Human Motion Prediction under Unexpected Perturbation

Jiangbei Yue[1] Baiyi Li[1] Julien Pettré[2] Armin Seyfried[3] He Wang[4 *]
[1]University of Leeds, UK [2]INRIA Rennes, France
[3]Forschungszentrum Jülich, Germany [4]University College London, UK

Abstract

We investigate a new task in human motion prediction, which is predicting motions under unexpected physical perturbation potentially involving multiple people. Compared with existing research, this task involves predicting less controlled, unpremeditated and pure reactive motions in response to external impact and how such motions can propagate through people. It brings new challenges such as data scarcity and predicting complex interactions. To this end, we propose a new method capitalizing differentiable physics and deep neural networks, leading to an explicit Latent Differentiable Physics (LDP) model. Through experiments, we demonstrate that LDP has high data efficiency, outstanding prediction accuracy, strong generalizability and good explainability. Since there is no similar research, a comprehensive comparison with 11 adapted baselines from several relevant domains is conducted, showing LDP outperforming existing research both quantitatively and qualitatively, improving prediction accuracy by as much as 70%, and demonstrating significantly stronger generalization.

1. Introduction

Human motion prediction aims to predict the future movements given the past motions, which has been heavily studied in computer vision [17, 67–69]. Deviating from existing research, we are interested in a new task setting: predicting human motions, on both individual and group levels, under unexpected physical perturbation. On the individual level, physical perturbation causes reactive motions as opposed to active motions. On the group level, such perturbations can propagate through people while possibly being intensified, *e.g.* a push at the back of a line of people could be transferred all the way to the front. These motions have not been investigated. Incorporating physical perturbation potentially extends motion prediction to new application domains *e.g.* balance recovery in biomechanics [4, 18], reactive motions for character animation [3, 14], crowd crush induced by pushing [6, 53], humanoid robots [24, 26], *etc.*

Incorporating physical perturbation in prediction imposes new challenges. First, the motions are purely reactive and less controlled such that they are less smooth and less coordinated among body parts. Furthermore, this perturbation can propagate through people when they are packed and the space to recover balance is restricted, such that an attempt to recover balance relies on pushing others. Last but not least, unlike existing research, the data for motion prediction under perturbation is extremely scarce. Not only is it rare to capture full-body motions under such circumstances, but it is also difficult to record the interactions between people, *e.g.* forces of pushes.

Before deep learning, many areas have formulated this problem, which can be broadly divided into two categories. The first is physics-based where human bodies are simplified into connected rigid bodies [37, 53]. The reaction to push is solved via optimization to compute what forces are needed to recover balance [40, 42], or through carefully tuning feed-forward controllers [37, 38]. These methods, despite aiming to mimic the balance recovery of humans, do not learn from human data and therefore cannot predict human motions. Alternatively, reactions to perturbation can be learned from data via regression [60], optimization [66], or reinforcement learning [63]. Comparatively, this type of method tends to generate more human-like motions, but they are not designed for prediction.

Recently, deep learning [44, 56, 67, 68] have dominated human motion prediction, but they cannot be adapted for our problem. First, most datasets only contain single-body motions without external perturbation. Even when multiple people are captured, it is not under unexpected perturbation. To predict push propagation, one would still need to measure information *e.g.* contact forces between people, ground friction, muscle forces, *etc.*, which are all absent. This data scarcity essentially rules out most deep-learning methods. Furthermore, there is also little work in modelling the physical/bio-mechanical interactions that can potentially propagate through people. Current research includes motion forecasting, generation and synthesis. Most

*Corresponding author, he_wang@ucl.ac.uk.

motion forecasting methods [17, 33, 67] are for a single person, with a few recent exceptions [44, 68, 69] but not involving perturbation. Alternatively, our problem could be formulated as motion generation conditioned on external perturbation. However, current methods [8, 46, 49–51, 73] again do not explicitly model close interactions among multiple people caused by perturbation. Theoretically, motion synthesis [36, 54, 55, 64] is a possibility, which potentially can predict motions under perturbation. But they require dense control signals to guide the synthesis, or/and extensive physical simulation. Therefore, it requires manual labor or/and is difficult to scale to many people.

To address the aforementioned challenges, we need a model that has *high data efficiency*, *strong generalizability* and can model *interactions between people*. In other words, this model needs to be able to learn from a small number of samples, can predict accurately in situations similar to the data, and is capable of generating plausible motions in drastically different scenarios. To this end, we propose a new deep-learning model for human motion prediction under unexpected perturbation. To address the data scarcity, we propose a scalable differentiable physics (DP) model for the human body, to learn the balance strategy and interaction propagation between people, inspired by recent DP research [15, 59, 71]. However, naively following existing DP approaches means we would need to make the full-body simulation differentiable for each individual. Not only is motion intrinsically indifferentiable due to *e.g.* foot contact, but full-body physical models are too computationally expensive to scale. Therefore, we propose a latent DP space where the full-body physics is reduced into a differentiable inverted pendulum model (IPM) [19, 25, 29, 41], and the full-body poses are mapped to and recovered from the IPM. At the low level, the IPM governs body physics and learns key forces such as ground friction and balance recovery. As the IPM is simple, the required data is small. At the high level, we use neural networks to recover the full-body pose from the IPM, which also does not require much data as the IPM provides strong guidance. We refer to our model as the Latent Differentiable Physics (LDP) model. Note different from other latent physics models where the dimensionality reduction is implicit [47, 62], ours is explicit and physically meaningful (*i.e.* mapping from full-body to IPM).

We show LDP can learn from very limited data and perform well under many widely used metrics. Since there is no similar work to our best knowledge, we adapt a wide range of baseline methods in the most relevant areas (motion forecasting, motion generation and motion synthesis), in single-person and multi-people scenarios, for comparison. The results demonstrate that LDP outperforms them both quantitatively and qualitatively. Notably, our model exhibits remarkable generalizability. It can accommodate unseen out-of-distribution perturbations, group sizes, and group formations, potentially extending our research beyond human motion prediction into broader areas, *e.g.* crowd simulation. Furthermore, owing to the explicit physics model, our model possesses a distinctive feature: explainability, providing plausible explanations for the predicted motion. Formally, our contributions include:

- A new task: human motion prediction under unexpected perturbation. To our best knowledge, this is the first deep-learning paper addressing this problem.
- A novel differentiable physics model in human motion prediction that explicitly considers physical interactions.
- A new differentiable IPM model that learns body physics under complex interactions.
- A novel differentiable interaction model that can learn interactions and interaction propagation.

