

DROP: Dynamics Responses from Human Motion Prior and Projective Dynamics

YIFENG JIANG, Stanford University, United States of America

JUNGDMAM WON, Seoul National University, South Korea

YUTING YE, Meta Reality Labs Research, United States of America

C. KAREN LIU, Stanford University, United States of America

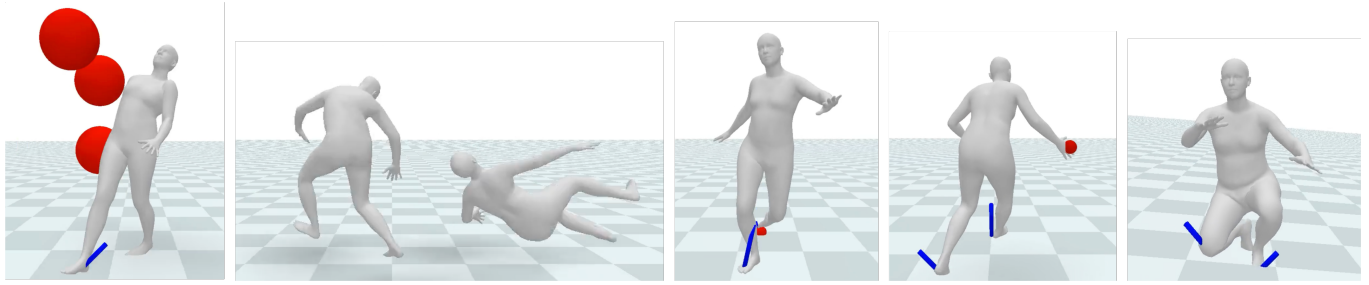


Fig. 1. DROP synthesizes dynamic reaction and recovery motion in response to a variety of perturbations. From left to right, responses to projectiles, two-character collision, tripped by an obstacle, target following, and responses to a tilting platform.

Synthesizing realistic human movements, dynamically responsive to the environment, is a long-standing objective in character animation, with applications in computer vision, sports, and healthcare, for motion prediction and data augmentation. Recent kinematics-based generative motion models offer impressive scalability in modeling extensive motion data, albeit without an interface to reason about and interact with physics. While simulator-in-the-loop learning approaches enable highly physically realistic behaviors, the challenges in training often affect scalability and adoption. We introduce DROP, a novel framework for modeling Dynamics Responses of humans using generative motion prior and Projective dynamics. DROP can be viewed as a highly stable, minimalist physics-based human simulator that interfaces with a kinematics-based generative motion prior. Utilizing projective dynamics, DROP allows flexible and simple integration of the learned motion prior as one of the projective energies, seamlessly incorporating control provided by the motion prior with Newtonian dynamics. Serving as a model-agnostic plug-in, DROP enables us to fully leverage recent advances in generative motion models for physics-based motion synthesis. We conduct extensive evaluations of our model across different motion tasks and various physical perturbations, demonstrating the scalability and diversity of responses.

CCS Concepts: • **Computing methodologies** → **Animation; Physical simulation**.

Authors' addresses: Yifeng Jiang, Stanford University, United States of America, yifengj@stanford.edu; Jungdam Won, Seoul National University, South Korea, jungdam@imo.snu.ac.kr; Yuting Ye, Meta Reality Labs Research, United States of America, yuting.ye@meta.com; C. Karen Liu, Stanford University, United States of America, karenliu@cs.stanford.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. 0730-0301/2023/12-ART \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

Additional Key Words and Phrases: **Generative Models, Physics-based Motion Synthesis, Hybrid Methods**

ACM Reference Format:

Yifeng Jiang, Jungdam Won, Yuting Ye, and C. Karen Liu. 2023. DROP: Dynamics Responses from Human Motion Prior and Projective Dynamics. *ACM Trans. Graph.* 1, 1 (December 2023), 12 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Creating virtual humans that can move autonomously and realistically stands as a key pursuit in character animation. Crucial to this endeavor is endowing these virtual characters with the ability to adapt to and dynamically interact with the environment. Characters imbued with such capabilities can enhance player immersion in games and AR/VR environments, provide a predictive human model for computer vision tasks, offer a safe and cost-efficient means to synthesize or augment data for robotics applications, and support healthcare and sports applications.

Recent advancements in character animation indicate that generative models, such as Variational Autoencoders (VAEs) or Diffusion Models ([Ling et al. 2020; Rempe et al. 2021; Tevet et al. 2023]), can efficiently learn human motion priors from large-scale mocap datasets. However, these learned kinematic models (motion priors) do not incorporate physics thus lack responses to perturbations in the environment. Previous attempts to incorporate physics to motion priors require either additional training of control policies using reinforcement or imitation learning, or re-training of motion priors with some notion of physics. Both approaches inevitably increase design and computational complexity, potentially hindering adoption and scalability of the methods. In contrast, this paper investigate the possibility to decouple the incorporation of physics from the training of kinematic models, exploiting scalability and modularity inherent from pre-trained generative models.

We introduce DROP, a motion-model-agnostic plug-in human physics simulator that adds dynamics capability to a pre-trained kinematic generative model, without the need to retrain the motion prior model or train another policy to control it. DROP simulates quasi-physics effects via an optimization-based implicit Euler integrator with projective dynamics, and incorporates them with an existing kinematic model (we use HuMoR [Rempe et al. 2021]). Our insight is that generative kinematic models inherently possess a capacity to generalize beyond the training distribution, particularly in recovering from out-of-distribution states, due to their design and training methodologies. However, unlike the recovery capabilities of simulator-in-the-loop methods, such as reinforcement learning, their recovery trajectories are oblivious to physics, making them often implausible and unreliable. Therefore, we develop a minimalist, physically plausible human simulator to collaborate with any autoregressive motion prior, directing its recovery trajectory towards physical validity, with no additional training.

Crucially, we employ projective dynamics [Bouaziz et al. 2014], an optimization-based (variational) simulation framework, to simulate our character which is modeled as a rigid stick figure aligning with the state representation of many kinematic models. Because an autoregressive motion prior provides a state-dependent manifold that models the probability distribution of the next state, we can seamlessly integrate the motion prior as one of the energy fields in the projective dynamics framework. It collaborates with other physics-based energy terms, such as contact and rigidity, to balance the equation of motion in the framework of variational integration. Hence, our work can also be viewed as an extension of example-based simulation [Jones et al. 2016; Martin et al. 2011], where the example manifold now represents state transitions instead of poses.

Our method opens avenues for various downstream applications that utilize motion priors, with minimal modifications and no further training. Changes to the physical properties of the skeleton, new environmental constraints, path or target tracking, can each be formulated as an additional energy term, while preserving existing projective energies. In our experiments, we present different plug-and-play examples to showcase the flexibility of our framework. Furthermore, we demonstrate that our model can generate diverse responses to environmental forces across various motion categories. Quantitative measures and ablation studies show that our method effectively regularizes the recovery of motion from perturbed, out-of-distribution states.

2 RELATED WORK

We review existing research in motion modeling, dynamic motion simulation, and hybrid methods that incorporate physics simulation with motion models.

2.1 Kinematic Motion Generation

Creating fluid and continuous motions for virtual characters is a core challenge in animation. The seminal work of Motion Graph [Arikan and Forsyth 2002; Kovar et al. 2002; Lee et al. 2002] arranges pre-recorded motion segments into a graph that can be traversed in runtime to produce seamless and extended motions. Although the results are high quality, the sparse graph connections limit

responsiveness. Subsequent methods such as Motion Fields [Lee et al. 2010] and Motion Matching [Clavet 2016] instead leverage k -nearest neighbors to determine the next pose for any given state, enabling instant adaptation to changes. These search-based algorithms unfortunately face difficulties when scaling to large, diverse, but unbalanced datasets such as AMASS [Mahmood et al. 2019].

Recent approaches use deep networks to model motion transitions from large datasets. Neural networks can accelerate the motion matching process [Holden et al. 2020], or automatically learn the best features for specific downstream tasks through regression models and supervised learning [Holden et al. 2017; Starke et al. 2022, 2019]. Generative models, such as (Conditional) Variational Autoencoders (VAE, CVAE) [Ling et al. 2020; Rempe et al. 2021], Generative Adversarial Networks (GAN) [Barsoum et al. 2018; Men et al. 2022], or Diffusion Models [Tevet et al. 2023], focus on learning the distribution of motion transitions to better represent the inherent variability in human motions. These models can generate the next state based on the predicted current state through sampling. In our work, we utilize a pre-trained generative model, HuMoR, but we enforce physical realism in the generated transitions and use projective dynamics to minimize deviation from the learned distribution. Our approach can synthesize novel connections amongst latent states, avoid divergence to unstable states, and guide the recovery back to the learned distribution when encountering unseen perturbations.

