistic motion jittering, are no-

The code of this work is available at https://github.com/zhangy76/PhysMoP.

tably alleviated. While current methods have made promising progress by integrating physics, they have not explored the way to directly synergize physics and neural networks. Specifically, these approaches primarily employ the Euler-Lagrange equations to preprocess data for training deep models or to impose constraints for refining predictions generated by deep models. In contrast, our work aims to directly integrate physics into a motion prediction model, seamlessly merging physics and deep learning without separate incorporation processes.

Recently, physics-informed deep learning has demonstrated significant promises in model prediction accuracy, training speed, and generalization [27]. Unknown partial differential equations [51] and intricate Lagrangian dynamics [11] can be accurately and efficiently solved and captured through neural networks. Inspired by these approaches, we present a novel approach to incorporate the Euler-Lagrange equations into a human motion prediction model by directly specifying and solving the equations through neural networks. We demonstrate that the proposed approach, PhysMoP, can significantly improve the prediction accuracy over existing methods and achieve more physically plausible estimates.

In summary, the main contributions of our work are:

- We propose PhysMoP, a novel approach that effectively incorporates physics principles into predictive models of human motion. PhysMoP is built upon a physics-based motion prediction model that encodes the Euler-Lagrange equations by explicitly specifying and solving them to estimate future motion.
- PhysMoP further utilizes a data-driven model to effectively guide long-term predictions and mitigate error accumulation in the physics-based estimations through a fusion model.
- Through experiments, we demonstrate that PhysMoP significantly outperforms existing works, with particular advantages in short-term human motion prediction.

2. Related Work

In this section, we review existing literature on data-driven techniques for predicting human motion and then discuss adjacent prior work that has leveraged physics to improve models of human dynamics, which have been employed for motion prediction and other related tasks.

Data-Driven Human Motion Prediction. Early efforts that rely on data to model motion patterns included statistical approaches [7, 31, 59, 62]. They made various assumptions about data dependencies and representation, such as the fixed variable relationships and the binary-valued units introduced in RBMs [57]. Their applicability is limited to simple movements that exhibit consistent patterns over time. Recently, with the availability of a larger amount of motion data, deep learning methods have proven to be superior at modeling complex motion patterns. To effectively capture the temporal dependency in body movements, RNN and their variations [10, 16, 19, 47] have been proposed. These traditional RNN-based temporal models, however, are not effective in capturing spatial information. To remedy this problem, several authors have employed CNN [9, 32], while others have used Graph Convolutional Networks (GCN) [4, 12, 34, 35] whereby human body movement is modeled as a graph with nodes represented by body joints. Furthermore, to better capture the dependency in both spatial and temporal domains, others have considered spatio-temporal graph convolution [56, 72]. Recently, Guo *et al.* [22] show better performance by utilizing Multi-Layer Perceptrons (MLPs). Xu *et al.* [64] further leverage geometric equivariance in motion data to improve the accuracy of conventional deep models. Nonetheless, as will be elaborated in Sec. 4, these approaches can suffer inherently from their failure to take advantage of the physics principles governing human motion.

Physics-Based Human Dynamics Modeling and Motion Prediction. Modeling human dynamics with physics requires the model to characterize the equations governing physical motion. Such physics-based models have been employed in different tasks to improve purely data-driven estimates. In monocular 3D human body reconstruction, some methods have formulated an optimization problem to jointly estimate unknown physical parameters in the Euler-Lagrange equations and improve the data-driven estimates [17, 18, 36, 52, 55, 63]. Others adopt learned policies to simulate realistic movements from initial data-driven estimates [24, 33, 40, 54, 68]. Instead of hinging on data-driven estimates and strive to refine them, we integrate physics principles directly into a prediction model to generate accurate future motion prediction. Furthermore, our approach stands apart from existing human motion prediction works that have incorporated physics mainly at a data level. In detail, Maeda *et al.* [42] employed a learned policy to eliminate physical artifacts exhibited in synthesized motion and then harnessed the enhanced quality to train motion prediction models, thereby improving model performance. Zhang *et al.* [71] employed the Euler-Lagrange equations to infer motion forces from observed motion and then used these forces as additional inputs to a motion prediction model, thereby enhancing the model's performance. In this paper, we integrate physics at a model level. We propose a novel framework that directly injects physics into a learning-based motion prediction model end-to-end.

3. Proposed Method

An overview of our proposed approach, PhysMoP, is illustrated in Fig. 1. Below we first introduce the relevant