2. Related Work

Human Motion Prediction. Compared with traditional statistical machine learning [31, 57], deep learning has dominated human motion prediction recently. It can be formulated as a sequence-to-sequence task modelled by Recurrent Neural Networks [13, 22, 43]. Also, human skeletons can be seen as graphs so that spatio-temporal graph convolutions can be employed [9, 10, 33, 34, 72]. Transformer-based methods [1, 5] use the attention mechanism to capture spatial and temporal correlations. Recently, there has been a surge of interest in multi-people motion prediction [44, 58, 68, 69]. MRT [58] models the social interactions between humans via a global encoder. JRFormer [68] exploits the joint relation representation for modelling the interactions where physical interactions are considered implicitly. However, existing methods share a common limitation - they do not consider unexpected perturbations, restricting their applications in predicting actively planned/controlled motions. Additionally, explicit physical interactions between people have often been overlooked in these methods. Our model extends the research to a more challenging scenario involving unexpected perturbation and perturbation propagation. The explicit physics knowledge in our model enables it to achieve better prediction, generalizability, and explainability.

Traditional Research on Balance Recovery Relevant research has been conducted in other fields where traditional methods mainly focus on modelling balance recovery strategies in response to perturbations [4, 6, 18, 24, 26, 42]. Brodie *et al.* [4] analyzed the biomechanical mechanisms in the balance recovery following an unexpected perturbation such as trips and slips. Chen *et al.* [6] studied the dynamics of individuals under pushing in crowds. A new controller was proposed to recover balance for bipedal robots under perturbation [42]. In parallel, some traditional methods aim to synthesize reactive motions to perturbation [3, 39, 40, 45, 60]. Arikan *et al.* [3] proposed an al-

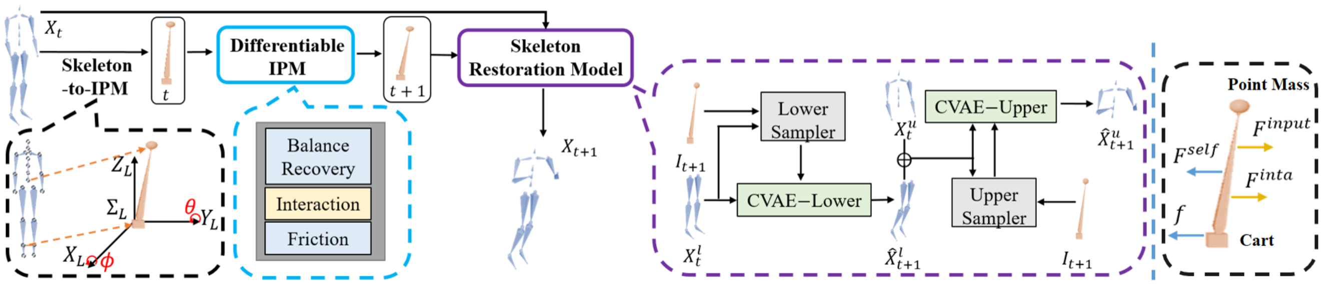

Figure 1. **Overview of our model.** Given a frame X_t, it is first mapped into the IPM space via Skeleton-to-IPM to get its IPM state I_t. Then I_t is simulated for one step via Differentiable IPM to compute I_{t+1}. Lastly, the full-body frame X_{t+1} is recovered from I_{t+1} via Skeleton Restoration Model. The IPM is shown in the right figure. The full-body state X is represented by joint positions.

gorithm for selecting and adjusting the motions from data to synthesize the motion for animating virtual characters being pushed. [39, 45] explored how to turn the given motions under perturbation into physically valid ones. Overall, traditional methods cannot predict motions under perturbation, either because they do not learn from data or have limited learning capacity. By contrast, we incorporate DP with deep neural networks to predict such human motions.

Differentiable Physics. DP is an emerging field focusing on combining traditional physics models with deep learning techniques, to provide high data efficiency and explainability. Consequently, many domains have investigated differentiable physics such as robotics [7, 30, 61], physics [20, 23], computer vision [70, 71], and computer graphics [15, 35]. We propose the first explicit latent differentiable physics model for human motion prediction under unexpected physical perturbation.

3. Methodology

Problem Definition. Given a motion with multiple people, we denote the skeletal pose of the nth person at frame t as $X_t^n \in \mathbb{R}^{J\times 3}$ where J is the joint number. Unlike existing research aiming to predict p frames $\{\hat{X}^n_{T-p+1:T}\}^N_{n=1}$ given k frames $\{X^n_{1:k}\}^N_{n=1}$ history, we minimize the required history due to limited data. Given the initial frame $\{X^n_0\}^N_{n=1}$ and the input forces F^{input}, we aim to predict the following T frames, by solving an initial problem:

$$\{\hat{X}^n_{1:T}\}^N_{n=1} = \mathcal{S}_\gamma(\{X^n_0\}^N_{n=1}, IPM_\eta(\mathcal{M}(\{X^n_0\}^N_{n=1}), F^{input})) \quad (1)$$

where $\{\hat{X}^n_{1:T}\}^N_{n=1}$ is the predicted T frames. $\mathcal{M}$ is a Skeleton-to-IPM mapping $\mathcal{M} : \mathcal{X} \to \mathcal{I}$ where $\mathcal{X}$ and $\mathcal{I}$ are the state space of skeleton poses (represented by joint positions) and the IPM respectively. IPM_η is a Differentiable IPM with learnable parameters η. Finally, $\mathcal{S}_\gamma$ is the inverse mapping, *i.e.* Skeleton Restoration Model, $\mathcal{S}_\gamma : \mathcal{I} \to \mathcal{X}$, reconstructing full-body skeleton pose from IPM states, with learnable parameters γ. An overview of our model is shown in Fig. 1. Given a motion, we map the full-body poses into their corresponding states of an IPM [19, 25, 29] as $\{I^n_0\}^N_{n=1} = \mathcal{M}(\{X^n_0\}^N_{n=1})$. By simulating the IPM forward in time via IPM_η, it can learn the key parameters η. The interaction forces between people are also learned simultaneously. Meanwhile, our Skeleton Restoration Model $\mathcal{S}_\gamma$ recovers the full-body poses from the predicted IPM states from IPM_η. For training, we minimize the mean squared error (MSE) between the predicted $\{\hat{X}^n_{1:T}\}^N_{n=1}$ and the ground-truth poses $\{X^n_{1:T}\}^N_{n=1}$:

$$Loss = MSE(\{\hat{X}^n_{1:T}\}^N_{n=1}, \{X^n_{1:T}\}^N_{n=1}) \quad (2)$$

where we need to specify $\mathcal{S}_\gamma$, IPM_η and $\mathcal{M}$ in Eq. (1). We give key equations and model information below and refer the readers to the supplementary material (SM) for details.

