Skeleton2Humanoid: Animating Simulated Characters for Physically-plausible Motion In-betweening

Yunhao Li
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
lyhsjtu@sjtu.edu.cn

Zhenbo Yu
Shanghai Jiao Tong University
Shanghai, China
yuzhenbo@sjtu.edu.cn

Yucheng Zhu
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
zyc420@sjtu.edu.cn

Bingbing Ni
Shanghai Jiao Tong University
Shanghai, China
nibingbing@sjtu.edu.cn

Guangtao Zhai
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
zhaiguangtao@sjtu.edu.cn

Wei Shen
MoE Key Lab of Artificial Intelligence, AI Institute
Shanghai Jiao Tong University
Shanghai, China
wei.shen@sjtu.edu.cn

Figure 1: Our Skeleton2Humanoid system can directly synthesize a complete humanoid character transition motion in a physics simulator (Bottom) given past keyframes and a future keyframe (Top). Our system can produce both accurate and physically-plausible character motions.

ABSTRACT

Human motion synthesis is a long-standing problem with various applications in digital twins and the Metaverse. However, modern deep learning based motion synthesis approaches barely consider the physical plausibility of synthesized motions and consequently they usually produce unrealistic human motions. In order to solve this problem, we propose a system “Skeleton2Humanoid” which performs physics-oriented motion correction at test time by regularizing synthesized skeleton motions in a physics simulator. Concretely, our system consists of three sequential stages: (I) test time motion synthesis network adaptation, (II) skeleton to humanoid matching and (III) motion imitation based on reinforcement learning (RL). Stage I introduces a test time adaptation strategy, which improves the physical plausibility of synthesized human skeleton motions by optimizing skeleton joint locations. Stage II performs an analytical inverse kinematics strategy, which converts the optimized human skeleton motions to humanoid robot motions in a physics simulator, then the converted humanoid robot motions can be served as reference motions for the RL policy to imitate. Stage III introduces a curriculum residual force control policy, which drives the humanoid robot to mimic complex converted reference motions in accordance with the physical law. We verify our system on a typical human motion synthesis task, motion-in-betweening. Experiments on the challenging LaFAN1 dataset show our system can outperform prior methods significantly in terms of both physical plausibility and accuracy. Code will be released for research purposes at: https://github.com/michaelliyunhao/Skeleton2Humanoid.
CCS CONCEPTS
• Computing methodologies → Motion capture.

KEYWORDS
3D motion in-betweening; inverse kinematics; reinforcement learning; 3D animation

ACM Reference Format:
Yunhao Li, Zhenbo Yu, Yucheng Zhu, Bingbing Ni, Guangtao Zhai, and Wei Shen. 2022. Skeleton2Humanoid: Animating Simulated Characters for Physically-plausible Motion In-betweening. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, Lisbon, Portugal, 10 pages. https://doi.org/10.1145/3503161.3548093

1 INTRODUCTION

Synthesizing both accurate and realistic virtual human motions has been a widely explored but challenging task in computer vision and graphics [18, 30] with various applications in digital twins and the Metaverse. Recently, deep learning sheds light onto a way to generate accurate human motions and has been applied to various motion synthesis tasks, such as human motion prediction [1, 4, 8, 16, 22, 27–29, 48, 51], human motion completion [14, 17, 41] and human motion in-betweening [6, 7, 13, 52]. Although they have shown great performance on synthesizing accurate human body motions with small skeleton joint errors comparing with ground truth motions, they fail to model the motions under the physics laws. Consequently, the synthesized motions are usually physically implausible. For example, the synthesized feet often penetrate the ground, the body joints are rotated with impossible angles, the whole body motions are unsmooth, the synthesized feet slide back and forth while they should be in static and touch the ground. These synthesized artifacts significantly limit the application of motion synthesis on the virtual human animation and the incoming Metaverse because they easily make humans feel unrealistic.

Utilizing humanoid characters in a physics simulator to optimize motions is a promising solution because the physics simulator can guarantee the physical plausibility of the generated motions. Prior works [33, 34, 42] utilized reinforcement learning (RL) to actuate the humanoid character to imitate various reference mocap data for creating physical character animation. Inspired by them, Recent works [26, 50] also attempted to utilize RL to imitate motions synthesized by deep neural networks, in the format of skeletons or SMPL [25] models, aiming at producing physically-plausible motions for 3D pose estimation. However, these methods are only validated on simple motions such as walking and talking in the Human3.6m dataset and cannot generalize well to complex motions or irregular motions. In addition, RL based imitation requires transferring synthesized human skeleton motions to humanoid motions, where a humanoid character should be carefully designed to exactly match the human skeletons in terms of both shapes and the kinematics tree. This limits RL based imitation to transfer motions between skeleton and humanoid with different shapes and kinematics trees.