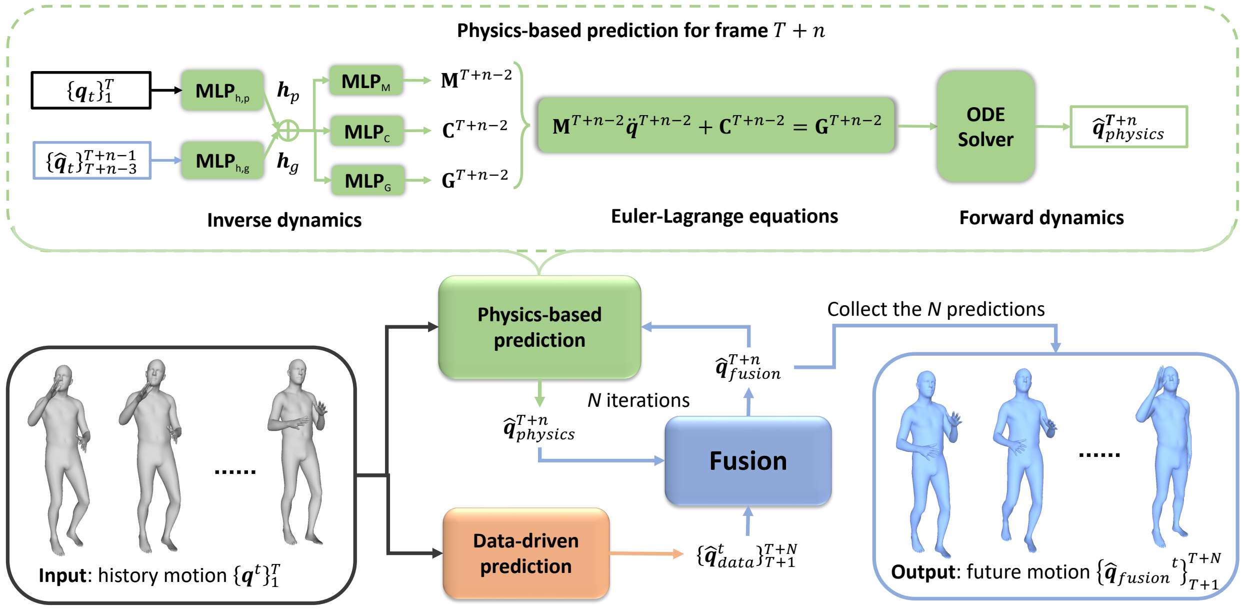
Physics-based prediction for frame $T+n$
$\{q_t\}_1^T$
$\text{MLP}_{h,p}$
h_p
$\{\hat{q}_t\}_{T+n-3}^{T+n-1}$
$\text{MLP}_{h,g}$
h_g
MLP_M
MLP_C
MLP_G
$\mathbf{M}^{T+n-2}$
$\mathbf{C}^{T+n-2}$
$\mathbf{G}^{T+n-2}$
$\mathbf{M}^{T+n-2}\ddot{q}^{T+n-2}+\mathbf{C}^{T+n-2}=\mathbf{G}^{T+n-2}$
ODE Solver
$\hat{q}_{physics}^{T+n}$
Inverse dynamics
Euler-Lagrange equations
Forward dynamics
Physics-based prediction
Collect the N predictions
$\hat{q}_{fusion}^{T+n}$
N iterations
$\hat{q}_{physics}^{T+n}$
Fusion
......
Data-driven prediction
$\{\hat{q}_{data}^t\}_{T+1}^{T+N}$
Input: history motion $\{q^t\}_1^T$
Output: future motion $\{\hat{q}_{fusion}{}^t\}_{T+1}^{T+N}$

Figure 1. **Overview of the proposed approach PhysMoP.** PhysMoP includes a physics-based model that effectively incorporates the Euler-Lagrange equations to estimate next frame's body configuration by capturing the forward and inverse dynamics process. Meanwhile, PhysMoP includes a data-driven branch to capture long-term dependency and a fusion model to leverage the data-driven prediction as guidance for the physics-based model to alleviate the problem of error accumulation.

physics knowledge characterized by PhysMoP in Sec. 3.1. The main component of PhysMoP is a physics-based motion prediction model that iteratively estimates future body configuration from input history motion and three previous estimates as will be elaborated in Sec. 3.2. Meanwhile, as will be introduced in Sec. 3.3, PhysMoP alleviates the error accumulated over time in the iterative prediction by including a data-driven model to capture long-term dependency and guide the physics-based motion prediction model through a fusion model. Lastly, in Sec. 3.4, we discuss the training and testing procedure of PhysMoP.

3.1. Physics Principles in Human Motion

In this section, we first elaborate the formulation of the Euler-Lagrange equations. Then, we introduce the forward and inverse dynamics process, by which those equations are used to model human dynamics.

The Euler-Lagrange Equations. The Euler-Lagrange equations are formulated in a generalized coordinate system, which consists of variables that fully specify the configuration of a physical system. In the context of modeling human motion, based on the successful human model SMPL [37], a human body configuration can be described using low-dimensional body pose and body shape parameters. Specifically, SMPL represents the human body through a 3D mesh model composed of 6890 vertices. The body pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{24\times 3}$ correspond to joint angles, defining the rotations of 23 body joints and a root rotation. On the other hand, the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ are coefficients for body shape bases, controlling variations in body attributes, such as width, height, and more. Given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the vertex and body joint positions of a 3D human can be obtained through forward kinematics. In the task of human motion prediction, the body shape of a subject remains unchanged. The motion trajectory in a world frame can be fully specified by the body pose parameters $\boldsymbol{\theta}$ along with the body translation parameters $\mathbf{T} \in \mathbb{R}^3$. Therefore, we define the generalized coordinate as:

$$\mathbf{q} = \{\boldsymbol{\theta}, \mathbf{T}\}, \tag{1}$$

where $\mathbf{q} \in \mathbb{R}^{75}$. During implementation, $\boldsymbol{\theta}$ represents the Euler angles of body joints. Unlike existing approaches that use a fixed rotation order to compute the Euler angles, we follow [70] and consider the biomechanically constrained joint angle ranges to determine the rotation order, thereby preventing duplicate angle solutions.

Given the generalized coordinate system defined above, we denote the generalized velocity and acceleration at frame t as $\dot{\mathbf{q}}^t \in \mathbb{R}^{75}$ and $\ddot{\mathbf{q}}^t \in \mathbb{R}^{75}$, respectively. The body dynamics, which are governed by the Euler-Lagrange equation, can be described as:

$$\mathbf{M}^t\ddot{\mathbf{q}}^t + \mathbf{C}^t = \mathbf{G}^t, \tag{2}$$

where $\mathbf{M}^t \in \mathbb{R}^{75\times 75}$ represents the generalized inertia matrix, which is determined by the generalized position $\mathbf{q}^t$ and the physical parameters including body mass and inertia. $\mathbf{C}^t \in \mathbb{R}^{75}$ represents the generalized bias force, including Coriolis, centrifugal, and gravitational forces. $\mathbf{C}^t$ is dependent on the generalized position, velocity, and physical properties of the human body. Lastly, $\mathbf{G}^t \in \mathbb{R}^{75}$ represents the generalized forces, which include both external forces (e.g., ground reaction forces) and internal forces (e.g., joint actuations that drive the rotation of different body joints).