3.1. Latent Physics Space for Full-body Motions

3.1.1 Background and Skeleton-to-IPM Mapping

We first introduce IPM_η and $\mathcal{M}$ in Eq. (1). Differentiable physics (DP) has shown extremely high data efficiency because physics can act as a strong inductive bias and eliminates the reliance on large amounts of training data [11, 59, 71]. For our model, a key design choice is to choose a DP model that has the right level of granularity while being scalable. Among many possible choices from full-body physics [2] to simple rods [65], we choose the Inverted Pendulum Model (IPM) [19, 25, 29] as it can fully capture balance loss and recovery while being scalable.

Our IPM has a massless rod mounted to a cart with a point mass at the end of the rod (Fig. 1 right). Denoting its state $\mathcal{I} \ni I_t = [x_t, y_t, \theta_t, \phi_t] \in \mathbb{R}^4$ at time step t where $[x, y]$ is the coordinates of the cart in the xy-plane and $[\theta, \phi]$ is the rotation angles of the rod around Y_L axis and X_L axis in the local coordinate system Σ_L, respectively. Our full-body pose X is represented by 22 joint positions. $\mathcal{M}$ in Eq. (1) is defined as (Fig. 1 left): the hip joint is mapped onto the point mass, and the midpoint between the two ankle joints is mapped onto the center of the cart. The point mass and the cart jointly determine the two angles $[\theta, \phi]$.

Next, IPM_η is defined. Given the initial IPM state, we can simulate it in time by solving Eq. (3) repeatedly [41]:

$$M(I_t, l_t)\ddot{I}_t + C(I_t, \dot{I}_t, l_t) + G(I_t, l_t) = F_t^{net} \quad (3)$$

where $M \in \mathbb{R}^{4\times4}$, $C \in \mathbb{R}^{4\times1}$ and $G \in \mathbb{R}^{4\times1}$ are the inertia matrix, the Centrifugal/Coriolis matrix, and the external force such as gravity, which are all functions of state I_t, its first-order derivative $\dot{I}_t$ and the rod length l_t. While the standard IPM has a fixed rod length, we allow it to change as the distance between the hip and the middle of two ankles can drastically change in human motions. Therefore, we also predict l_t at each time step. Overall given the net force $F_t^{net} \in \mathbb{R}^4$ and the rod length l_t, we can solve Eq. (3) for the next state I_{t+1} via a semi-implicit scheme $\dot{I}_{t+1} = \dot{I}_t + \triangle t \ddot{I}_t$ and $I_{t+1} = I_t + \triangle t \dot{I}_{t+1}$, where $\triangle t$ is the time step.

Finally, the learnable parameters η in IPM_η parameterize F_t^{net} and the rod length l_t, where the formulation differs between single-person and multi-people, and will be elaborated later. It's notable that F_t^{net} in Eq. (3) is the generalized force. Using the generalized force (instead of the Euler force) keeps the motion equation simple, and its entries have explicit physical meanings as shown later.

3.1.2 Single-Person Prediction via Differentiable IPM

Under single-person, we only consider Balance-Recovery and Friction (blue blocks in Fig. 1 Differentiable IPM) when predicting F^{net}. Specifically, we consider three forces:

$$F_t^{net} = F_t^{self} + f_t + F_t^{input} \quad (4)$$

where F_t^{self}, f_t, and F_t^{input} are the balance recovery force, the ground friction and the external perturbation. The Balance-Recovery module learns F_t^{self} which is further decomposed into $F_t^{self} = F_t^{self-pd} + F_t^{self-nn}$. This decomposition is because F_t^{self} is the muscle force at the hinge of the rod which serves two purposes. The first one is to give a feed-forward torque $F_t^{self-pd}$ to react to perturbation for balance recovery, and the second is to give a torque correction $F_t^{self-nn}$ for tracking observed motions. In generalized forces, we parameterize $F_t^{self-pd}$ by proportional derivative (PD) control:

$$F_t^{self-pd} = K_p e_t + K_d \dot{e}_t, e_t = s_d - s_t \quad (5)$$

where e_t is the PD state error, K_p and K_d are the control parameters. Different from the IPM state, the current PD state is $s_t = [\dot{x}_t, \dot{y}_t, \theta_t, \phi_t]$ and the desired PD state s_d is $[0, 0, 0, 0]$. In other words, we assume people tend to recover to the upright body pose and zero linear velocity after unexpected perturbation, which is a widely accepted assumption [28, 32, 48]. However, $F_t^{self-pd}$ only captures the general balance recovery strategy. To mimic the data, we parameterize $F_t^{self-nn}$ with a Long Short Term Memory (LSTM) network:

$$F_t^{self-nn} = LSTM([\theta_t, \phi_t, \dot{x}_t, \dot{y}_t, \dot{\theta}_t, \dot{\phi}_t, M]), \quad (6)$$

where M is the mass of the person.

Ground friction f_t is the main reason for successful self-balance and therefore needs to be explicitly considered. In generalized forces, friction affects the IPM motion via damping [41]. So we parameterize $f_t = -\mu[\dot{x}_t, \dot{y}_t, 0, 0]$, where the parameter μ is a learnable positive scalar and shared by all people for simplicity. The damping force only directly influences the cart motion. Finally, to compute Eq. (3), we also need to predict the change of the rod length l_t, where we employ a multi-layer perception (MLP):

$$\triangle l_t = MLP([\theta_t, \phi_t, \dot{x}_t, \dot{y}_t, \dot{\theta}_t, \dot{\phi}_t, F_t^{self}, M, l_t]) \quad (7)$$

where l_t is the rod length at time step t. We predict the rod length at the next time step by $l_{t+1} = l_t + \triangle l_t$. Finally, after obtaining the prediction of F_t^{net} and l_t at every time step t, we can calculate the next IPM state by solving Eq. (3) via the semi-implicit scheme mentioned above.

3.1.3 Multi-people with Differentiable Interaction

When there is more than one person, the complexity increases quickly. The main reason is that the interaction propagation among people is: (1) complex, *e.g.* complicated contact positions/duration/forces. (2) hard to capture in data. Therefore, we propose to consider them as latent variables that cannot be directly observed. But again large amounts of data would be needed if we only relied on data to infer these variables. Therefore we model the interactions in the reduced IPM space, rather than the original space, so that it becomes a Differential Interaction Model (DIM).