To address these issues, we propose Skeleton2Humanoid, a novel system which is able to improve the physical plausibility of the motions synthesized from motion synthesis networks, though the transfer from human skeleton motions to humanoid character motions. Our Skeleton2Humanoid system consists of three sequential stages: (I) Test Time Motion Synthesis Network Adaptation: We adapt the motion synthesis network with a few gradients on the test data using two new self-supervised losses, a foot contact consistency loss and a motion smoothness loss, which can improve the physical plausibility of the predicted motions. (II) Skeleton to Humanoid Matching: We match the synthesized human skeleton motions to humanoid character motions by a novel general analytical inverse kinematic method. Inverse kinematics is able to convert human skeleton motions to humanoid motions even when the body structure is different from the human skeleton. (III) Motion Imitation base on RL: Finally, we animate the humanoid character to imitate various synthesized motions. Specifically, based on recent work [26, 49], we propose a curriculum residual force control humanoid control policy (CRP) by introducing a curriculum learning paradigm that dynamically adjusts a residual force scale during RL training, which can improves asymptotic RL performance on imitating various synthesized motions. To verify the effectiveness of our Skeleton2Humanoid system, we select “motion in-betweening” task, as it is a recent proposed challenging motion prediction task [7, 13] for evaluation. Motion in-betweening aims at predicting the transition motions between the past given keyframes and a provided future keyframe. Experiments on challenging LaFAN1 dataset show the superiority of our Skeleton2Humanoid system.

The main contributions of this paper are as follows: (1) We present Skeleton2Humanoid, a new system that converts human skeleton motions to humanoid character motions to produce physically plausible motions. (2) Our proposed test time adaptation stage can further improve the prediction accuracy and physical plausibility on large mocap dataset LaFAN1 for the motion in-betweening task. With test time adaptation, we achieve a new benchmark accuracy on the motion in-betweening task. (3) Our proposed curriculum residual force control policy enables finer character control and outperforms prior arts on motion imitation. (4) Our whole Skeleton2Humanoid system significantly improves the performance of human in-betweening motions on physical plausibility and achieves comparable motion prediction accuracy.

2 RELATED WORK

Human/character motion synthesis: Motion synthesis is a general term which contains several tasks including motion prediction, in-betweening and completion. Motion prediction aims at predicting future human motions given past motions. Deterministic motion prediction estimates a single accurate motion and prior works used various network architectures including recurrent neural network [8, 16, 29], graph convolution network [20] or transformer [27] to model human motions. Stochastic motion prediction produces diverse future human motions by utilizing generative model such as VAE [12, 35, 48, 51], GAN [2, 11, 37]. Motion completion and in-betweening aim at filling gaps of motion with predefined keyframe constraints. Current works utilized convolution networks [14, 17, 44, 52], recurrent networks [5, 13] or transformers [7] to synthesize accurate and consistent results. For instance, Harvey et al. [13] proposed a transition generation technique based on recurrent neural networks for motion in-betweening task. Duan et al. [7] utilized transformer architecture to model human motions in a sequence-to-sequence manner for the motion in-betweening task.
3 APPROACH

3.1 System Overview

The human motion in-betweening task can be formulated as: given the past 10 human skeleton poses \(\overline{m}_{-9:0}\) and a future skeleton keyframe \(\overline{m}_T\) at time \(T\), we want to recover the ground truth transition motions \(\overline{m}_{1:T-1}\). Given a pretrained typical motion in-betweening network \([13]\), our Skeleton2Humanoid performs a physics-oriented motion correction consists of 3 stages as presented in Fig. 2 to optimize synthesized in-betweening motions. Stage I optimizes the pretrained motion in-betweening network at test time to predict more physically-plausible skeleton transition motions \(\overline{m}_{1:T-1}\). Then Stage II transfers the optimized skeleton motions \(\overline{m}_{1:T-1}\) to humanoid motions \(\overline{m}_{1:T-1}\) through analytical inverse kinematics. Finally, Stage III learns a curriculum residual force control policy to imitate the transferred humanoid motions \(\overline{m}_{1:T-1}\) to produce physically-plausible humanoid motions \(\overline{m}_{1:T-1}\).