Forward and Inverse Dynamics. When employing the Euler-Lagrange equations to model and analyze human motion, two essential processes are involved: forward dynamics and inverse dynamics. Forward dynamics focuses on solving the Euler-Lagrange equations to predict the next frame's 3D body configuration, given the physical parameters $\mathbf{M}^t$, $\mathbf{C}^t$, and the forces $\mathbf{G}^t$. Specifically, starting from the fully specified Eq. 2, we first solve for $\ddot{\mathbf{q}}$, which is then utilized to determine future motion using Euler's Method, expressed as follows:

$$\begin{aligned}\ddot{\mathbf{q}}^t &= \mathbf{M}^{t^{-1}}(\mathbf{G}^t - \mathbf{C}^t)\\ \dot{\mathbf{q}}^{t+1} &= \dot{\mathbf{q}}^t + \ddot{\mathbf{q}}^t\Delta t\\ \mathbf{q}^{t+1} &= \mathbf{q}^t + \dot{\mathbf{q}}^t\Delta t\end{aligned} \tag{3}$$

where Δt represents the time interval between frames, $\mathbf{M}^{t^{-1}}$ is the inverse of the generalized inertia matrix. Inverse dynamics, on the other hand, aims at estimating the unknown physical parameters from observed motion. In this work, we propose to incorporate the Euler-Lagrange equations into model prediction by explicitly capturing the forward and inverse dynamics processes through a physics-based motion prediction model, as introduced below.

3.2. Physics-Based Human Motion Prediction

As illustrated in the top part of Fig. 1, the physics-based human motion prediction model predicts the future motion state $\hat{\mathbf{q}}^{T+n}_{physics}$ for frame $T+n$ by taking two inputs: the complete input history motion $\{\mathbf{q}^t\}_1^T$ and the motion states at three frames before $T+n$, denoted as $\{\hat{\mathbf{q}}^t\}_{T+n-3}^{T+n-1}$.

The physics-based model first characterizes the inverse dynamics by using neural networks to estimate the unknown physical parameters. Two separate Multi-Layer Perceptrons (MLP) are employed to extract $\mathbf{h}_p$, features related to the physical properties of the subject, and $\mathbf{h}_g$, features related to the geometry information near the current frame:

$$\mathbf{h}_p = \mathtt{MLP}_{h,p}(\{\mathbf{q}^t\}_1^T), \tag{4a}$$

$$\mathbf{h}_g = \mathtt{MLP}_{h,g}(\{\hat{\mathbf{q}}^t\}_{T+n-3}^{T+n-1}). \tag{4b}$$

We only consider three previous frames to extract the geometry features as information in three time frames can fully specify the Euler-Lagrange equations at certain time t. These extracted features are then concatenated and passed through three additional MLPs to respectively predict the unknown physical parameters in Eq. 2:

$$\mathbf{M}^{T+n-2} = \mathtt{MLP}_M(\mathbf{h}_p \oplus \mathbf{h}_g), \tag{5a}$$

$$\mathbf{C}^{T+n-2} = \mathtt{MLP}_C(\mathbf{h}_p \oplus \mathbf{h}_g), \tag{5b}$$

$$\mathbf{G}^{T+n-2} = \mathtt{MLP}_G(\mathbf{h}_p \oplus \mathbf{h}_g). \tag{5c}$$

Here, our design is aimed to fully leverage the expressivity of neural networks. We rigorously adhere to the physics equations to be characterized, specifically Eq. 2, in order to preserve the relationships among these physical parameters. Given the predicted physical parameters, the physics-based model employs an ODE solver to predict $\hat{\mathbf{q}}^{T+n}_{physics}$, where forward dynamics is performed following Eq. 3.

For training of the physics-based model, we utilize

$$\mathcal{L}_{physics} = \textstyle\sum_{T+1}^{T+N} \|\mathbf{q}^t - \hat{\mathbf{q}}^t_{physics}\| + \lambda \sum_{T+1}^{T+N} \|\mathbf{J}^t - \hat{\mathbf{J}}^t_{physics}\|, \tag{6}$$

where $\mathbf{q}^t$ is the ground truth generalized position and $\mathbf{J}^t$ are the ground truth 3D body joint positions computed using forward kinematic based on the generalized positions $\mathbf{q}^t$. On the other hand, $\hat{\mathbf{q}}^t_{physics}$ and $\hat{\mathbf{J}}^t_{physics}$ are the corresponding estimates generated by the physics-based model.

To predict future motion, we can iteratively apply the physics-based motion prediction model to generate configuration at the next frame from the previous estimates and the input history motion. However, the physics-based model generates future motion estimates primarily considering the information near the current frame. The accumulation of errors at each frame can be nontrivial as the prediction horizon increases. To address this issue, we introduce a data-driven model and a fusion model to alleviate the problem of error accumulation, as described below.