Our DIM models a differentiable interaction force between any two IPMs and is learned in the Interaction module (the yellow block in Fig. 1 Differentiable IPM). The overall net force on an IPM in multi-people then becomes:

$$F_t^{net} = F_t^{self} + F_{t,n}^{inta} + f_t + F_t^{input} \quad (8)$$

where F_t^{self}, f_t and F_t^{input} are the same as Eq. (4). Note all forces are learned and shared among all people, so that we can generalize to an arbitrary number of people later. $F_{t,n}^{inta} \in \mathbb{R}^4$ is the new interaction force:

$$F_{t,n}^{inta} = \sum_{j\in\Omega_{t,n}} F_{t,nj}^{inta} = \sum_{j\in\Omega_{t,n}} F_{t,nj}^{inta-bs} + F_{t,nj}^{inta-nn} \quad (9)$$

where $\Omega_{t,n}$ is the neighborhood of the person n at time t. $F_{t,nj}^{inta}$ is the interaction force applied onto person n from her/his neighbor $j \in \Omega_{t,n}$. We model two factors in $F_{t,nj}^{inta}$: $F_{t,nj}^{inta-bs}$ and $F_{t,nj}^{inta-nn}$. The first $F_{t,nj}^{inta-bs}$ represents a consistent and trackable repulsive tendency when two IPMs get close, while $F_{t,nj}^{inta-nn}$ captures the variations of the repulsion. So we expect $F_{t,nj}^{inta-bs}$ to capture most of the interaction while $F_{t,nj}^{inta-nn}$ being a supplement. To this end, we separate the dimensions of an IPM state $I = [x, y, \theta, \phi]$ into two groups $[x, y]$ and $[\theta, \phi]$ and treat them separately as $F_{nj}^{bs-xy} \in \mathbb{R}^2$ and $F_{nj}^{bs-\theta\phi} \in \mathbb{R}^2$, such that $F_{t,nj}^{inta-bs} = [F_{nj}^{bs-xy}, F_{nj}^{bs-\theta\phi}]^{\mathbf{T}}$, where we omit the time subscript t and the superscript $inta$ for simplicity.

For F_{nj}^{bs-xy}, we define a repulsive potential energy between two close IPMs which leads to a repulsive force:

$$F_{nj}^{bs-xy}(r_{nj}) = -\nabla_{r_{nj}} \mathcal{U}[b(r_{nj})], \;\; \mathcal{U}[b] = ue^{-\frac{b}{\sigma}} \quad (10)$$

$$b = \frac{1}{2}\sqrt{(\|r_{nj}\| + \|r_{nj} - \triangle t\dot{r}_{jn}\|)^2 - \|\triangle t\dot{r}_{jn}\|^2}. \quad (11)$$

where $r_{nj} = r_n - r_j$ is the relative position of the carts of a person and his/her neighbor j, *i.e.* r_n is the vector $[x, y]$ in the IPM state I_n. The $\mathcal{U}[b]$ is the repulsive potential with elliptical equipotential lines, and u and σ are hyperparameters. b is the semi-minor axis of the ellipse where $\dot{r}_{jn} = \dot{r}_j - \dot{r}_n$ is the relative velocity.

For $F_{nj}^{bs-\theta\phi}$, we treat it as a force with a constant magnitude (tunable hyperparameter) and apply it on θ and ϕ independently. Although the magnitude is constant, its directions can vary in different situations. We explain it for θ and the same principle applies to ϕ. On the high level, we need to decide the direction of $F_{nj}^{bs-\theta\phi}$ based on the states of two close IPMs. θ can be positive, zero and negative. For two IPMs, this produces a total of 9 possible states, which we detail in the SM.

After defining $F_{t,nj}^{inta-bs}$, we explain $F_{t,nj}^{inta-nn}$ which should capture the variation of interactions. Unlike $F_{t,nj}^{inta-bs}$ where we can define an explicit form, we learn $F_{t,nj}^{inta-nn}$ via an MLP:

$$F_{nj}^{nn} = MLP([x_{nj}, y_{nj}, \theta_n, \phi_n, \theta_j, \phi_j, \dot{x}_{nj}, \dot{y}_{nj}, \dot{\theta}_{nj}, \dot{\phi}_{nj}]) \quad (12)$$

where $x_{nj} = x_n - x_j$ and $\dot{x}_{nj} = \dot{x}_n - \dot{x}_j$. y_{nj}, $\dot{y}_{nj}$, $\dot{\theta}_{nj}$ and $\dot{\phi}_{nj}$ are computed in a similar fashion.

3.2. Skeleton Restoration Model

To predict full-body motion, we recover the full-body pose from the predicted IPM states. This is divided into two steps as shown in Fig. 1. We first recover the lower body from the IPM state, then recover the upper body from both the IPM state and the recovered low body. There are two reasons for this design. First, the Skeleton-to-IPM mapping dictates that the IPM has a higher correlation with the lower body than with the upper body. Also, the dynamics of the lower body and the upper body are relatively independent [27, 49], *i.e.* similar low-body motions can correspond to different upper-body motions, *e.g.* different styles in walking. Therefore, we use two models to recover the lower body and the

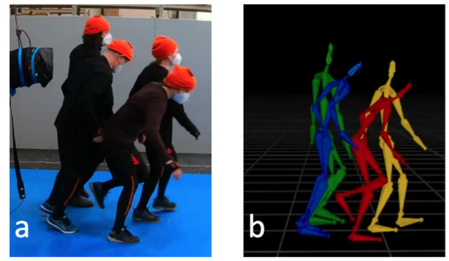

Figure 2. FZJ Push [12]. The blue agent was pushed by the punch bag and then he pushed other people.

upper body, respectively. Overall, although the Skeleton Restoration Model involves deep neural networks, the required data is small as there is strong IPM guidance.

Lower Body Restoration. We use a Conditional Variational Autoencoder (CVAE) [46, 49, 64] (CVAE-Lower in Fig. 1) to learn a Normal distribution of the lower body X_{t+1}^l in the latent space conditioned on X_t^l. During inference, since X_{t+1}^l is unavailable, we train a sampler (Lower Sampler) to sample the latent space to generate the next frame $\hat{X}_{t+1}^l$. The Lower Sampler network is an MLP. It takes as input X_t^l, I_{t+1}, and outputs a latent code of CVAE-Lower which is then decoded. Overall, CVAE-Lower takes as input the current lower body X_t^l and the predicted IPM state I_{t+1}, to predict the next lower body $\hat{X}_{t+1}^l$, essentially reconstructing the lower body under the IPM guidance.