2.2 Dynamic Motion Simulation

Contrary to motion models that reproduce motions similar to those in the dataset, physics-based motion control naturally yields realistic dynamics responses to forces and changes in the environment. Techniques using optimization [da Silva et al. 2008; Muico et al. 2009; Ye and Liu 2010a] and reinforcement learning [Liu and Hodgins 2017; Peng et al. 2018] have proven effective in simulating a single example sequence under moderate perturbations. The source of these examples can be motion capture, hand animation, or outputs from motion synthesis [Bergamin et al. 2019; Park et al. 2019]. To expand the controller from a single example to a diverse motion dataset, methods akin to generative motion models are adopted in control policies. These include employing Motion Graphs with mixture-of-expert networks [Won et al. 2020], supervised policy learning with differentiable simulation [Fussell et al. 2021], using Conditional Variational Autoencoders (CVAEs) [Won et al. 2022; Yao et al. 2022], or using Generative Adversarial Networks (GANs) [Peng et al. 2022, 2021; Xu and Karamouzas 2021]. However, the complexity of the training process and computation time also scale with the dataset diversity. Additionally, these control policies tend to produce stiff physical responses [Xie et al. 2023], and creating human-like compliance may require tuning simulation parameters and modifying the training process for each different downstream task [Lee et al. 2022]. For realistic recovery behavior, specific controllers can be learned from captured recovery motions [Shiratori et al. 2009]. In contrast, our method serves as a decoupled add-on module to a generic pre-trained kinematic motion model, and does not require any task-specific learning of the motion model or a control policy.

2.3 Hybrid Simulation and Kinematic Methods

Closest to our method are techniques that enhance kinematic motions with physical and dynamics responses at the moment of interaction. Zordan and his colleagues [2002; 2005] blend simulated passive motions with kinematic motions at the onset and offset of interaction. It produces convincing responses but the transitions are not always realistic. Ye and Liu [2008] apply passive simulation in a torque space that does not interfere with the input motion, but their method cannot handle balance. Example data of people being pushed can be adapted to runtime perturbations. Arikan et al. [2005] use simulated motions to retrieve and modify the closest example. Similarly, Ye and Liu [2010b] learn a latent dynamic model from perturbation and recovery examples, and apply simulation to search in the latent space of recovery motions. Motion Field [Lee et al. 2010] also provides an add-on method for perturbation recovery. While effective, these methods are restricted by situations that can be captured safely and realistically. They also do not scale well to diverse dynamic situations. In comparison, our method does not require the motion dataset to contain examples of people being pushed or falling over, provided enough diversity in the data and we can effectively identify and connect relevant spaces within the data during interactions. Our method also does not switch in and out of simulation.

3 ALGORITHM OVERVIEW

We introduce a human motion simulator based on projective dynamics, an optimization-based (variational) implicit Euler integrator. Projective dynamics provides a flexible interface to incorporate the generative motion prior as a projective energy term. Optimization-based simulation provides stability under large time steps, which facilitates integrating existing motion priors operating at arbitrary frequencies. Using this framework, we first create a minimalist uncontrolled ragdoll model, adhering to basic human kinematic and dynamic invariants, such as contact, rigidity, and range of motion (RoM), with each invariant represented as an energy for the variational integrator.

The learned motion prior then serves as an additional energy term, effectively supplying control forces to the character. The choice of projective dynamics over other optimization integrator variants is crucial, as it permits the incorporation of any class of autoregressive generative motion prior, so long as its output can be represented as a state-dependent manifold. For a more detailed discussion on this design choice, see Sec. 9.1.

Further, we introduce a simple and effective correction technique to reduce the use of "magic force" from the motion prior energy term by regulating the linear and angular momentum of the center-of-mass (CoM).

4 MINIMAL CHARACTER SIMULATION WITH PROJECTIVE DYNAMICS

We begin by constructing an uncontrolled ragdoll simulator. As shown in Fig. 2, our simulation character consists of N mass particles, each positioned at a human joint (excluding hands, toes, and head) with rest-pose locations provided by the SMPL human model with an average body shape [Loper et al. 2015]. Massless sticks

connect the joint particles, modeled as strong springs similar to recent work [Chen et al. 2022] that extends the IPC [Li et al. 2020a] framework to articulated rigid-bodies. Consequently, the state $\mathbf{x}$ of our simulated ragdoll can be fully determined by joint particle locations, with $\mathbf{x} \in \mathbb{R}^{3N}$ and $N = 22$ according to SMPL.

Using an optimization-based (or variational) integrator [Gast et al. 2015; Liu et al. 2013; Martin et al. 2011], we solve for the next state $\mathbf{x}_{t+1}, \mathbf{v}_{t+1}$ from $\mathbf{x}_t, \mathbf{v}_t$, and external forces $\mathbf{f}_t$ by:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2h^2} (\mathbf{x} - \mathbf{y})^T \mathbf{M} (\mathbf{x} - \mathbf{y}) + \sum_i E_i(\mathbf{x}; \mathbf{x}_t, \mathbf{v}_t), \quad (1)$$

$$\mathbf{v}_{t+1} = (\mathbf{x}_{t+1} - \mathbf{x}_t)/h, \quad (2)$$

where $\mathbf{y} = \mathbf{x}_t + h\mathbf{v}_t + h^2\mathbf{M}^{-1}\mathbf{f}_t + h^2\mathbf{g}$, $\mathbf{g}$ represents the gravity vector, and h is the simulation step length (30Hz in this work to match the learned kinematics motion prior). The first term in Eq. 1, $E_{mom} = \frac{1}{2h^2} (\mathbf{x} - \mathbf{y})^T \mathbf{M} (\mathbf{x} - \mathbf{y})$, arises from integrating backward (implicit) Euler discretization over $\mathbf{x}$. This "momentum" energy term holds second order (Newtonian) dynamics, allowing environment to change the system states through forces $\mathbf{f}_t$. The $\{E_i\}$ comprise energy terms for contact, RoM, and rigidity constraints, representing invariants we want to adhere to regardless of whether our character is controlled.

The projective dynamics method alternates between slack variable projection and solving steps at each optimization iteration. The projection step independently constructs a slack variable for each E_i in Eq. 1 (termed "local" step), while the solving step considers all $\{E_i\}$ collectively (termed "global" step). In practice, the local projection step can employ any routine, even non-differentiable ones like sampling, which provides flexibility for our framework.

In the local step, we independently construct a manifold C_i for each E_i and define E_i as [Bouaziz et al. 2014]:

$$E_i(\mathbf{x}) = \frac{w_i}{2} \min_{\mathbf{p}_i \in C_i} \|\mathbf{A}_i \mathbf{x} - \mathbf{p}_i\|^2, \quad (3)$$

where C_i is the manifold when $\mathbf{A}_i \mathbf{x} \in C_i$ we have $E_i(\mathbf{x}) = 0$. Eq. (3) requires solving the slack variable $\mathbf{p}_i$ by projecting $\mathbf{A}_i \mathbf{x}$ onto C_i . For example, consider the case where $\mathbf{A}_i = \mathbf{I}$, and C_i contains only one constant point $\bar{\mathbf{x}}$. In this case, we can easily obtain the slack variable $\mathbf{p}_i$ to always be $\mathbf{p}_i = \bar{\mathbf{x}}$. Consequently, $E_i(\mathbf{x}) = \|\mathbf{x} - \bar{\mathbf{x}}\|^2$ essentially creates a spring force (∇E) that continually attracts the optimization variable $\mathbf{x}$ to the single-point manifold $\bar{\mathbf{x}}$.

In the global step, treating all solved $\mathbf{p}_i$'s as constants, $\sum_i E_i$ becomes a quadratic function. Finding the minimizer $\mathbf{x}^* = \operatorname{argmin} (E_{mom} + \sum_i E_i)$ requires solving the linear system of equations:

$$(\mathbf{M}/h^2 + \sum_i w_i \mathbf{A}_i^T \mathbf{A}_i) \mathbf{x} = \mathbf{M} \mathbf{y}/h^2 + \sum_i w_i \mathbf{A}_i^T \mathbf{p}_i \quad (4)$$

obtained from $\nabla E = 0$.

4.1 The Energy Terms

Rigidity. Following previous works, we approximate rigid bones [Chen et al. 2022] with strong springs [Liu et al. 2013]:

$$E_1 = \frac{w_1}{2} \min_{\mathbf{d} \in C_1} (\mathbf{x}^T \mathbf{L} \mathbf{x} - \mathbf{x}^T \mathbf{J} \mathbf{d}), \quad (5)$$

where C_1 is the manifold of rest-length spring directions, and $\mathbf{d}$ is the vectors of springs projected back to their rest length. For a detailed derivation of $\mathbf{L}$ and $\mathbf{J}$, we direct readers to [Liu et al. 2013].