3.3. Integrating Motion Prediction Guidance

Data-Driven Model. Existing methods have demonstrated promising progress on utilizing data to capture long-term dependencies, where future human motion is directly predicted using neural networks from input history motion $\{\mathbf{q}^t\}_1^T$. Building upon the model proposed by [22], we introduce an data-driven human motion prediction model based on MLP to capture long-term dependencies and generate future data-driven estimates as:

$$\{\hat{\mathbf{q}}^t_{data}\}_{T+1}^{T+N} = \mathtt{MLP}_{data}(\{\mathbf{q}^t\}_1^T). \tag{7}$$

For training the data-driven model, we utilize the following loss functions:

$$\mathcal{L}_{data} = \textstyle\sum_{T+1}^{T+N} \|\mathbf{q}^t - \hat{\mathbf{q}}^t_{data}\| + \lambda \sum_{T+1}^{T+N} \|\mathbf{J}^t - \hat{\mathbf{J}}^t_{data}\|, \tag{8}$$

where $\hat{\mathbf{J}}^t_{data}$ are 3D body joint positions computed from $\hat{\mathbf{q}}^t_{data}$ using forward kinematics.

Fusion Model. As shall be seen in Sec. 4, the physics-based model is superior to data-driven approaches at short prediction horizons, but data-driven approaches are better at longer time horizons due to the abovementioned error accumulation issue. Therefore, we introduce an additional fusion model that combines the data-driven prediction and the physics-based prediction optimally. Specifically, at a future frame t, the output of the fusion model is:

$$\hat{\mathbf{q}}^t_{fusion} = (1 - \hat{w}^t)\hat{\mathbf{q}}^t_{physics} + \hat{w}^t\hat{\mathbf{q}}^t_{data}, \tag{9}$$

where $\hat{w}^t$ is a scalar fusion weight. Note that, instead of directly fusing the two estimates in one round, we perform iterative fusion to fully leverage the physics-based estimation. $\hat{\mathbf{q}}^t_{physics}$ is therefore the physics-based prediction generated using previous *fusion* estimates. Executing fusion from $t = T + 1$ to $t = T + N$ results in $\{\hat{\mathbf{q}}^t_{fusion}\}_{T+1}^{T+N}$.

To estimate the fusion weights, we utilize an MLP that takes as input the data-driven estimates and the physics-based estimates $\hat{\mathbf{q}}^t_{physics,p}$ (purely using physics without fusion), and outputs the weights at different time frames:

$$\{\hat{w}^t\}_{T+1}^{T+N} = \texttt{MLP}_{fusion}(\{\hat{\mathbf{q}}^t_{data}, \hat{\mathbf{q}}^t_{physics,p}\}_{T+1}^{T+N}). \tag{10}$$

As the errors are accumulated over time, we perform a time position encoding by adding a time index vector $\{t\}_{T+1}^{T+N}$ to the data-driven and physics-based estimates. The resulting values are then used for predicting the weights.

The loss function for training the fusion model is:

$$\begin{aligned}\mathcal{L}_{fusion} = &\sum_{T+1}^{T+N} \|\mathbf{q}^t - \hat{\mathbf{q}}^t_{fusion}\| + \lambda \sum_{T+1}^{T+N} \|\mathbf{J}^t - \hat{\mathbf{J}}^t_{fusion}\| \\ &+ \lambda_{reg} \sum_{T+1}^{T+N} |\hat{w}^t|\end{aligned} \tag{11}$$

where $\hat{\mathbf{J}}^t_{fusion}$ are the 3D body joint positions computed based on $\hat{\mathbf{q}}^t_{fusion}$. We introduce a regularization term on the fusion weights to encourage the prediction to rely more on the physics-based estimation.

3.4. Model Training and Testing Strategy

Model Training. Training of PhysMoP is two-stage. In the first stage, we train the physics-based model and the data-driven model by minimizing Eq. 6 and Eq. 8, respectively. For training the physics-based model, we employ a strategy inspired by [8, 60] where we use ground truth three previous estimates as input to the physics-based model for faster convergence. Once the training of the physics-based and data-driven models converges, we fix their model weights and proceed to the second stage, where we train the fusion model by minimizing Eq. 11.

Model Testing. During testing, given input history motion $\{\mathbf{q}^t\}_1^T$, we use the trained physics-based and data-driven model to respectively generate the physics-based and data-driven estimates. We then utilize the two estimates to compute the fusion weights and iteratively apply the fusion following Eq. 9 to obtain the final future motion estimates.

4. Experiment

We validate our proposed approach following the standard protocol used by [22, 41, 45, 46] and [56, 72]. Below we present the detailed experiment settings. In Sec. 4.1, we demonstrate the superior performance compared to State-of-the-Arts (SOTAs), followed by a qualitative evaluation in Sec. 4.2. Finally, we discuss our ablation study in Sec. 4.3.

Datasets. We employ three Motion Capture (MoCap) datasets for training and evaluation: Human3.6M [25], AMASS [43], and 3DPW [61]. Human3.6M consists of motion sequences of 7 subjects (S1, S5-9, S11) performing 15 daily actions. Testing utilizes S5, while the rest are for training and validation. AMASS is a collection of multiple MoCap datasets, including a larger number of subjects and actions. Training utilizes its training subset, and evaluation is performed on AMASS-BMLrub [58]. Lastly, 3DPW is a dataset collected from unconstrained environments, including complex activities like uphill walking and running for a bus. We evaluate the model trained on AMASS on the test set of 3DPW to assess its generalization performance.

Implementation. The input history length T is 25, and the output future prediction length N is 25. In contrast to previous works that exclude body translation and rotation, we consider them during both training and testing, creating a more challenging setting. For training loss weights, we set $\lambda = 2$ and $\lambda_{reg} = 1$. We utilize the Adam optimizer with a weight decay of $1e^{-4}$. The initial learning rate is set to $3e^{-4}$, and we apply a learning rate decay of 0.9 after every 500 training steps. For Human3.6M, we train the data-driven, physics-based, and fusion models for 30, 10, and 10 epochs, respectively. For AMASS, we train the data-driven, physics-based, and fusion models for 5, 2, and 2 epochs, respectively. Please refer to Supp. A for details of the model architecture and other experiment settings.