In our Skeleton2Humanoid framework, \(\overline{m}_T\) and \(\hat{m}_T\) are skeleton motions, and \(\overline{m}_T\) is represented by \(\overline{m}_T \triangleq (\overline{q}_T, \overline{r}_T, \overline{p}_T)\), where \(\overline{q}_T\) and \(\overline{r}_T\) denote body joint angles in quaternions and root translation, \(\overline{p}_T\) denotes 3d joint positions calculated by forward kinematics. Similarly, \(\hat{m}_T \triangleq (\hat{q}_T, \hat{r}_T, \hat{p}_T)\), \(\hat{m}_T\) represents 3d skeleton positions calculated by inverse kinematics. In addition, \(m_{1:T}\) and \(\hat{m}_{1:T}\) are humanoid motions, \(\hat{m}_{1:T}\) is represented by \(\hat{m}_{1:T} \triangleq (\hat{q}_{1:T}, \hat{r}_{1:T}, \hat{p}_{1:T})\), where \(\hat{q}_{1:T}, \hat{r}_{1:T}\) and \(\hat{p}_{1:T}\) denote joint angles in euler angles, root translation and 3d joint positions of the reference humanoid motions. Similarly, \(m_{1:T} \triangleq (q_{1:T}, r_{1:T}, p_{1:T})\).

3.2 Test Time Motion In-betweening Network Adaptation

3.2.1 Adaptation for Physically-plausible Skeleton Motion. Previous human motion in-betweening model \([13]\) has shown great performance on synthesizing accurate human motions. However, it still suffers from implausibility on testing data because of the
The training loss is formulated as:

\[ L_{\text{train}} = L_{\text{inbetween}} + \alpha L_{\text{contact}}, \]

where \( L_{\text{inbetween}} \) is the sum of all the training losses used in [13], \( L_{\text{contact}} \) is the foot contact consistency loss, and \( \alpha \) is a constant coefficients.

During testing, we only use the self-supervised losses to update the motion in-betweening network which is trained on the training set given the test data. The testing loss is formulated as:

\[ L_{\text{test}} = L_{\text{contact}} + \beta L_{\text{smooth}}; \]

where \( \beta \) is a constant coefficients. After test time adaptation, the motion in-betweening network is able to produce optimized plausible skeleton motions \( \hat{m}_{1:T-1} \).

### 3.2.2 Implementation Detail

Now, we describe the training phase and the testing phase of our test time adaptation method, respectively.

During training, as shown in Fig. 3, we select the typical motion in-betweening network [13] as our baseline to generate transition human skeleton motions. The training process is the same as [13].

The training loss is formulated as:

\[ L_{\text{inbetween}} = \sum_{t=0}^{T-2} \sum_{f \in F} \| \dot{p}_{\text{foot}}^f - \dot{p}_f^t \|_2 \cdot \text{contact}_t^f, \]

![Figure 4: Illustration of skeleton to humanoid matching process in left elbow.](image)

where \( \dot{p}_{\text{foot}}^f \) is the predicted 3D position of the foot joint at time \( t \), \( F \) is a subset of joints which contains all foot-related joints, i.e. "left ankle", "right ankle", "left front feet" and "right front feet". \( \text{contact}_t^f \) represents the predicted contact probability for the foot joint at time \( t \).

The motion smoothness loss: We encourage the root joint positions along the time axis to be smooth. The motion smoothness loss is:

\[ L_{\text{smooth}} = \sum_{t=0}^{T-2} \| \dot{r}_t \|_2, \]

where \( r_t \) is the 3D position of root joint at time \( t \).

By utilizing these two losses, we can optimize the motion in-betweening network which is trained on the training set given the test set. In the following section, we will introduce our method in detail.

#### 3.3 Skeleton to Humanoid Matching

After acquiring optimized skeleton motions \( \hat{m}_{1:T-1} \) from stage I, we then transfer the skeleton motions \( \hat{m}_{1:T-1} \) to corresponding humanoid motions \( \hat{m}_{1:T-1} \) which can be imitated by RL. To achieve this, recent works [26, 50] constructed a humanoid model which is exactly the same as the skeleton or the SMPL model in shapes and kinematics tree, hence it is easy to transfer the motions to humanoid motions with no error because they can directly transfer the axis-angles of the SMPL model to the corresponding euler angles of humanoid. However, we emphasize that it is not suitable for all cases in practice.

First, this method is not general and cumbersome because it needs to carefully construct a humanoid model and it is not possible to flexibly transfer motions between human skeleton and humanoid with different shapes and kinematics trees. Second, constructing an exactly same humanoid model is an ideal solution because the pose of humanoid \( m_f \) in the physics simulator is represented by euler angles \( \tilde{q}_t \) with specific ranges which corresponds body joints constraints. For example, the euler angle of the shoulder joint in z axis is usually between -90 and 90 degrees. However, transferring the quaternions of the skeleton to the corresponding euler angles of humanoid in a specific range is actually an ill-posed problem. The range of the euler angle in each axis is between -180 and 180 degrees, hence one rotation can be represented by many different euler angles. For human skeletons with strange poses or strange kinematics trees, it is difficult to transfer the quaternions to the corresponding euler angles in a specific range.