Contact. To mitigate penetration, we utilize a simplistic contact model:

$$E_2 = \frac{w_2}{2} \min_{\mathbf{x}_c \in C_2} (\mathbf{x}^T \mathbf{S} \mathbf{x} - \mathbf{x}^T \mathbf{S} \mathbf{x}_c), \quad (6)$$

where $\mathbf{S} \in \mathbb{R}^{3N \times 3N}$ is a diagonal selector matrix. Joint particles $\mathbf{x}^j, j = 1, \dots, N$ that will be in collision with the environment trigger corresponding entries in $\mathbf{S}$ to 1. $\mathbf{x}_c$ is the projection of the current configuration $\mathbf{x}_t$, along contact normal directions, back to the nearest penetration-free configuration. Consequently, E_2 effectively prompts all particles that are on the verge of penetration to exit the collision boundary in the normal direction, yet remain static in the tangential direction. There is a static friction assumption commonly seen in human motion synthesis. While integrating a more advanced IPC-style penetration-free contact model [Lan et al. 2022] should improve our framework, we leave that for future work.

Range of Motion (RoM). We introduce another energy term to regulate the full body pose to remain within the joint ranges of motion:

$$E_3 = \frac{w_3}{2} \min_{\mathbf{x}_r \in C_3} \|\mathbf{x} - \mathbf{x}_r\|^2, \quad (7)$$

Here, $\mathbf{x}_r$ is the projection of the current pose back to the nearest pose within joint limits, with the routine involving a neural-net model, VPoser [Pavlakos et al. 2019]. We provide further details in Sec. 7.

5 CASTING MOTION PRIOR AS PROJECTIVE ENERGY

The kinematic motion prior serves as the "control" of the human character. The straightforward alternative for incorporating a motion prior is to train a model from massive motion data in the form of $E_{kin}(\tilde{\mathbf{x}}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1}) \in \mathbb{R}$ (or equivalently $E_{kin}(\tilde{\mathbf{x}}_{t+1}, \mathbf{x}_t, \mathbf{v}_t)$), where a realistic transition result in a low energy value and high energy vice versa, attracting the system state $\mathbf{x}_{t+1}$ back to the natural motion manifold. Training such differentiable energy functions as Energy-Based Models (EBM) [LeCun et al. 2006] is possible, but we opt for projective dynamics which allows us to directly take an existing motion prior not necessarily in the form of an EBM, without the need for retraining.

A general motion prior model typically takes the following form:

$$\tilde{\mathbf{x}}_{t+1} = \mathcal{M}(\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{z}). \quad (8)$$

This model aims to generate the next motion frame given the current state, an optional short motion history, and a noise vector $\mathbf{z}$ sampled from a predefined prior distribution (e.g., Gaussian). $\mathcal{M}$ is often a deep neural network. The model is considered "generative" since different sequences of $\mathbf{z}$ vectors would, over time steps, produce various motions, capturing the variability in natural human motion. We only consider *autoregressive* motion models, which predict one frame at a time, as we build online dynamic simulation and response, where changes in the environment (e.g., force perturbation) are unknown beforehand.

A generative model in such a form provides a straightforward way to approximate its transition manifold: if we sample a set of

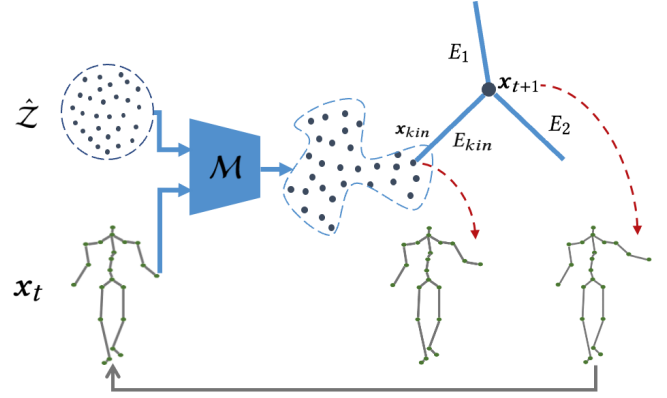


Fig. 2. DROP: casting motion prior as one of the projective energies. Physics-based energies $\{E_i\}$ drive the simulated character with projective dynamics. Given a motion prior (e.g. conditional VAE decoder), an additional data-driven energy E_{kin} is incorporated with the slack variable $\mathbf{x}_{kin}$ being the projection of the solver state $\mathbf{x}$ onto the motion prior manifold.

$\hat{\mathbf{z}}$, we can estimate the transition manifold of $\mathcal{X}_{t+1}$ conditioned on known history frames with the set of $\hat{\mathbf{x}}_{t+1}$ samples:

$$\hat{\mathcal{X}}_{t+1} := \{\hat{\mathbf{x}}_{t+1} | \hat{\mathbf{x}}_{t+1} = \mathcal{M}(\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \hat{\mathbf{z}}) \text{ and } \hat{\mathbf{z}} \sim \mathcal{N}_0\}, \quad (9)$$

where $\mathcal{N}_0$ is the prior distribution (we denote μ_0 as its mode). This then naturally fits into the projective dynamics framework with E_{kin} defined as:

$$E_{kin} = \frac{w_{kin}}{2} \min_{\mathbf{x}_{kin} \in \hat{\mathcal{X}}_{t+1}} \|\mathbf{x} - \mathbf{x}_{kin}\|^2, \quad (10)$$

where we obtain $\mathbf{x}_{kin}$ in the local step by simply searching in $\hat{\mathcal{X}}_{t+1}$ the nearest point to the current solver state $\mathbf{x}$.

In conjunction with the other physics-based energy terms, we present the main simulation loop (Algorithm 1, Fig. 2). At each simulation time step, we first batch-sample the motion prior to estimate $\hat{\mathcal{X}}_{t+1}$. Subsequently, we execute a few (3, in our implementation) projective dynamics iterations, where we alternate between solving for the projection slack variables $\mathbf{p}_i$ for each physics-based E_i , searching for $\mathbf{x}_{kin}$ for E_{kin} , and then finding the minimizer of (now quadratic) $E_{mom} + E_{kin} + \sum_i E_i$ with the solved $\mathbf{p}_i$'s and $\mathbf{x}_{kin}$ treated as known constants.

The success of our method depends on two hypotheses on the motion prior. First, E_{kin} from the motion prior should roughly agree with other energies $\{E_i\}$ when the transitions are in distribution of the training data. That is, they should all be small when simulating an in-distribution, unperturbed motion. This assumption is reasonable if we design the $\{E_i\}$ according to real-world physics and the motion prior is learned from real-world data.

Second, the motion prior has certain level of inherent ability to return to training distribution from a state that is out-of-distribution, albeit via nonphysical paths due to no data coverage. This assumption can be achieved because most autoregressive motion priors are trained by methods that enhance robustness, preventing error accumulation from causing significant drifts over long-sequence

generation. For instance, MotionVAE [Ling et al. 2020] demonstrated that scheduled sampling [Bengio et al. 2015] during training allows the model to recover to the training distribution over time. Likewise, robustness is inherent in Diffusion Models due to the large noise added during training.

Under these two assumptions, the main task of physics-based energies $\{E_i\}$ would be regulating the recovery path to be more physically valid. This is achieved by directing the solved state $\mathbf{x}$ to compromise between E_{kin} and $\{E_i\}$, as well as the momentum correction method introduced in the next section.

ALGORITHM 1: DROP Simulation Step

```

Input: Motion Prior  $\mathcal{M}$ , Current State  $\mathbf{x}_t, \mathbf{x}_{t-1}$ 
 $\hat{\mathbf{z}} \sim \mathcal{N}_0$ 
 $\hat{\mathbf{x}}_{t+1} = \mathcal{M}(\mathbf{x}_t, \mathbf{x}_{t-1}, \hat{\mathbf{z}})$ ,  $\mathbf{x}_{t+1} = \mathcal{M}(\mathbf{x}_t, \mathbf{x}_{t-1}, \mu_0)$ 
for  $l = 0, 1, \dots, \text{numProjectiveDynamicsIter}$  do
     $\mathbf{x}_{kin} = \text{argmin}_{\mathbf{x} \in \hat{\mathbf{x}}_{t+1}} \|\mathbf{x} - \mathbf{x}_{t+1}\|^2$ 
     $E_{kin} = \frac{w_{kin}}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_{kin}\|^2$  // Note,  $E_{kin} = 0$  when  $l = 0$ 
    Routines to calculate other projection variables  $\mathbf{d}, \mathbf{x}_c, \mathbf{x}_r$ 
    Solve  $\mathbf{x}_{t+1} = \text{argmin} E = \text{argmin}[E_{mom} + E_{kin} + \sum_i E_i]$ 
end
 $\mathbf{x}_{t+1} = \text{SoftCorrection}(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1})$ 

```

6 ROOT WRENCH METRIC AND SOFT CORRECTION

While a passive ragdoll simulated in maximum coordinates with " $E_{mom} + \sum_i E_i$ " adheres to physical validity, the introduction of E_{kin} can introduce non-physical forces into the system. One non-physical behavior of our controlled character that cannot be regulated by $\{E_i\}$, is the fictitious "puppet-like" root force and torque (wrench) to unrealistically regain balance, ignoring the under-actuated nature of human movement.