Evaluation Metrics. To measure the motion prediction quality, we report Mean Per Joint Position Error (MPJPE) at different future time stamps. MPJPE is computed as the mean 3D Euclidean distance between the predicted and ground truth joint positions after aligning the root joint.

4.1. Comparison with State-of-the-Arts (SOTAs)

First, we report PhysMoP's improvements over SOTA through the evaluation of short-term (<500ms) and long-term (>500ms) human motion prediction on Human3.6M, AMASS, and 3DPW. Then, we highlight PhysMoP's superior performance on short-term motion prediction via action-wise evaluation on Human3.6M.

Improvements Over State-of-the-Arts (SOTAs). We first discuss the evaluation on Human3.6M. Existing methods adopt two different evaluation protocols: Human3.6M-P1 and Human3.6M-P2. Human3.6M-P1 considers the target time frame [22, 41, 45, 46], while Human3.6M-P2 considers the average of all frames up to the target time frame [56, 72]. We report the evaluation results under both protocols in Tab. 1 and compare with existing methods under the same protocol. As shown, PhysMoP outperforms SOTA on both short-term and long-term motion prediction. Specifically, siMLPe [22] achieves the best performance among purely data-driven methods by utilizing MLP to model motion data. EqMotion [64] reaches a better model performance than siMPLe by further integrating domain knowledge about the geometric equivariance in motion data. PhysMoP significantly outperforms them by incorporating physics knowledge. Under Human3.6M-P1, PhysMoP reduces MPJPE achieved by siMLPe from 109.4mm and EqMotion from 106.9mm to 103.9mm at 1000ms. The error reduction becomes much more significant for shorter prediction horizons. For example, PhysMoP achieves MPJPE of just 2.1mm — a decrease of 78.1% from siMLPe's MPJPE of 9.6mm at 80ms. When evaluating under Human3.6M-P2, the enhancements introduced by PhysMoP remain consistent, demonstrating its superiority over existing approaches. Moreover, PhysMoP is similarly superior to existing data-driven methods when evaluated on other datasets such as AMASS-BMLrub and 3DPW, as illustrated in Tab. 2. For either within-dataset (AMASS-BMLrub) or cross-dataset (3DPW) evaluation, PhysMoP achieves consistent improvements over SOTAs on both short-term and long-term motion prediction accuracy. Particularly, for a prediction horizon of 80ms, siMLPe achieves MPJPE of 10.8mm on AMASS-BMLrub and 12.1mm on 3DPW, while PhysMoP achieves 0.6mm and 0.7mm, respectively – reductions of nearly 95%.

Existing methods utilize various neural networks to model body movements. In contrast, PhysMoP, by effectively incorporating physics, achieves superior performance on different datasets, both within-dataset and cross-dataset. To provide deeper insights into the short-term prediction capabilities of PhysMoP, we now delve into its action-wise human motion prediction performance on Human3.6M.

Superior Short-Term Prediction Performance. Tab. 3 presents MPJPE of short-term motion prediction (<=400ms) on different actions in Human3.6M. To highlight the improvements of PhysMoP over existing methods, we also report the relative error reduction (RED). As demonstrated, PhysMoP is superior across all actions to a degree that increases as the time horizons become shorter. Interestingly, the advantage of PhysMoP at the 400ms horizon is smallest for actions like "Walking" for which the average acceleration is largest (about twice that of the average across all actions), and can be worse than for existing approaches (in this case, our method yields MPJPE of 42.4mm vs. 39.2 for EqMotion). It is reasonable that the activities for which the physics-based model is least effective are those for which human muscular activity shifts on the fastest time scales.

Human3.6M-P1	MPJPE (↓)							
Time (ms)	80	160	320	400	560	720	880	1000
LTD [46]	12.2	25.4	50.7	61.5	79.6	93.6	105.2	112.4
Hisrep [45]	10.4	22.6	47.1	58.3	77.3	91.8	104.1	112.1
MSR-GCN [12]	11.3	24.3	50.8	61.9	80.0	-	-	112.9
SPGSN† [35]	10.4	22.3	47.1	58.3	77.4	-	-	109.6
ST-DGCN [41]	10.6	23.1	47.1	57.9	76.3	90.7	102.4	109.7
siMLPe [22]	9.6	21.7	46.3	57.3	75.7	90.1	101.8	109.4
EqMotion† [64]	9.1	20.1	43.7	55.0	73.4	-	-	106.9
Ours	**2.1**	**7.6**	**28.4**	**43.8**	**72.9**	**86.2**	**96.1**	**103.9**

Human3.6M-P2	MPJPE (↓)							
Time (ms)	80	160	320	400	560	720	880	1000
STS-GCN [56]	10.1	17.1	33.1	38.3	50.8	60.1	68.9	75.6
STG-GCN [72]	10.1	16.9	32.5	38.5	50.0	-	-	72.9
siMLPe [22]	4.5	9.8	22.0	28.1	39.3	49.2	57.8	63.7
Ours	**1.4**	**3.7**	**11.6**	**17.2**	**30.5**	**41.9**	**51.1**	**57.1**

Table 1. **Evaluation of prediction accuracy over various horizons on Human3.6M.** Human3.6M-P1 (top) and Human3.6M-P2 (bottom) stand for computing MPJPE (unit of mm) following the protocol used by [22,41,45,46] and [56,72], respectively. Results of other works are obtained from [22] and the respective paper (†).