Upper Body Restoration. Similarly, we also use a CVAE named CVAE-Upper, except this time we use both the lower body predicted by CVAE-Lower $\hat{X}_{t+1}^l$ and the current upper body X_t^u as the condition. A sampler (Upper Sampler) is also used to take as input I_{t+1}, $\hat{X}_{t+1}^l$ and X_t^u, and sample the latent space of CVAE-Upper, which is then decoded to predict the upper body at the next frame $\hat{X}_{t+1}^u$.

3.3. Training with Auxiliary Losses

In summary, the learnable parameters of our model include: the LSTM (Eq. (6)), the MLPs (Eq. (7), Eq. (12)), the ground friction coefficent μ, CVAE-Lower, CVAE-Upper, Lower Sampler and Upper Sampler. Other than the main loss in Eq. (2), we also use other auxiliary losses such as foot sliding, IPM state MSE, *etc*. We also pre-train some components for initialization. Due to space limit, all details including training/prediction algorithms, implementation details, parameters, code, data, *etc*. are in the SM.

4. Experiments

4.1. Dataset and Metrics

Data for our problem is extremely scarce compared with other human motion prediction research. The only publicly available dataset, to our best knowledge, is a new dataset

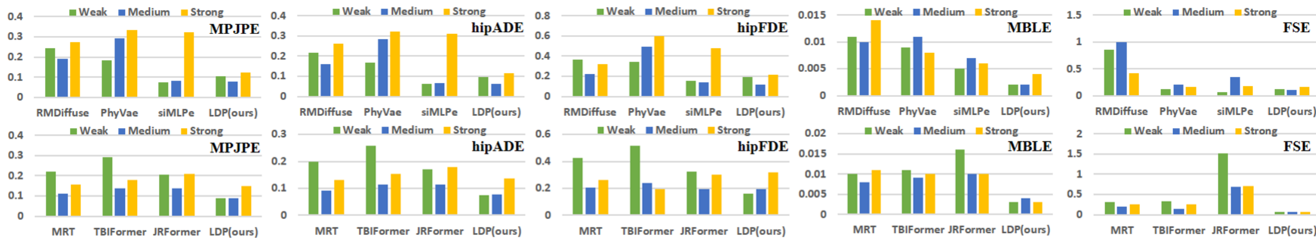

Figure 3. Perturbations with different magnitudes in single-person (top) and multi-people (bottom).

Method	MPJPE	hipADE	hipFDE	MBLE	FSE
A2M	0.403	0.386	0.730	0.019	0.200
ACTOR	0.362	0.338	0.591	0.020	0.434
MDM	0.500	0.424	0.686	0	2.567
RMDiffuse	0.228	0.202	0.299	0.011	0.790
PhyVae	0.260	0.249	0.460	0.009	0.170
siMLPe	0.130	0.117	0.226	0.006	0.182
EqMotion	0.296	0.270	0.543	0.064	1.552
Ours	0.097	0.086	0.171	0.002	0.131
MRT	0.162	0.140	0.282	0.010	0.256
DuMMF	0.312	0.285	0.480	0	3.194
TBIFormer	0.204	0.177	0.305	0.010	0.234
JRFormer	0.181	0.152	0.260	0.012	0.932
Ours	0.106	0.092	0.218	0.003	0.069

Table 1. Metrics in single-person (top) and multi-people (bottom).

[12] named FZJ Push. The dataset includes standing individuals, groups of four, and groups of five, with one person pushed by a punching bag unexpectedly and the push is propagated through the group. In total, the dataset includes only 45 single-person motions and 63 multi-people motions. This is considerably less than data normally used for human motion prediction. As shown later, the necessity of a model with high data efficiency is crucial. The motion is recorded at 60 Hz. Shown in Fig. 2 a, a hanging punch bag is operated by a person to give pushes of various magnitudes to one person in the group. Then the skeletal motions (Fig. 2 b) are recorded. There is a pressure sensor measuring the pushing forces on the punching bag. However, the pushing forces between people are not recorded. We discard redundant data such as frames in waiting.

For evaluation, we adopt five widely used metrics [49, 67, 69]: Mean Per Joint Position Error (MPJPE) in meters, Average Displacement Error at the hip (hipADE) in meters, Final Displacement Error at the hip (hipFDE) in meters, Mean Bone Length Error (MBLE) in meters, and Foot Skating Error (FSE) in centimeters. Details and justifications for these metrics are in the SM.

4.2. Baselines

There is no similar work in human motion prediction to our best knowledge, so we carefully review a wide spectrum of research in motion prediction, synthesis and generation, and choose the latest methods in each field for comparison. Specifically, we choose 11 baselines: A2M [16], ACTOR [46], MDM [51], RMDiffuse [73], PhyVae [64], siMLPe [17] and EqMotion [67] for the single-person scenario, and MRT [58], DuMMF [69], TBIFormer [44] and JRFormer [68] for the multi-people scenario. The specific adaptation varies according to the baseline, and we give the details in the SM. One notable difference is our model only requires the first frame with the perturbation force during inference, while the other methods tend to require much more information such as multiple frames.

4.3. Quantitative Results

The single-person comparison is shown in Tab. 1 top. Despite requiring the minimal information, our model still achieves the best performance on all metrics except the MBLE. MDM obtained 0 MBLE because its parameterization is joint angle based, *i.e.* no bone-length change incurred. A joint angle parameterization could also work with our model but in practice, we find a joint-position-based parameterization works better. Across different metrics, LDP outperforms the best baseline by as much as 25.38%, 26.50%, 24.34%, 66.67%, and 22.94% on MPJPE, hipADE, hipFDE, MBLE, and FSE respectively, excluding the MBLE of MDM. We tend to attribute the higher performance to the explicit physics-based inductive biases embedded in the design of LDP. Furthermore, we look into performances under perturbations with different magnitudes (weak, medium and strong) in Fig. 3 top, where we only include the best three baselines and leave the full comparison in SM. Stronger pushes lead to stronger responses and tend to be harder to predict. This is especially obvious in metrics related to motion tracking, *i.e.* MPJPE, hipADE and hipFDE, where as the push becomes stronger, the errors become larger. Comparatively, LDP consistently outperforms other baselines, demonstrating its effectiveness in strong perturbations. In addition, compared with weak and medium pushes, LDP has a slower error increment under

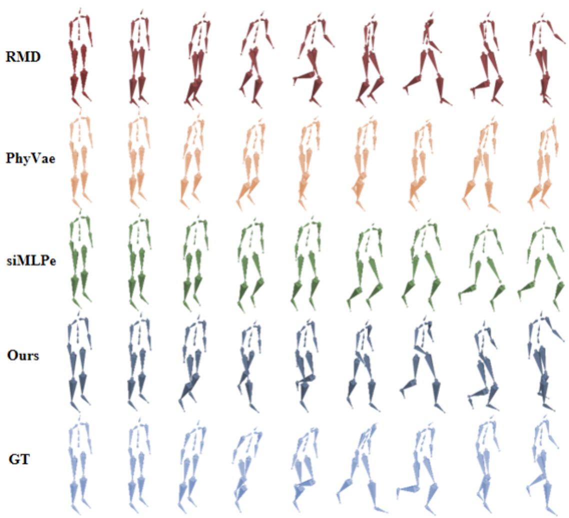

Figure 4. Visual Results in the Single-person scenario.

strong pushes, in contrast to the more volatile performances of other baselines, showing better generalizability. Overall, LDP either ranks as the best or is close to the top performance across metrics and perturbation levels.