Before we can attempt to minimize the use of such "magical forces", we need to develop a metric capable of quantifying them. To reduce unnatural balancing, this metric should ensure that the movement of the center of mass (CoM) is feasible via contact forces with the ground. Given that E_{kin} can generate large non-physical forces and our contact model is very simplified, we cannot obtain meaningful contact forces directly from the simulation. Rather, we shall solve another optimization problem to find *any* contact force that can explain, to the greatest extent possible, the current changes in CoM momentum. It is important to note that these contact forces are *not* simulated within the system but are used solely as slack variables in this metric. We borrow such a metric from previous works [Zhang et al. 2021; Zheng and Yamane 2013]:

$$\min_{\mathbf{c}} \|\mathbf{F}\mathbf{c} - \mathbf{b}\|^2, \text{ subject to } \mathbf{c} \geq \mathbf{0}, \quad (11)$$

Here, $\mathbf{F}$ is the contact force basis constructed from the Coulomb friction cone assumption, $\mathbf{c}$ are the unilateral contact force coefficients for each basis, and $\mathbf{b} = (\dot{\mathbf{P}}, \dot{\mathbf{L}})$ is the change in linear and angular momentum around CoM, estimated using finite-differencing and subtracting the effect from gravity and external environmental forces. The minimization is solved with non-negative least squares (NNLS). We direct readers to [Zhang et al. 2021] for technical details.

Once $\mathbf{c}$ is solved, we can interpret $\|\mathbf{b} - \mathbf{F}\mathbf{c}\|^2$ as another quadratic energy term where $\mathbf{F}\mathbf{c}$ is constant ($\mathbf{F}$ can be assumed to only depend on $\mathbf{x}_t$ and not the $\mathbf{x}$ to be solved). Specifically, $\mathbf{b}$ is an affine function of $\mathbf{x}$, given that:

$$\dot{\mathbf{P}} = \sum_j m^j \mathbf{a}^j = \frac{1}{h} \sum_j m^j \left(\frac{\mathbf{x}^j - \mathbf{x}_t^j}{h} - \mathbf{v}_t^j \right), \quad (12)$$

$$\dot{\mathbf{L}} = \sum_j m^j \mathbf{r}^j \times \mathbf{a}^j = \frac{1}{h} \sum_k m^j [\mathbf{r}^j]_{\times} \left(\frac{\mathbf{x}^j - \mathbf{x}_t^j}{h} - \mathbf{v}_t^j \right). \quad (13)$$

In these equations, m^j represents the mass of the j -th joint particle, $\mathbf{a}^j$ denotes its acceleration, and $\mathbf{r}^j$ is its distance to the center of mass (CoM) location (moment arm), which remains constant if evaluated, similarly to $\mathbf{F}$, at the previous step $\mathbf{x}_t$. The skew-symmetric matrix is denoted by $[\mathbf{r}^j]_{\times}$. Here, central difference method is used for approximating $\mathbf{a}^j$. These equations enable us to write $\|\mathbf{b} - \mathbf{F}\mathbf{c}\|^2$ succinctly as $\|\mathbf{B}\mathbf{x} - \mathbf{a} - \mathbf{F}\mathbf{c}\|^2$, where $\mathbf{B} \in \mathbb{R}^{6 \times 3N}$ and $\mathbf{a} \in \mathbb{R}^6$ can be determined through manipulation of the equations above.

One might be inclined to directly incorporate $\|\mathbf{B}\mathbf{x} - \mathbf{a} - \mathbf{F}\mathbf{c}\|^2$ as an extra energy term in E during projective dynamics iterations, where solving $\mathbf{c}$ using non-negative least squares (NNLS) would serve as the projection step. However, we found this approach to be unstable in practice. In early (e.g. first) projective dynamics iterations, the solved $\mathbf{c}$ can be significantly off when $\mathbf{x}$ is still far from the final solution. This could occasionally lead to an excessively large update in $\mathbf{x}$, given that there are infinitely many $\mathbf{x}$ that can minimize $\|\mathbf{B}\mathbf{x} - \mathbf{a} - \mathbf{F}\mathbf{c}\|^2$. Instead, we choose to apply a post-processing correction, so we only evaluate $\mathbf{c}$ when $\mathbf{x}$ has stabilized to solution once all iterations are complete. Specifically, we solve for:

$$\mathbf{x}_{t+1}, \cdot = \text{argmin}_{\mathbf{x}, \epsilon} \|\mathbf{x} - \mathbf{x}_{pd}\|^2 + w_p \|\epsilon\|^2, \text{ subject to } \mathbf{B}\mathbf{x} - \mathbf{a} - \mathbf{F}\mathbf{c} = \epsilon, \quad (14)$$

where $\mathbf{x}_{pd}$ represents the solution $\mathbf{x}$ obtained after all projective dynamics iterations, w_p stands for the strength of the correction, and ϵ is the slack variable for correction tolerance. $\mathbf{c}$ here is first solved from Eq. 11 by evaluating $\mathbf{b}$ on $\mathbf{x}_{pd}$ (i.e., $\mathbf{b} = \mathbf{B}\mathbf{x}_{pd} - \mathbf{a}$). Eq. 14 can be solved precisely by writing out the KKT matrix. We found that this approach is more stable than directly solving the unconstrained problem $\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}_{pd}\|^2 + w_p \|\mathbf{B}\mathbf{x} - \mathbf{a} - \mathbf{F}\mathbf{c}\|^2$, which requires adding the low-rank $\mathbf{B}^T \mathbf{B}$ to the system matrix.

It is crucial for such a correction to be gentle, since the root wrench metric $\|\mathbf{b} - \mathbf{F}\mathbf{c}\|$ involves finite differencing for the second derivative and can be fairly imprecise, particularly for highly dynamic motions. We adjust w_p to prevent an excessively large correction $\|\mathbf{x} - \mathbf{x}_{pd}\|$. Conversely, to prevent minor corrections from accumulating over time and degrading the naturalness of the motion, we skip the post-correction when $\|\mathbf{b} - \mathbf{F}\mathbf{c}\|$ is already small.

7 IMPLEMENTATIONS

In our algorithm, the kinematic energy term, E_{kin} , is based on a recent autoregressive generative model, HuMoR [Rempe et al. 2021]. HuMoR is trained on the large-scale AMASS dataset [Mahmood et al. 2019], excluding motions involving terrains. When the next

step joint positions $\mathbf{x}_{t+1}$ have been solved, the input to HuMoR is replaced with this solved position before querying HuMoR again for sampling the transition manifold at the next time step. HuMoR, along with many kinematic models, also considers joint rotations $\mathbf{q}$ as input, in addition to joint positions. Therefore, before querying HuMoR again, we need to determine joint rotations $\mathbf{q}_{t+1}$ from the solved joint positions $\mathbf{x}_{t+1}$. In this work, we adopt a swing-twist decomposition strategy [Li et al. 2021], where the swing components of the joint rotations are completely determined by the bone vectors calculated from $\mathbf{x}$. As for the twist component, during the construction of the approximate manifold, the collection of samples $\hat{\mathbf{X}}_{t+1}$ are paired with corresponding $\hat{\mathbf{Q}}_{t+1}$ samples also predicted by the HuMoR model. We adopt the twist given by $\mathbf{q}_{kin}$, which is the corresponding $\mathbf{q}$ paired with the selected sample $\mathbf{x}_{kin}$ in the final projective dynamics iteration. For the leaf body nodes (head, hands, toes), whose rotations $\mathbf{x}$ does not inform, we directly prompt HuMoR to predict (so they might appear weird in out-of-distribution cases). We add a kinematic spring to damp leaf node rotations toward the identity (rest pose). Finally, although the HuMoR model also redundantly predicts joint Cartesian velocities $\mathbf{v}$ and autoregressively uses them, different from $\mathbf{x}$, we choose not to overwrite $\mathbf{v}$ using solved $\mathbf{x}$ with finite-differencing, as we found overwriting $\mathbf{v}$ can destabilize HuMoR's output.

For range-of-motion energy E_3 , the VPoser model operates on $\mathbf{q}$ instead of $\mathbf{x}$, so we adopt the same swing-twist decomposition inverse kinematics. The VPoser model is a VAE with encoder represented as $\mathbf{z} = E(\mathbf{q})$ and decoder $\mathbf{q} = D(\mathbf{z})$. Following previous works [Rempe et al. 2021; Weng et al. 2023], $\|\mathbf{z}\|$ can be considered as the validity of the pose, and we heuristically define $\|\mathbf{z}\| < 6.0$ as the range of motion boundary. Thus, the projection $\mathbf{q}_r$ of a pose $\mathbf{q}$ back to the range of motion is as follows:

$$\mathbf{q}_r = \begin{cases} D(E(\mathbf{q})/\|\mathbf{z}\| \cdot 6.0), & \text{if } \|\mathbf{z}\| > 6.0, \\ \mathbf{q}, & \text{otherwise,} \end{cases} \quad (15)$$

after which we obtain $\mathbf{x}_r$ for Eq. 7 through forward kinematics.