AMASS-BMLrub	MPJPE (↓)							
Time (ms)	80	160	320	400	560	720	880	1000
convS2S [32]	20.6	36.9	59.7	67.6	79.0	87.0	91.5	93.5
LTD [46]	11.0	20.7	37.8	45.3	57.2	65.7	71.3	75.2
Hisrep [45]	11.3	20.7	35.7	42.0	51.7	58.6	63.4	67.2
siMLPe [22]	10.8	19.6	34.3	40.5	50.5	57.3	62.4	65.7
Ours	**0.6**	**2.1**	**9.0**	**13.0**	**23.9**	**37.1**	**50.0**	**61.4**

3DPW	MPJPE (↓)							
Time (ms)	80	160	320	400	560	720	880	1000
convS2S [32]	18.8	32.9	52.0	58.8	69.4	77.0	83.6	87.8
LTD [46]	12.6	23.2	39.7	46.6	57.9	65.8	71.5	75.5
Hisrep [45]	12.6	23.1	39.0	45.4	56.0	63.6	69.7	73.7
siMLPe [22]	12.1	22.1	38.1	44.5	54.9	62.4	68.2	72.2
Ours	**0.7**	**3.6**	**15.1**	**19.8**	**30.3**	**43.5**	**58.6**	**70.9**

Table 2. **Evaluation of prediction accuracy over various time horizons on AMASS-BMLrub (top) and 3DPW (bottom).** The model is trained on the AMASS training set and 3DPW is utilized for cross-dataset evaluation. Results of other works are obtained from [22]. The unit of MPJPE is mm.

Time (ms)	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Action	Walking				Eating				Smoking				Discussion			
Hisrep [45]	10.0	19.5	34.2	39.8	6.4	14.0	28.7	36.2	7.0	14.9	29.9	36.4	10.2	23.4	52.1	65.4
ST-DGCN [41]	10.2	19.8	34.5	40.3	7.0	15.1	30.6	38.1	6.6	14.1	28.2	34.7	10.0	23.8	53.6	66.7
siMLPe [22]	9.9	-	-	39.6	5.9	-	-	36.1	6.5	-	-	36.3	9.4	-	-	64.3
EqMotion [64]	9.0	17.5	32.6	**39.2**	6.3	13.6	28.9	36.5	5.5	11.3	23.0	**29.3**	8.2	18.9	42.1	53.9
Ours	**2.6**	**9.0**	**29.0**	42.4	**1.3**	**4.9**	**19.9**	**31.1**	**1.3**	**4.9**	**19.5**	31.6	**2.0**	**7.6**	**31.0**	**48.3**
(RED, %)	71.1	48.6	11.0	-	78.0	64.0	30.7	13.9	76.4	56.6	15.2	-	75.6	59.8	26.4	10.4
Action	Directions				Greeting				Phoning				Posing			
Hisrep [45]	7.4	18.4	44.5	56.5	13.7	30.1	63.8	78.1	8.6	18.3	39.0	49.2	10.2	24.2	58.5	75.8
ST-DGCN [41]	7.2	17.6	40.9	51.5	15.2	34.1	71.6	87.1	8.3	18.3	38.7	48.4	10.7	25.7	60.0	76.6
siMLPe [22]	6.5	-	-	55.8	12.4	-	-	77.3	8.1	-	-	48.6	8.8	-	-	73.8
EqMotion [64]	6.3	15.8	38.9	50.1	-	-	-	-	7.4	16.7	36.9	47.0	8.2	18.9	43.4	57.5
Ours	**1.6**	**6.0**	**23.8**	**37.6**	**2.9**	**10.2**	**37.0**	**56.1**	**1.7**	**6.2**	**23.6**	**37.0**	**2.2**	**8.1**	**30.0**	**46.3**
(RED, %)	74.6	62.0	38.8	25.0	76.6	66.1	42.0	27.4	77.0	62.9	36.0	21.3	73.2	57.1	30.9	19.5
Action	Purchases				Sitting				Sitting down				Taking photo			
Hisrep [45]	13.0	29.2	60.4	73.9	9.3	20.1	44.3	56.0	14.9	30.7	59.1	72.0	8.3	18.4	40.7	51.5
ST-DGCN [41]	12.5	28.7	60.1	73.3	8.8	19.2	42.4	53.8	13.9	27.9	57.4	71.5	8.4	18.9	42.0	53.3
siMLPe [22]	11.7	-	-	72.4	8.6	-	-	55.2	13.6	-	-	70.8	7.8	-	-	50.8
EqMotion [64]	-	-	-	-	8.1	18.0	41.2	52.9	13.0	26.5	56.2	70.7	-	-	-	-
Ours	**2.5**	**9.1**	**36.0**	**56.1**	**1.7**	**6.2**	**23.7**	**37.4**	**2.7**	**9.6**	**31.6**	**46.8**	**1.6**	**6.0**	**24.4**	**38.3**
(RED, %)	80.0	68.3	40.1	23.5	79.0	65.6	42.5	29.3	79.2	63.8	43.8	33.8	81.0	68.3	41.9	28.1
Action	Waiting				Walking dog				Walking together				Average			
Hisrep [45]	8.7	19.2	43.4	54.9	20.1	40.3	73.3	86.3	8.9	18.4	35.1	41.9	10.4	22.6	47.1	58.3
ST-DGCN [41]	8.9	20.1	43.6	54.3	18.8	39.3	73.7	86.4	8.7	18.6	34.4	41.0	10.3	22.7	47.4	58.5
siMLPe [22]	7.8	-	-	53.2	18.2	-	-	83.6	8.4	-	-	41.2	9.6	21.7	46.3	57.3
EqMotion [64]	7.6	17.4	39.9	51.1	-	-	-	-	7.8	16.1	30.6	**37.1**	9.1	20.1	43.7	55.0
Ours	**1.9**	**7.0**	**25.8**	**39.8**	**3.6**	**12.5**	**45.0**	**68.9**	**2.1**	**7.2**	**25.4**	39.5	**2.1**	**7.6**	**28.4**	**43.8**
(RED, %)	75.0	59.8	35.3	22.1	80.2	68.2	38.6	20.2	73.1	55.3	17.0	-	77.0	62.2	35.0	20.4