The results under the multi-people scenario are shown in Tab. 1 bottom. The MBLE of DuMMF is 0 because it employs joint-angle-based parameterization. Multi-people is a challenging task for all methods. On all metrics, LDP outperforms all baselines by at least 34.57%, 34.29%, 16.15%, 70%, and 70.51% on MPJPE, hipADE, hipFDE, MBLE, and FSE, respectively, (excluding the MBLE of DuMMF). Moreover, we show detailed analysis under perturbations with different magnitudes in Fig. 3 bottom, with the three best baselines. One challenge in multi-people is to predict the onset and duration of interactions. The baseline methods need to learn the interactions by purely fitting the data, while our method learns them as a latent physical process. Consequently, none of the baselines can predict well, *e.g.* they predict moving without being pushed or not moving while being pushed, while our model can learn to predict the interactions and their propagation well. Overall, our model achieves or is close to the best performance across metrics and perturbation levels.

4.4. Qualitative Results

We visually compare our methods with the best three baselines under single-person in Fig. 4. Our prediction has the highest quality and is the most similar to the ground truth. RMDiffuse severely violates bone lengths, especially around ankles, and generates jittering motions. PhyVae predicts walking but with rather small steps. siMLPe predicts only a single step. The multi-people scenario is much harder (Fig. 5), where both individual reactions and interactions need to be predicted. MRT and TBIFormer suffer from serious intersections between individuals. JRFormer predicts merely subtle movements that deviate considerably from the ground truth. Our model generates the most similar prediction to the ground truth.

Method	MPJPE	hipADE	hipFDE	MBLE	FSE
no IPM, Full	0.217	0.195	0.341	0.007	0.196
no IPM, Low-up	0.206	0.184	0.320	0.009	0.313
IPM, Full	0.110	0.094	0.242	0.004	0.126
IPM, Low-up	0.106	0.092	0.218	0.003	0.069

Table 2. Ablation study with (1) IPM and no IPM, (2) Full body and Lower-up body pose reconstruction.

Explainability In Fig. 5 bottom, we show the learned net forces on the second person (from left), to provide plausible explanations of the predicted motion. This person remains still initially under zero net force, then experiences a push from the first person, resulting in forces in x and θ, and small forces in y and ϕ. Then the third person is pushed by the second, resulting in the change of the net force on the second person from positive to negative in x and θ. Finally, the second person recovers the balance. Our model predicts the motion results from plausible forces, and therefore possess strong explainability.

4.5. Generalization

LDP can easily generalize to out-of-distribution scenarios, *e.g.* unseen pushes, more people, different formations, *etc*. Since there is no ground truth, we show the visual result of a challenging generalization scenario in Fig. 6, where 13 people stand in a diamond formation and 3 of them indicated by the orange arrows are pushed. Note the data only contain up to 5 people in simple formations such as one or two lines. So this 13-people formation is totally out-of-distribution. However, our model can still generate plausible motions for the entire group, given only the initial poses and the perturbation forces, demonstrating strong generalizability. More experiments can be found in the SM.

4.6. Ablation Study

The Differentiable IPM and the Skeleton Restoration Model are two key components of our model. We conduct the ablation study to assess the effectiveness of them. There are four combinations: with/without IPM, and full-body restoration or separate restoration (first lower body then upper body). When the IPM is absent, the next frame is directly predicted by either one full-body CVAE (Full) or two CVAEs with one for the lower body and the other for the upper body (Low-up). Without IPM, there are also no samplers (Lower Sampler and Upper Sampler in Fig. 1) so we need to directly sample in the latent space of the CVAEs. We randomly sample the latent space 3 times when predicting the next frame

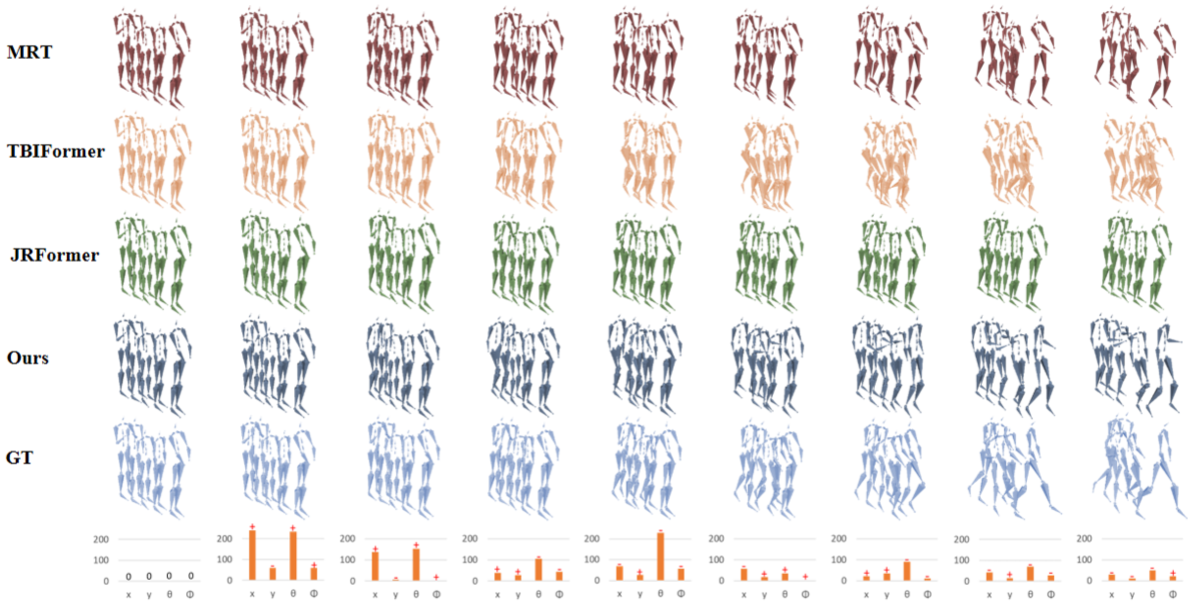

Figure 5. Multi-people comparison. The last row shows the learned net force on the second (from the left) person. The bar height indicates the magnitude and the sign indicates the direction, where the people move in the positive direction of the x-axis.