Our current prototype does not utilize the detailed skinned geometry of SMPL for collision detection and contact resolution. Instead, we simply position manually-scaled spheres at the joints to approximate the collision shape. The weights for the energy terms do not have a significant impact on the overall quality of the motion. We set w_1, w_2, w_3, w_{kin} to 1.0, 3.0, 3.0, 3.0 respectively. We solve projective dynamics for three iterations at each simulation step. We simply set $\mathbf{M}$ to $\mathbf{I}$, assuming all joint mass particles have unit mass. While we can craft a more realistic $\mathbf{M}$, note that it will not affect the data-driven control term E_{kin} since it operates on positions.

Our current unoptimized Python implementation of the method operates at 10 FPS, which is thrice slower than real-time, given that the system simulates with a time step of 1/30 second. The performance bottleneck arises from the swing-twist IK (at each projective dynamics iteration for the RoM energy) and data transfer between CPU and GPU (for the neural-nets). Since our system matrix is small ($3N = 66$), optimized full-GPU code should most likely bring the performance beyond real-time. Notably, the sampling process for the kinematic energy can be easily batched on the GPU. Our code will be open-sourced upon publication.

8 EVALUATIONS

Our experiments are designed to demonstrate three main objectives: (1) Our system's capability to generate a diverse array of responses to dynamic changes in a variety of motion categories; (2) The ease of motion editing and a variety of downstream applications enabled by our system; (3) The importance of the components proposed in our method.

All demonstrations in the supplementary video are generated by the existing HuMoR model combined with our human simulator. No additional motion control or planning are used, except for the modifications on energy terms and forces as detailed below. As such, the character's responses and recovery motions to unexpected events are stochastically created due to the generative nature of the HuMoR model.

8.1 Diverse Responses Across Various Motion Categories

In the basic setting, the only environmental variation comes from an external perturbation force $\mathbf{f}$ (Eq. 1), which is applied either to the torso, foot, or the hand (Fig. 3). Our supplementary video illustrates the varied responses of our simulated human to both smaller (Video 1'01) and larger (Video 1'33) forces. We randomly initialize our system from poses within the CMU dataset [cmu 2003] to encompass a broad range of motion categories.

Our method can also model responses to perturbations when performing highly dynamic tasks, such as back flips (Video 5'15). In this specific demo, the system is asked to initially track a specific CMU motion until just before the perturbation occurs. In other words, initially, the system follows a mock motion prior that is deterministic and consistently transitions to the next frame in the MoCap database, only starting to track the actual motion prior (HuMoR) immediately after the onset of the external force. We employ this workaround since if we let HuMoR progresses randomly on its own without guidance, it tends to quickly converge to the motion categories that comprise the majority of its unstructured and unbalanced training data, i.e. locomotion.

8.2 Downstream Motion Editing

In our supplementary video, we showcase an array of downstream tasks that can be accomplished by our system in a straightforward manner.

Projectiles. We can enable our character to physically respond to projectile balls thrown at it by setting the external forces to be proportional to the depth of penetration between the projectiles and the collision spheres (Sec. 7) on the character body. This means that $\mathbf{f} = -\sum_j k \mathbf{d}_j$, where $\mathbf{d}_j$ is the penetration vector between the projectile sphere and the joint collision sphere, with this summed over all joints. The projectiles are then simulated according to Newton's second law with $-\mathbf{f}$ (Video 0'29, 2'19).

Two-character collision. While it is possible to solve for two characters in one system, for ease of prototyping, we can also simulate each character independently, treating the other character's body-collision spheres as if they were collidable objects in the environment. Since both characters are based off stochastic motion priors, a variety of interesting behaviors arise when two characters

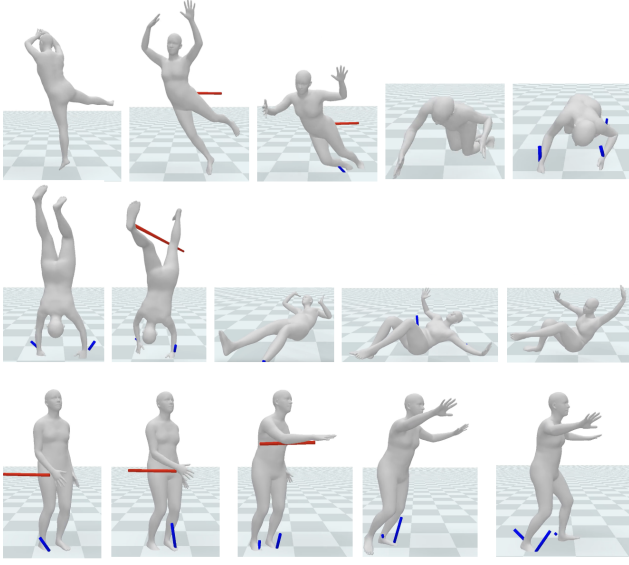


Fig. 3. Our method can model responses to perturbations at different body parts. Red lines are pushing forces and blue lines represent contact forces calculated during root wrench correction.

come closer to each other (Video 4’12). When approaching with a low speed, the characters try to dodge each other upon collision or attempt to taunt the other character out of way. When running fast into each other, we observe in one case where one character falls due to collision, but the other character regain balance and quickly runs away, leaving the first character struggling on the ground.

Direction following. Since the rollout of a motion prior will only generate random movements, previous generative kinematic models [Ling et al. 2020] have shown that a high-level DRL policy can be trained to guide the kinematic generative model to follow a specific path or direction. With our framework, basic and greedy direction tracking can be achieved zero-shot by adding a center-of-mass-velocity energy, without the need for training an additional model. Specifically, we add the following energy to E :

$$E_{com}(\mathbf{x}) = \left\| \sum_j m^j \frac{\mathbf{x}^j - \mathbf{x}_t^j}{h} - \mathbf{v}_{tar} \sum_j m^j \right\|^2, \quad (16)$$

where $\mathbf{v}_{tar}$ is the target velocity, m^j and $\mathbf{x}^j$ are j -th joint’s mass and position, same as in Eq. 12. E_{com} can also be considered as an additional soft physical constraint, that encourages E_{kin} to project to poses in the transition manifold that align more closely with $\mathbf{v}_{tar}$. In the supplementary video (3’28), we make the character follow a moving red dot that changes the moving direction every 4 seconds. When the red dot moves at a slow speed ($\|\mathbf{v}_{tar}\| = 1.2$ m/s), the character walks and turns accordingly. When we speed up the red dot ($\|\mathbf{v}_{tar}\| = 2.5$ m/s), the character switches to running so it can catch up the target. Using the same technique, in the projectile demos (Video 0’29, 2’19), we also ask the character to attempt to

dodge the projectile balls thrown at it by setting the direction of $\mathbf{v}_{tar}$ to be maximally distant to the directions of the balls.

Force-based constraints. One way to set positional constraints, such as a target in space for the right hand, is to add an external force $\mathbf{f}$ to the target joint (e.g., right hand), proportional to the constraint violation vector, pulling the target joint towards the position constraint. In the previous direction following demo (Video 3’28), when the character is asked to walk or run along a square trajectory (CoM tracking), we also add such a force-based constraint to its right hand, mimicking the character being led to walk or run. As shown in the video, this positional constraint is more closely satisfied at a lower center-of-mass speed (walking) and is violated more at a high speed (running).

Energy-based constraints. We also experimented with setting positional constraints by adding a local energy term that penalizes the distance $E_{constr} = \|\mathbf{x}^j - \mathbf{x}_{tar}^j\|^2$ between the joint and its target position. In our video (1’58), we emulate a character getting tripped by immovable obstacles on the ground by fixing the heel joint for 1 second when it is about to touch an obstacle on the ground.

Tilting platform. Tilting the ground plane does not necessitate the addition of energy terms to the system. The character’s changes in its behaviors are automatic due to unexpected collision with the ground. As the ground tilts, the character stops walking, crouches down, and sticks its arms out for balance (Video 2’41). This emergent behavior is entirely generated by the generative model, showcasing HuMoR’s inherent generalization ability as its training data only contains motions on a flat ground.

Stiff knee. Our method is also capable of modeling dynamics variations intrinsic to the character’s skeleton model. By adding a strong spring between the left hip joint and the left heel joint, similar to the ones we used to model bones, we can simulate a walking motion with a stiff knee (Video 4’03).

8.3 Ablation Study

To demonstrate the significance of the components proposed in our method for generating realistic response motions, we performed the following ablation studies (Video 4’41). Without HuMoR, the character would behave like a passive rag doll without control. Without the range of motion energy E_3 , the system is prone to generating unnatural poses that fall outside of the typical human range of motion. Without the collision energy term E_2 , the character might interpenetrate with the ground during perturbations and recoveries. Without the soft correction for root wrench, the character uses puppet-like ghost force to recover from large perturbations without utilizing contact.