Table 3. **Action-wise evaluation of short-term motion prediction on Human3.6M.** MPJPE (unit of mm) is computed using the protocol used by [22,41,45,46] (Human3.6M-P1). Results of other works are obtained from the respective papers. The values of MPJPE are in mm; thus smaller values are better. Besides MPJPE, we report the relative error REDuction (RED), which is calculated as the error reduction achieved by PhsyMoP (Ours) relative to the second best (marked by underline). RED is caculated in percentage (%). Larger RED indicates larger improvements achieved by PhysMoP over existing approaches.

4.2. Qualitative Evaluation

We showcase two examples of 3D human motion predicted by PhysMoP in Fig. 2. The motion sequences are from the Human3.6M test set. As illustrated, PhysMoP generates favourable results at different prediction horizons. We note that existing methods do not model global rotation, while PhysMoP can further recover accurate global rotation.

4.3. Ablation Study

This section includes the ablation study to illustrate the effectiveness of different components of PhysMoP.

Benefits of the Physics-Based, Data-Driven, and Fusion Model. We individually evaluate the physics-based, data-driven, and fusion model and report the results in Tab. 4. As shown, the physics-based model, when not integrating the motion prediction guidance given by the data-driven and fusion model, generates precise short-term motion prediction but its long-term prediction can suffer at longer time stamps ("Physics" in Tab. 4). The data-driven model, on the other hand, has poor short-term motion prediction performance but it demonstrates better performance in long-term prediction than the physics-based model ("Data" in Tab. 4). By combining the two results using the proposed fusion model, PhysMoP takes the advantages of the two models. We also demonstrate that the estimates obtained by our proposed fusion strategy is better than the heuristic method, that is taking the averaging of the physics-based and data-driven estimates ("Vanilla" in Tab. 4). To further study the effec-

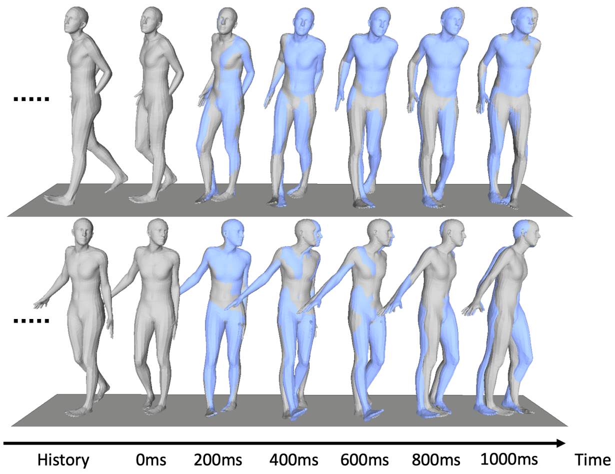

Figure 2. **Qualitative evaluation of PhysMoP.** The motion sequences are from Human3.6M test set. The ground truth and estimated future motion is marked in grey and blue colors, respectively. Larger overlaps between the two configurations indicate smaller prediction errors.

Human3.6M	Time (ms)	Physics	Data	Fusion Vanilla	 PhysMoP
MPJPE (↓)	80	2.1	4.4	4.1	2.1
	160	7.6	10.6	10.7	7.6
	320	28.4	33.8	34.6	28.4
	400	42.2	47.8	46.7	43.8
	560	72.8	68.7	68.2	72.9
	720	102.6	85.0	85.1	86.2
	880	129.8	97.3	97.1	96.1
	1000	148.2	103.3	103.4	103.9
ACCL (↓)	-	1.8	8.2	3.3	2.3

Table 4. **Ablation study on different components of PhysMoP.** "Physics" and "Data" represents the data-driven and physics-based model, respectively. "Vanilla" means generating prediction by directly taking the average of the data-driven and physics-based estimates. Computing MPJPE (mm) at different future time stamps follows [22, 41, 45, 46] (Human3.6M-P1). "ACCL" stands for the acceleration error averaged over time to measure the physical plausibility of the predicted motion.

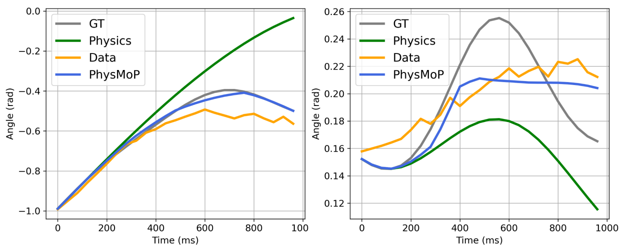

Figure 3. **Joint angles estimated by PhysMoP with comparison to the data-driven and physics-based model.** The testing motion sequences are from Human3.6M test set. The curves are estimated joint angles of left leg with comparison to the ground truth at different future time stamps.

tiveness of PhysMoP, we compare the joint angles estimated by the data-driven model, physics-based model, and PhysMoP with examples illustrated in Fig. 3. As shown, the joint angles estimated by the data-driven model (orange curves) exhibit excessive jittering even at the starting future time stamps (such as <400ms). By contrast, the physics-based estimates offer a smoother trajectory and higher accuracy within shorter time horizons, almost aligning perfectly with the ground truth (grey curves). However, the physics-based estimates may deviate from the ground truth over time, leading to a worse prediction accuracy than the data-driven estimates. By leveraging the fusion model, PhysMoP's estimates are closer to the ground truth. The differences in the estimated joint angles over time emphasize the advantages of PhysMoP in addressing the limitations of purely data-driven approaches by effectively incorporating physics.