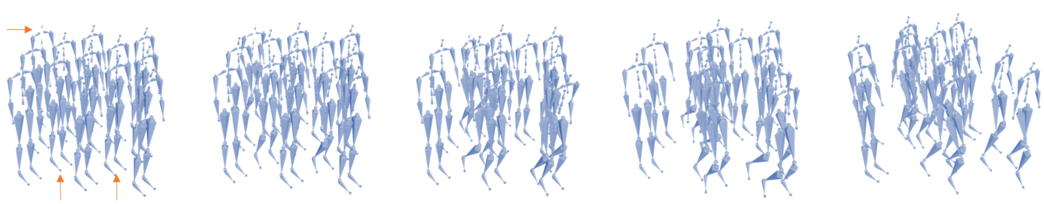

Figure 6. A 13-person group in a diamond formation with three people (indicated by orange arrows) being pushed.

and average the results. In contrast, with IPM, we can train the samplers to only sample once to predict the next frame.

Results are shown in Tab. 2. When there is no IPM, the performance deteriorates significantly across all metrics. With the IPM guidance, all metrics are significantly improved. Further, the Low-up separation of the body improves the performance further across all metrics under the IPM guidance, especially on the FSE. However, it exhibits limited effectiveness without the IPM guidance, even resulting in a bad FSE. This is because IPM states have strong correlations with the lower body, without which the Low-up is unable to improve the performance significantly even when the lower body is separately predicted.

5. Conclusion

We proposed a new task, human motion prediction under unexpected perturbation, which extends human motion prediction into new application domains. To this end, we have identified and overcome new challenges *e.g.* data scarcity and interaction modelling, by proposing a new class of deep learning models based on differentiable physics. Our model outperforms existing methods despite requiring far less information and shows strong generalization to unseen scenarios. One limitation is our method requires explicit modelling of the physical process, making the model not as general as black-box deep neural nets that can be plug-and-play on data. However, we argue this is mainly driven by the data scarcity. Also, it brings stronger generalizability and interpretability. In future, we will investigate more general physics models that can potentially accommodate more diversified physical interactions between people. A big difference between other existing datasets [21, 52] and the dataset FZJ Push is the former is active motions while the latter is passive balance recovery. We will also explore LDP on action motions in future.

Acknowledgements

The project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 899739 CrowdDNA.

References

[1] 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. 2

[2] Mazen Al Borno, Martin De Lasa, and Aaron Hertzmann. Trajectory optimization for full-body movements with complex contacts. *IEEE transactions on visualization and computer graphics*, 19(8):1405–1414, 2012. 3

[3] Okan Arikan, David A Forsyth, and James F O'Brien. Pushing people around. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 59–66, 2005. 1, 2

[4] Matthew A Brodie, Yoshiro Okubo, Daina L Sturnieks, and Stephen R Lord. Optimizing successful balance recovery from unexpected trips and slips. *Journal of Biomechanical Science and Engineering*, 13(4):17–00558, 2018. 1, 2

[5] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 226–242. Springer, 2020. 2

[6] Changkun Chen, Tong Lu, Weibing Jiao, and Congling Shi. An extended model for crowd evacuation considering crowding and stampede damage under the internal crushing. *Physica A: Statistical Mechanics and its Applications*, page 129002, 2023. 1, 2

[7] Siwei Chen, Xiao Ma, and Zhongwen Xu. Imitation learning via differentiable physics. *arXiv preprint arXiv:2206.04873*, 2022. 3

[8] Wenheng Chen, He Wang, Yi Yuan, Tianjia Shao, and Kun Zhou. Dynamic future net: Diversified human motion generation. In *ACM Multimedia (ACM MM)*, pages 2131–2139, 2020. 2

[9] Qiongjie Cui and Huaijiang Sun. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4801–4810, 2021. 2

[10] 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. 2

[11] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. *Advances in Neural Information Processing Systems*, 34:887–899, 2021. 3

[12] Sina Feldmann and Juliane Adrian. Forward propagation of a push through a row of people. *Safety science*, 164:106173, 2023. 5, 6

[13] 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. 2

[14] Thomas Geijtenbeek and Nicolas Pronost. Interactive character animation using simulated physics: A state-of-the-art review. In *Computer graphics forum*, pages 2492–2515. Wiley Online Library, 2012. 1

[15] Deshan Gong, Zhanxing Zhu, Andrew J Bulpitt, and He Wang. Fine-grained differentiable physics: a yarn-level model for fabrics. In *The International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 6

[17] 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, 6

[18] Elizabeth T Hsiao-Wecksler. Biomechanical and age-related differences in balance recovery using the tether-release method. *Journal of Electromyography and Kinesiology*, 18 (2):179–187, 2008. 1, 2

[19] Jaepyung Hwang, Jongmin Kim, Il Hong Suh, and Taesoo Kwon. Real-time locomotion controller using an inverted-pendulum-based abstract model. In *Computer Graphics Forum*, pages 287–296. Wiley Online Library, 2018. 2, 3

[20] A. Iollo, M. Ferlauto, and L. Zannetti. An aerodynamic optimization method based on the inverse problem adjoint equations. *Journal of Computational Physics*, 173(1):87–115, 2001. 3

[21] 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. 8

[22] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[23] Y. Jarny, M.N. Ozisik, and J.P. Bardon. A general optimization method using adjoint equation for solving multidimensional inverse heat conduction. *International Journal of Heat and Mass Transfer*, 34(11):2911–2919, 1991. 3

[24] Cengiz Kahraman, Muhammet Deveci, Eda Boltürk, and Seda Türk. Fuzzy controlled humanoid robots: A literature review. *Robotics and Autonomous Systems*, 134:103643, 2020. 1, 2

[25] Shuuji Kajita, Fumio Kanehiro, Kenji Kaneko, Kazuhito Yokoi, and Hirohisa Hirukawa. The 3d linear inverted pendulum mode: A simple modeling for a biped walking pattern generation. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, pages 239–246. IEEE, 2001. 2, 3