8.4 Analysis on the Energy Trajectory

Fig. 6 displays a synthesized motion wherein during normal walking, the character is perturbed by an external force at the pelvis (represented by the red line in the rendering), takes a few recovery steps, and returns to normal walking. During this trajectory, we note that at the onset of perturbation, the RoM and Momentum energies experience increased violation, making the solved motion

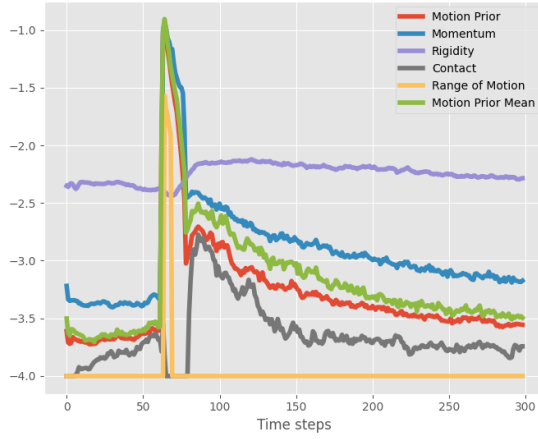


Fig. 4. Median trajectory of different energy terms in our method, across 250 random motions all being perturbed from steps (frames) 63 to 78. Our method guides the motions to return to training data distribution over time, as evidenced by decreasing motion prior energy.

to deviate from the distribution of motion prior and driving E_{kin} to increase as well. Subsequently, during the recovery steps, E_{kin} remains relatively high, suggesting that the motion prior has not yet returned to the AMASS training distribution. The Contact energy is high here to prevent the recovery steps from appearing implausible (i.e., penetrating the floor too much).

To illustrate that our system can guide the motion prior's recovery back towards the original training data distribution, we extend Fig. 6 to analyzing 250 random motions, all perturbed at the same point of time in the trajectory (Fig. 4). A logarithmic y-scale is used for visual clarity, and the medians per frame across the 250 motions are plotted. Similar to Fig. 6, Fig. 4 shows that during the unperturbed stage, the kinematic energy largely aligns with physics-based energy terms, and hence all energy levels are low. Both kinematic and other energy terms increase at the point of perturbation, causing the motion to deviate from the motion prior as a compromise between the energies. The median energies gradually decrease, reflecting that it takes time for motions to recover to training distribution (e.g. get up and resume normal walking). We also plot a hypothetical energy term, "motion prior mean" (green line in Fig. 4), which represents the kinematic energy level if E_{kin} was calculated from a single-point manifold, as given by the mean/mode of the HuMoR's decoder, instead of having access to a full manifold. The results demonstrate that, especially during recovery (around frames 100 to 150), using the full manifold reduces E_{kin} by at least 30% (note the logarithmic scale), indicating that the system prefers having the flexibility of choosing from a manifold over a single point.

Finally, for an ablation system without the Contact and RoM energy terms (we retain the Rigidity energy, effectively allowing the generalized coordinate representation to always hold), Fig. 5 shows that both kinematic and physics-based energies are at least 5

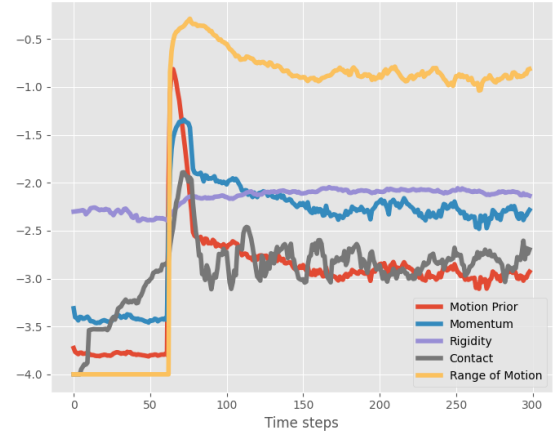


Fig. 5. Median trajectory of different energy terms in an ablation method without RoM and contact energies being used in simulation, across 250 random motions all being perturbed from steps (frames) 63 to 78. Final energies of both motion prior term and the physics terms are at least 5x higher than our full method, and show no signs of further decrease.

times higher at the end of the trajectory and do not demonstrate a trend towards final recovery.

8.5 Analysis on Root Wrench Metric and Correction

In Sec. 6, we introduced a *soft* correction mechanism to regulate the usage of the "root magic force". Given that the measurement of this metric involves twice finite differencing, it can be numerically sensitive. Fig. 7 Top provides evidence that even for ground-truth mocap motions, such as jumping and running, the root wrench metric can have a non-negligible value. Intuitively, even a small inaccuracy in MoCap data can have a substantial impact on the metric, because finite differencing involves multiplying by $1/h^2$.

We also demonstrate that our simple inertia approximation using mass points at joints does not significantly affect the estimation of the root wrench metric. This can be observed by comparing it with a more detailed inertia approximation that treats all body parts as manually-scaled cylinders (scaled to roughly match the skinned SMPL body shape [Loper et al. 2015]). As can be seen in Fig. 7 Top-left and Top-right, the cylindrical approximation yields similar root wrench metric for ground-truth motions.

In Fig. 7 Bottom, we present an illustrative example of how our soft correction method reduces the root wrench metric for a synthesized motion of perturbed running. During this motion, the red line represents what the root metric would have been without our correction mechanism, while the blue line shows the actual value for this motion, post-correction. Our method only brings down the metric to a similar range as seen in ground-truth motions, instead of completely eliminating it, as the metric is not exact. As such, some motions that visually use non-physical root forces, but has the metric in range, will not be corrected (Sec. 9.2).

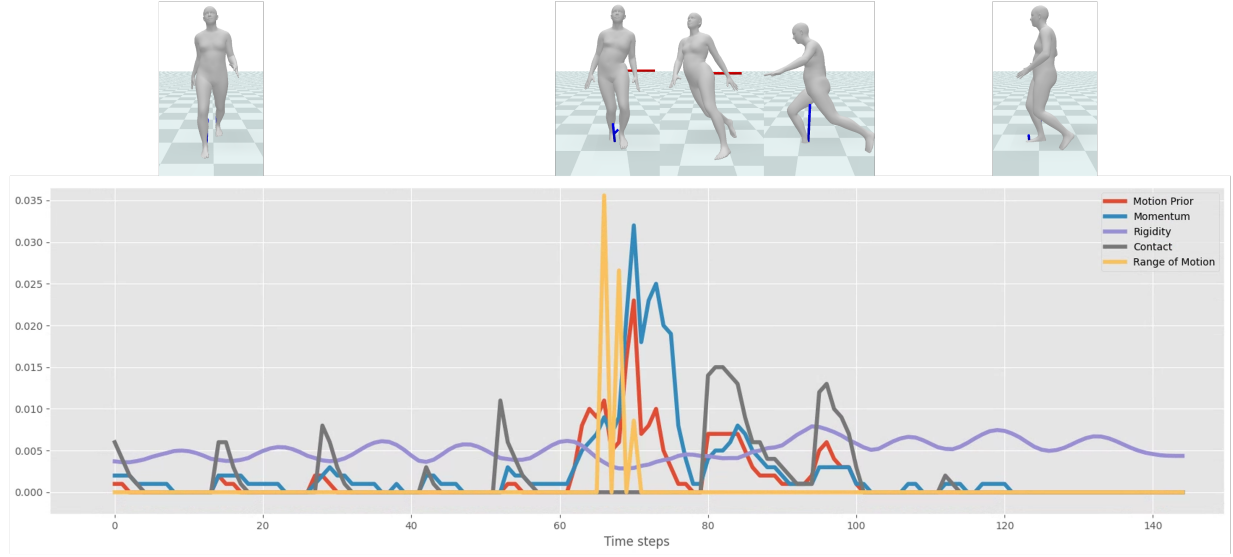


Fig. 6. Contributions of each energy term during synthesizing a motion of perturbed walking. External force happens at frame 63. At each step, the solve state compromises between motion prior and other energies, regulating the recovery of motion prior back to training data distribution to be physically valid.