Improved Physical Plausibility. Incorporating physics also results in more realistic prediction, where the estimates retain better physical plausibility. We demonstrate PhysMoP's improved physical plausibility through the evaluation of acceleration error (please refer [68] for the calculation details). We compute the average acceleration error over all joints and future frames and report the results in Tab. 4. ACCL has a unit of mm/frame^2, with smaller values indicating better physical plausibility. PhysMoP significantly reduces the acceleration error achieved by the data-driven model from 8.2 to 2.3. This further highlights the advantage of incorporating physics principles into the motion prediction model.

5. Conclusion

We have demonstrated that, by incorporating fundamental physics principles into predictive models of human motion, the prediction accuracy can be improved significantly over horizons of up to one second. At extremely short prediction horizons of 80 msec or a bit more, the improvement can be over a factor of 10 in some cases. Specifically, we introduced PhysMoP, a method that incorporates the Euler-Lagrange equations by explicitly capturing the forward and inverse dynamics using neural networks. PhysMoP utilizes data-driven and fusion models to guide the physics-based prediction, thereby reducing the error accumulation that plagues purely physics-based approaches and generating predictions that leverage human behavioral patterns gleaned from data while also observing physical laws.

References

[1] NSF Grant 0196217. CMU graphics lab motion capture database. 1

[2] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) 2015*, June 2015. 1

[3] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 1

[4] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 1, 2

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 1

[7] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, 2000. 1, 2

[8] Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers. *arXiv preprint arXiv:2202.03376*, 2022. 5

[9] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017. 1, 2

[10] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1423–1432. IEEE, 2019. 1, 2

[11] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020. 2

[12] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11467–11476, 2021. 1, 2, 6

[13] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104, 2020. 1

[14] Roy Featherstone. *Rigid body dynamics algorithms*. Springer, 2014. 1

[15] Advanced Computing Center for the Arts and Design. Accad mocap dataset. 1

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 1, 2

[17] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13190–13200, 2022. 1, 2

[18] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2022. 2

[19] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12116–12125, 2019. 1, 2

[20] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the european conference on computer vision (ECCV)*, pages 786–803, 2018. 1

[21] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 1

[22] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023. 1, 2, 4, 5, 6, 7, 8

[23] Ludovic Hoyet, Kenneth Ryall, Rachel McDonnell, and Carol O'Sullivan. Sleight of hand: perception of finger motion from reduced marker sets. In *Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games*, pages 79–86, 2012. 1

[24] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6417–6426, 2022. 1, 2

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1, 5

[26] Deepak Kumar Jain, Masoumeh Zareapoor, Rachna Jain, Abhishek Kathuria, and Shivam Bachhety. Gan-poser: an improvised bidirectional gan model for human motion prediction. *Neural Computing and Applications*, 32(18):14579–14591, 2020. 1

[27] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. 2

[28] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015. 1

[29] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506, 2021. 1

[30] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019. 1

[31] Andreas M Lehrmann, Peter V Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014. 1, 2

[32] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5226–5234, 2018. 1, 2, 6

[33] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In *European Conference on Computer Vision*, pages 479–496. Springer, 2022. 1, 2

[34] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3316–3333, 2021. 1, 2

[35] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European Conference on Computer Vision*, pages 18–36. Springer, 2022. 2, 6

[36] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8649, 2019. 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 3

[38] Eyes JAPAN Co Ltd. Eyes japan mocap dataset. 1

[39] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: quantization-based 3d human motion generation and forecasting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 417–435. Springer, 2022. 1

[40] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *arXiv preprint arXiv:2206.09106*, 2022. 2

[41] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 1, 5, 6, 7, 8

[42] Takahiro Maeda and Norimichi Ukita. Motionaug: Augmentation with physical correction for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6427–6436, 2022. 1, 2

[43] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1, 5

[44] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016. 1

[45] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 474–489. Springer, 2020. 1, 5, 6, 7, 8

[46] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 1, 5, 6, 7, 8

[47] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 1, 2

[48] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2276–2284, 2021. 1

[49] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022. 1

[50] Soohwan Park, Hoseok Ryu, Seyoung Lee, Sunmin Lee, and Jehee Lee. Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)*, 38(6):1–11, 2019. 1

[51] Maziar Raissi, Paris Perdikaris, and George 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*, 378:686–707, 2019. 2

[52] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 71–87. Springer, 2020. 2

[53] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6457–6466, 2022. 1

[54] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 1, 2

[55] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 2

[56] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021. 1, 2, 5, 6

[57] Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning*, pages 1025–1032, 2009. 1, 2

[58] Nikolaus F Troje. Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of vision*, 2(5):2–2, 2002. 5

[59] Raquel Urtasun, David J Fleet, Andreas Geiger, Jovan Popović, Trevor J Darrell, and Neil D Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th international conference on Machine learning*, pages 1080–1087, 2008. 1, 2

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[61] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 5

[62] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007. 1, 2

[63] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11532–11541, 2021. 1, 2

[64] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1410–1420, 2023. 2, 6, 7

[65] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 1

[66] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020. 1

[67] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022. 1

[68] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7159–7169, 2021. 2, 8

[69] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 1

[70] Yufei Zhang, Hanjing Wang, Jeffrey O Kephart, and Qiang Ji. Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9020–9032, 2023. 3

[71] Zhibo Zhang, Yanjun Zhu, Rahul Rai, and David Doermann. Pimnet: Physics-infused neural network for human motion prediction. *IEEE Robotics and Automation Letters*, 7(4):8949–8955, 2022. 1, 2

[72] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6447–6456, 2022. 1, 2, 5, 6