[26] Duško Katić and Miomir Vukobratović. Survey of intelligent control techniques for humanoid robots. *Journal of Intelligent and Robotic Systems*, 37:117–141, 2003. 1, 2

[27] Petar Kormushev, Dragomir N Nenchev, Sylvain Calinon, and Darwin G Caldwell. Upper-body kinesthetic teaching of a free-standing humanoid robot. In *2011 IEEE International Conference on Robotics and Automation*, pages 3970–3975. IEEE, 2011. 5

[28] Arthur D Kuo. An optimal control model for analyzing human postural balance. *IEEE transactions on biomedical engineering*, 42(1):87–101, 1995. 4

[29] Taesoo Kwon and Jessica K Hodgins. Momentum-mapped inverted pendulum models for controlling dynamic human motions. *ACM Transactions on Graphics (TOG)*, 36(1):1–14, 2017. 2, 3

[30] Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023. 3

[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. 2

[32] Dustin Li and Heike Vallery. Gyroscopic assistance for human balance. In *2012 12th IEEE International Workshop on Advanced Motion Control (AMC)*, pages 1–6. IEEE, 2012. 4

[33] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2020. 2

[34] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Multiscale spatio-temporal graph neural networks for 3d skeleton-based motion prediction. *IEEE Transactions on Image Processing*, 30:7760–7775, 2021. 2

[35] Junbang Liang, Ming Lin, and Vladlen Koltun. Differentiable cloth simulation for inverse problems. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[36] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 2

[37] Libin Liu, KangKang Yin, Michiel Van de Panne, Tianjia Shao, and Weiwei Xu. Sampling-based contact-rich motion control. In *ACM SIGGRAPH 2010 papers*, pages 1–10. 2010. 1

[38] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. In *Computer Graphics Forum*, pages 415–423. Wiley Online Library, 2015. 1

[39] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 2, 3

[40] Igor Mordatch, Emanuel Todorov, and Zoran Popović. Discovery of complex behaviors through contact-invariant optimization. *ACM Transactions on Graphics (ToG)*, 31(4):1–8, 2012. 1, 2

[41] Reza Olfati-Saber. *Nonlinear control of underactuated mechanical systems with application to robotics and aerospace vehicles*. PhD thesis, Massachusetts Institute of Technology, 2001. 2, 4

[42] Christian Ott, Maximo A Roa, and Gerd Hirzinger. Posture and balance control for biped robots based on contact force optimization. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pages 26–33. IEEE, 2011. 1, 2

[43] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2

[44] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17121–17130, 2023. 1, 2, 6

[45] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018. 2, 3

[46] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2, 5, 6

[47] Siyuan Shen, Yin Yang, Tianjia Shao, He Wang, Chenfanfu Jiang, Lei Lan, and Kun Zhou. High-order differentiable autoencoder for nonlinear model reduction. *ACM Transactions on Graphics (TOG)*, pages 1–15, 2021. 2

[48] Jacek Stodolka, Marian Golema, and Juliusz Migasiewicz. Balance maintenance in the upright body position: Analysis of autocorrelation. *Journal of Human Kinetics*, 50:45–52, 2016. 4

[49] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–10, 2022. 2, 5, 6

[50] Xiangjun Tang, Linjun Wu, He Wang, Bo Hu, Xu Gong, Yuchen Liao, Songnan Li, Qilong Kou, and Xiaogang Jin. Rsmt: Real-time stylized motion transition for characters. In *ACM SIGGRAPH*, pages 1–10, 2023.

[51] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 6

[52] 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. 8

[53] Chongyang Wang, Shunjiang Ni, and Wenguo Weng. Modeling human domino process based on interactions among individuals for understanding crowd disasters. *Physica A: Statistical Mechanics and its Applications*, 531:121781, 2019. 1

[54] He Wang, Kirill A Sidorov, Peter Sandilands, and Taku Komura. Harmonic parameterization by electrostatics. *ACM Transactions on Graphics (TOG)*, 2013. 2

[55] He Wang, Edmond SL Ho, and Taku Komura. An energy-driven motion planning method for two distant postures.

IEEE Transactions on Visualization and Computer Graphics (TVCG), 2014. 2

[56] He Wang, Edmond SL Ho, Hubert PH Shum, and Zhanxing Zhu. Spatio-temporal manifold learning for human motions via long-horizon modeling. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2019. 1

[57] Jack Wang, Aaron Hertzmann, and David J Fleet. Gaussian process dynamical models. *Advances in neural information processing systems*, 18, 2005. 2

[58] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 2, 6

[59] Kun Wang, Mridul Aanjaneya, and Kostas Bekris. Sim2sim evaluation of a novel data-efficient differentiable physics engine for tensegrity robots. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1694–1701. IEEE, 2021. 2, 3

[60] Xiaolin Wei, Jianyuan Min, and Jinxiang Chai. Physically valid statistical models for human motion generation. *ACM Transactions on Graphics (TOG)*, 30(3):1–10, 2011. 1, 2

[61] Keenon Werling, Dalton Omens, Jeongseok Lee, Ioannis Exarchos, and C. Karen Liu. Fast and feature-complete differentiable physics for articulated rigid bodies with contact. *CoRR*, abs/2103.16021, 2021. 3

[62] Steffen Wiewel, Moritz Becher, and Nils Thuerey. Latent space physics: Towards learning the temporal evolution of fluid flow. In *Computer graphics forum*, pages 71–82. Wiley Online Library, 2019. 2

[63] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 39(4):33–1, 2020. 1

[64] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 2, 5, 6

[65] Yujiang Xiang, Hyun-Joon Chung, Joo H Kim, Rajankumar Bhatt, Salam Rahmatalla, Jingzhou Yang, Timothy Marler, Jasbir S Arora, and Karim Abdel-Malek. Predictive dynamics: an optimization-based novel approach for human motion simulation. *Structural and Multidisciplinary Optimization*, 41:465–479, 2010. 3

[66] 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

[67] 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. 1, 2, 6

[68] Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. Joint-relation transformer for multi-person motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9816–9826, 2023. 1, 2, 6

[69] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 6

[70] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *The European Conference on Computer Vision (ECCV)*, pages 376–394, 2022. 3

[71] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory forecasting with explainable behavioral uncertainty. *arXiv preprint arXiv:2307.01817*, 2023. 2, 3

[72] Chuanqi Zang, Mingtao Pei, and Yu Kong. Few-shot human motion prediction via learning novel motion dynamics. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 846–852, 2021. 2

[73] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. *arXiv preprint arXiv:2304.01116*, 2023. 2, 6