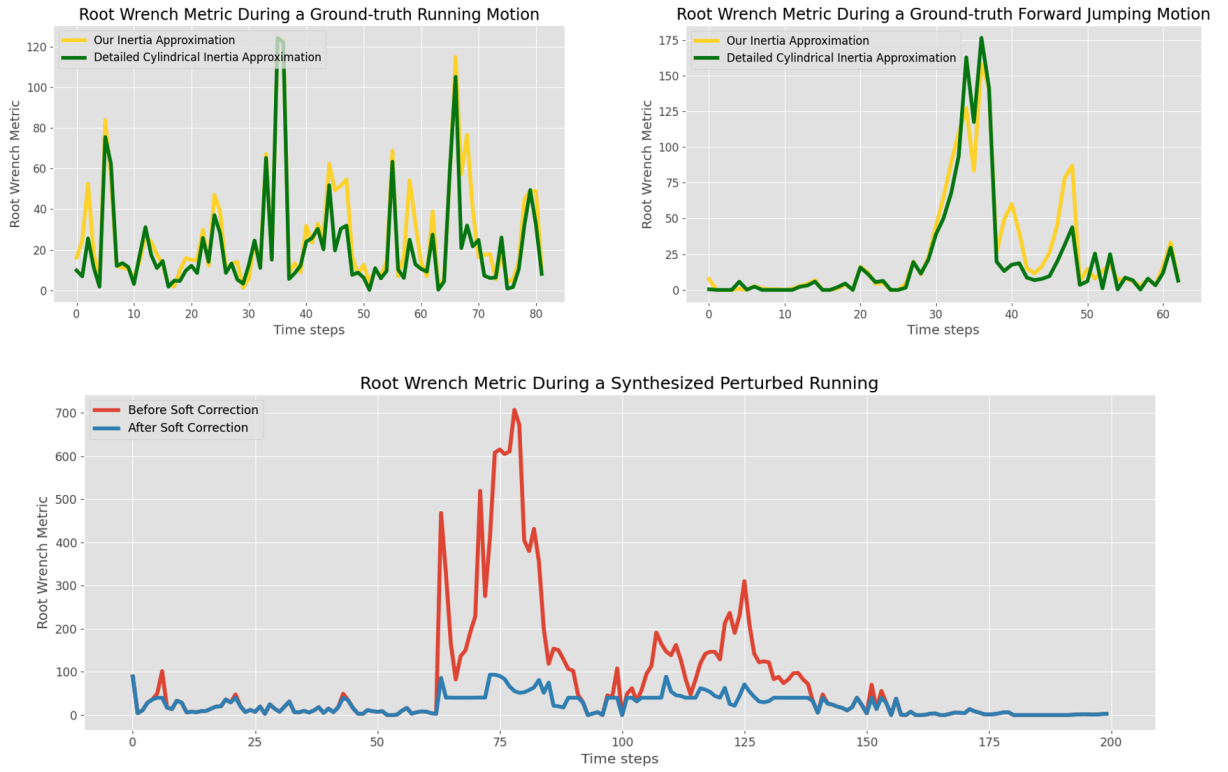


Fig. 7. Up: Root wrench usage metric is non-negligible even during ground-truth MoCap trajectories, shown running (left) and jumping (right) motions. Bottom: Our soft correction technique aims to bring the metric down to a similar range as ground-truth motions.

8.6 Comparisons with Reinforcement-Learning-Based Methods

We explore an alternative framework to generate physics-based human motion without using deep reinforcement learning (DRL). By decoupling the training to be only kinematics, our method scales to large datasets much more easily and with less compute. In the accompanying video (0'48, 1'20), we present a few non-rigorous qualitative comparisons to the state-of-the-art DRL method ASE [Peng et al. 2022]. We focus on the responses of the two methods to small and large external forces, since when without perturbation our method generates motions with quality on-par to HuMoR. Though ASE is trained on a much smaller dataset which specifically contains perturb and get-up motions, our method can generate more compliant and realistic (in terms of physical capabilities of real humans) responses. Our method also generates more diverse responses thanks to being trained on the large-scale AMASS data. On the other hand, motions from ASE can sometimes be more physical than ours since it is trained in a full-blown physics engine.

8.7 Failure Cases

It is challenging to define failure quantitatively for this work, as the kinematic model will continue to drive the character towards random poses even if it enters a unstable state. Also, with our physics-based energies in similar forms with common physicality metrics, quantitative physical metrics could appear valid even in obvious failure cases from visual inspection. Our video (5'43) showcases several such qualitative failure cases, where the most common issue can be characterized as "super-human behavior". For example, the character could quickly spin near the floor while lying. In such cases, the motion prior clearly enters a disastrous state from which it cannot recover even with guidance of the system. However, since the character still makes intermittent contact with the floor, the motion would be deemed valid by the soft correction routine. This behavior suggests that additional physical constraints, such as torque limits, may need to be added to the system, since such fast spinning indicates an effectively huge amount of control torque being added to the system. We provide more discussion on current limitations in Sec. 9.2.

9 DISCUSSION

This paper introduces a minimal human physics simulator based on projective dynamics, that can be plugged into a learned generative kinematics motion prior to augment it with dynamic capabilities, without any further training beyond the pretrained motion prior. Thanks to the motion diversity inherited from the motion prior, simulating various dynamics responses to physical and environmental changes is surprisingly easy.

9.1 Alternative Formulations

Our current framework poses no constraint on the type of generative motion models, as long as they can be sampled to generate a set of next states for manifold approximation. We note that Eq. (10) cannot be used in vanilla optimization-based simulators as an energy term, due to construction of $\mathcal{X}_{t+1}$ involving non-differentiable

sampling. On the other hand, projective dynamics does not require an analytical or differentiable expression of Eq. (10).

It's technically also feasible to train an Energy-Based Model [LeCun et al. 2006], explicitly predicting the scalar distance (score) to training data distribution, given any motion transition as input. An EBM, being a closed-form differentiable neural network, can theoretically work with any energy-based simulation framework. In fact, that is precisely what we experimented with early on in this work before deciding to switch to projective dynamics - despite recent advances in training EBMs, we found they can still be unstable to train and hard to scale to a large dataset. In contrast, projective dynamics frees us to use any motion prior that scales well, such as VAEs [Rempe et al. 2021], discretized token-based Transformers [Li et al. 2020b], or more recent Diffusion Models [Tevet et al. 2023].

9.2 Limitations

Due to the sensitivity of the root wrench metric (Sec. 8.4), we choose to allow a small amount of root wrench, which may occasionally lead to puppet-like motions. Further research is required to base this metric on a time window rather than a single step, which may potentially mitigate its sensitivity resulted from finite differencing. The root wrench correction should also ideally be incorporated into the projective dynamics solver rather than as post-processing. Particularly at states where data coverage is poor, the motion prior could mandate a very unrealistic transition causing the root correction to solve for a very large correction step in response, which might violate previously solved energy terms causing artifacts such as foot sliding.

We note that our foot sliding or penetration metric (equivalently to the level of contact energy, Fig. 4 and 6) and physicality metric (equivalent to root wrench, Fig. 7) should be at least as good as HuMoR on unperturbed AMASS training motions, in which case our method degenerates to physics-based post-processing on top of kinematic models. During perturbation recoveries, our physics-based energy values become higher, which equivalently indicates quantitative physics-based metrics becoming worse.

As our method involves no dependency on existing physics engines or libraries, our current implementation from scratch involves several simplifications. Our method does not simulate toes, hands, or head, and only uses spheres at each joint as collision geometries. As result, our visualization (using fully skinned SMPL model) inevitably sees occasional foot and hand penetration, as well as self-penetration between body parts, even though our RoM energy provides some implicit mitigation to self-penetration. We also use a maximal-coordinate stick figure to ease the incorporation of projective dynamics, with the twist motion excluded in the projective dynamics solver and fully predicted kinematically from HuMoR. Extending the framework to articulated rigid bodies would provide a better representation.

ACKNOWLEDGMENTS

This work was supported by the Wu Tsai Human Performance Alliance at Stanford University, the Stanford Institute for Human-Centered AI (HAI), Meta Reality Labs, and NSF:CCRI 2120095. Jungdam Won was partially supported by the New Faculty Startup

Fund from Seoul National University and the ICT (Institute of Computer Technology) at Seoul National University. We would like to thank Seunghwan Lee, Andrew Spielberg, Tom Van Wouwe, Jackson Kuan-Chieh Wang, Albert Wu, Eris Jiayi Zhang, Alexander Winkler, Deepak Gopinath, and anonymous reviewers for helpful discussions.

REFERENCES

2003. Carnegie-Mellon Mocap Database. <http://mocap.cs.cmu.edu/>
- Okan Arıkan and David A. Forsyth. 2002. Interactive Motion Generation from Examples. *ACM Trans. Graph.* 21, 3 (jul 2002), 483–490. <https://doi.org/10.1145/566654.566606>
- Okan Arıkan, David A. Forsyth, and James F. O'Brien. 2005. Pushing People Around. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '05)*. 59–66.
- Emad Barsoum, John Kender, and Zicheng Liu. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1418–1427.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems* 28 (2015).
- Kevin Bergamin, Simon Clavet, Daniel Holden, and James Richard Forbes. 2019. DReCon: Data-Driven Responsive Control of Physics-Based Characters. *ACM Trans. Graph.* 38, 6, Article 206 (nov 2019), 11 pages. <https://doi.org/10.1145/3355089.3356536>
- Sofien Bouaziz, Sebastian Martin, Tiantian Liu, Ladislav Kavan, and Mark Pauly. 2014. Projective Dynamics: Fusing Constraint Projections for Fast Simulation. *ACM Trans. Graph.* 33, 4, Article 154 (jul 2014), 11 pages. <https://doi.org/10.1145/2601097.2601116>
- Yunuo Chen, Minchen Li, Lei Lan, Hao Su, Yin Yang, and Chenfanfu Jiang. 2022. A Unified Newton Barrier Method for Multibody Dynamics. *ACM Trans. Graph.* 41, 4, Article 66 (jul 2022), 14 pages. <https://doi.org/10.1145/3528223.3530076>
- Simon Clavet. 2016. Motion Matching and The Road to Next-Gen Animation. In *Game Developers Conference*.
- Marco da Silva, Yeui Abe, and Jovan Popović. 2008. Interactive Simulation of Stylized Human Locomotion. *ACM Trans. Graph.* 27, 3 (aug 2008), 1–10.
- Levi Fussell, Kevin Bergamin, and Daniel Holden. 2021. SuperTrack: Motion Tracking for Physically Simulated Characters Using Supervised Learning. *ACM Trans. Graph.* 40, 6, Article 197 (dec 2021), 13 pages. <https://doi.org/10.1145/3478513.3480527>
- Theodore F Gast, Craig Schroeder, Alexey Stomakhin, Chenfanfu Jiang, and Joseph M Teran. 2015. Optimization integrator for large time steps. *IEEE transactions on visualization and computer graphics* 21, 10 (2015), 1103–1115.
- Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. 2020. Learned Motion Matching. *ACM Trans. Graph.* 39, 4, Article 53 (aug 2020), 13 pages.
- Daniel Holden, Taku Komura, and Jun Saito. 2017. Phase-Functioned Neural Networks for Character Control. *ACM Trans. Graph.* 36, 4, Article 42 (jul 2017), 13 pages.
- Ben Jones, Nils Thuerey, Tamar Shinar, and Adam W. Bargteil. 2016. Example-Based Plastic Deformation of Rigid Bodies. *ACM Trans. Graph.* 35, 4, Article 34 (jul 2016), 11 pages. <https://doi.org/10.1145/2897824.2925979>
- Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion Graphs. *ACM Trans. Graph.* 21, 3 (jul 2002), 473–482.
- Lei Lan, Guanqun Ma, Yin Yang, Changxi Zheng, Minchen Li, and Chenfanfu Jiang. 2022. Penetration-Free Projective Dynamics on the GPU. *ACM Trans. Graph.* 41, 4, Article 69 (jul 2022), 16 pages. <https://doi.org/10.1145/3528223.3530069>
- Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fugie Huang. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).
- Jehee Lee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard. 2002. Interactive Control of Avatars Animated with Human Motion Data. *ACM Trans. Graph.* 21, 3 (jul 2002), 491–500.
- Seunghwan Lee, Phil Sik Chang, and Jehee Lee. 2022. Deep Compliant Control. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) (SIGGRAPH '22). Association for Computing Machinery, New York, NY, USA, Article 23, 9 pages. <https://doi.org/10.1145/3528233.3530719>
- Jongjoon Lee, Kevin Wampler, Gilbert Bernstein, Jovan Popović, and Zoran Popović. 2010. Motion Fields for Interactive Character Locomotion. *ACM Trans. Graph.* 29, 6, Article 138 (dec 2010), 8 pages.
- Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3383–3393.
- Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020b. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171* (2020).
- Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M. Kaufman. 2020a. Incremental Potential Contact: Intersection-and Inversion-Free, Large-Deformation Dynamics. *ACM Trans. Graph.* 39, 4, Article 49 (aug 2020), 20 pages. <https://doi.org/10.1145/3386569.3392425>
- Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. 2020. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 40–1.
- Libin Liu and Jessica Hodgins. 2017. Learning to Schedule Control Fragments for Physics-Based Characters Using Deep Q-Learning. *ACM Trans. Graph.* 36, 4, Article 42a (jul 2017), 14 pages. <https://doi.org/10.1145/3072959.3083723>
- Tiantian Liu, Adam W Bargteil, James F O'Brien, and Ladislav Kavan. 2013. Fast simulation of mass-spring systems. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 1–7.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.
- Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.
- Sebastian Martin, Bernhard Thomaszewski, Eitan Grinspun, and Markus Gross. 2011. Example-Based Elastic Materials. In *ACM SIGGRAPH 2011 Papers* (Vancouver, British Columbia, Canada) (SIGGRAPH '11). Association for Computing Machinery, New York, NY, USA, Article 72, 8 pages. <https://doi.org/10.1145/1964921.1964967>
- Qianhui Men, Hubert P.H. Shum, Edmond S.L. Ho, and Howard Leung. 2022. GAN-based reactive motion synthesis with class-aware discriminators for human-human interaction. *Computers & Graphics* 102 (2022), 634–645.
- Uldarico Muico, Yongjoon Lee, Jovan Popović, and Zoran Popović. 2009. Contact-Aware Nonlinear Control of Dynamic Characters. *ACM Trans. Graph.* 28, 3, Article 81 (jul 2009), 9 pages.
- Soohwan Park, Hoseok Ryu, Seyoung Lee, Sunmin Lee, and Jehee Lee. 2019. Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–11.
- Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.
- Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills. *ACM Trans. Graph.* 37, 4, Article 143 (jul 2018), 14 pages. <https://doi.org/10.1145/3197517.3201311>
- Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. 2022. ASE: Large-Scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters. *ACM Trans. Graph.* 41, 4, Article 94 (jul 2022), 17 pages. <https://doi.org/10.1145/3528223.3530110>
- Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. 2021. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Trans. Graph.* 40, 4, Article 144 (jul 2021), 20 pages. <https://doi.org/10.1145/3450626.3459670>
- Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. In *International Conference on Computer Vision (ICCV)*.
- Takaaki Shiratori, Brooke Coley, Rakié Cham, and Jessica K. Hodgins. 2009. Simulating Balance Recovery Responses to Trips Based on Biomechanical Principles. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '09)*. 37–46.
- Sebastian Starke, Ian Mason, and Taku Komura. 2022. DeepPhase: Periodic Autoencoders for Learning Motion Phase Manifolds. *ACM Trans. Graph.* 41, 4, Article 136 (jul 2022), 13 pages.
- Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. 2019. Neural State Machine for Character-Scene Interactions. *ACM Trans. Graph.* 38, 6, Article 209 (nov 2019), 14 pages.
- Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=SJ1kSyO2jwu>
- Zhenzhen Weng, Laura Bravo-Sánchez, and Serena Yeung. 2023. Diffusion-HPC: Generating Synthetic Images with Realistic Humans. *arXiv preprint arXiv:2303.09541* (2023).
- Jungdam Won, Deepak Gopinath, and Jessica Hodgins. 2020. A Scalable Approach to Control Diverse Behaviors for Physically Simulated Characters. *ACM Trans. Graph.* 39, 4, Article 33 (aug 2020), 12 pages. <https://doi.org/10.1145/3386569.3392381>
- Jungdam Won, Deepak Gopinath, and Jessica Hodgins. 2022. Physics-Based Character Controllers Using Conditional VAEs. *ACM Trans. Graph.* 41, 4, Article 96 (jul 2022), 12 pages. <https://doi.org/10.1145/3528223.3530067>
- Kaixiang Xie, Pei Xu, Sheldon Andrews, Victor Zordan, and Paul Kry. 2023. Too Stiff, Too Strong, Too Smart: Evaluating Fundamental Problems with Motion Control Policies. In *The Proceedings of the ACM in Computer Graphics and Interactive Techniques*.
- Pei Xu and Ioannis Karamouzas. 2021. A GAN-Like Approach for Physics-Based Imitation Learning and Interactive Character Control. *Proc. ACM Comput. Graph. Interact. Tech.* 4, 3 (2021). <https://doi.org/10.1145/3480148>

- Heyuan Yao, Zhenhua Song, Baoquan Chen, and Libin Liu. 2022. ControlVAE: Model-Based Learning of Generative Controllers for Physics-Based Characters. *ACM Trans. Graph.* 41, 6, Article 183 (nov 2022), 16 pages.
- Yuting Ye and C. Karen Liu. 2008. Animating responsive characters with dynamic constraints in near-unactuated coordinates. In *SIGGRAPH Asia '08: ACM SIGGRAPH Asia 2008 papers*. ACM, New York, NY, USA, 1–5.
- Yuting Ye and C. Karen Liu. 2010a. Optimal feedback control for character animation using an abstract model. In *SIGGRAPH '10: ACM SIGGRAPH 2010 papers*. ACM, 1–9.
- Yuting Ye and C. Karen Liu. 2010b. Synthesis of Responsive Motion Using a Dynamic Model. *Computer Graphics Forum* 29, 2 (2010), 555–562.
- He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. 2021. ManipNet: Neural Manipulation Synthesis with a Hand-Object Spatial Representation. *ACM Trans. Graph.* 40, 4, Article 121 (jul 2021), 14 pages. <https://doi.org/10.1145/3450626.3459830>
- Yu Zheng and Katsu Yamane. 2013. Human motion tracking control with strict contact force constraints for floating-base humanoid robots. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. 34–41. <https://doi.org/10.1109/HUMANOIDS.2013.7029952>
- Victor Brian Zordan and Jessica K. Hodgins. 2002. Motion Capture-Driven Simulations That Hit and React. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '02)*. 89–96.
- Victor Brian Zordan, Anna Majkowska, Bill Chiu, and Matthew Fast. 2005. Dynamic response for motion capture animation. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 697–701.