To solve this problem, inverse kinematics is a solution that can convert the 3d joint positions of skeleton into the euler angles of humanoid. Analytical Inverse kinematics calculates the rotation vectors at the end of a kinematic chain in a given position and can match motions between skeletons and humanoids with different shapes and kinematics trees. Previous works use analytical inverse kinematics [19, 45] to convert the 3d joint positions of skeleton to the axis-angles of the SMPL model. We extend their works to convert 3d joint positions of skeleton to the euler angles of humanoid via specific matching equations. Concretely, we select a suitable axis definition for each joint of the skeleton that can compute the corresponding euler angle of the humanoid model. We take the left elbow for instance to describe the detailed matching process in Fig. 4. Following [45], we consider the skeleton and the humanoid as articulated bodies with pairs of parent-children joints. Suppose the parent and the child coordinate system of joints in 3d skeleton is denoted as \([x_p, y_p, z_p]\) and \([x_c, y_c, z_c]\), then the parent and child coordinate system of the left elbow is calculated by Eqn. 5,
where each item (e.g., $x_c$) is a 3x1 vector in world coordinate system. $\mathbf{l}_{a-b}$ indicates the vector pointing from joint $a$ to joint $b$ and subscripts $l_s, l_e, l_w$ denote left shoulder, left elbow and left wrist, respectively. Specifically, $\mathbf{l}_{l_e-l_w}$ means the vector pointing from the left elbow joint to the left wrist joint. Then we can calculate the euler angle $E$ of the left elbow of the humanoid via Eqn. 6.

\[
\begin{align*}
\vec{N} &= [n_0, n_1, n_2] = [x_c \cdot x_p^T, y_c \cdot y_p^T, z_c \cdot z_p^T] \\
\vec{O} &= [o_0, o_1, o_2] = [y_c \cdot x_p^T, x_c \cdot y_p^T, z_c \cdot z_p^T] \\
\vec{A} &= [a_0, a_1, a_2] = [z_c \cdot x_p^T, x_c \cdot y_p^T, y_c \cdot z_p^T] \\
E_z &= \text{atan2}(a_1, a_0) \\
E_y &= \text{atan2}(n_0 \cdot \cos(E_z) + n_1 \cdot \sin(E_z)) \\
E_x &= \text{atan2}(a_1 \cdot \cos(E_z) + a_0 \cdot \sin(E_z)) \\
E &= [E_x, E_y, E_z]
\end{align*}
\]

where $E$ is the euler angle of the humanoid model in z-y-x rotation order. After applying the similar matching operation on the other joints of the humanoid, the whole skeleton motions $\tilde{\mathbf{m}}_{10:T-1}$ can be converted to humanoid motions $\mathbf{m}_{10:T-1}$. In addition, because the matching process is applied frame by frame, the values of the converted euler angles usually flip and become discontinued due to the $\text{atan2}$ process in equation 6. We correct the values of euler angles by adding a temporally checking process, suppose $E_z^t$ is a euler angle of a joint in z axis at time $t$,

\[
\begin{align*}
E_z^{t+1} &= E_z^{t+1} + 2\pi, & \text{if } E_z^t > 0 \text{ and } |E_z^{t+1} - E_z^t| > \lambda \\
E_z^{t+1} &= E_z^{t+1} - 2\pi, & \text{if } E_z^t < 0 \text{ and } |E_z^{t+1} - E_z^t| > \lambda,
\end{align*}
\]

where $\lambda$ is a threshold value for checking if the euler angles are discontinued or not and we set $\lambda = 5$. In the end, we can convert the synthesized skeleton motions $\tilde{\mathbf{m}}_{1:T-1}$ to humanoid motions $\mathbf{m}_{1:T-1}$.

### 3.4 Motion Imitation based on RL

After getting the converted humanoid motions $\mathbf{m}_{1:T-1}$, they are still physically implausible due to lack of physical constraint. Motivated by [26, 49, 50], we utilize reinforcement learning to learn a curriculum residual force control policy $\pi_{\text{CRP}}$ which can imitate $\mathbf{m}_{1:T-1}$ to generate physically-plausible humanoid motions $\mathbf{m}_{1:T-1}$. The motion imitation problem is usually formulated as a Markov decision process. Given a reference motion and current state $s_t$, the humanoid agent interacts with the physics simulator environment by action $a_t$ and receive reward $r_t$. The action is produced by a policy $\pi_{\text{CRP}}(a_t|s_t)$ conditioned on state $s_t$. The reward encourages the agent to act like the reference motion. When an action is taken, the current agent state $s_t$ changes to next state $s_{t+1}$. The goal of the motion imitation is to learn a policy that can maximizes the average cumulative rewards.

3.4.1 Motion Imitation Paradigm. Here we detail the states, actions, rewards and the training process.

**State.** Previous work [33, 49, 50] construct the humanoid state $s_t$ which contains the information of the character’s current pose and the information of the reference pose in the next frame. We utilize the information of the reference pose in the next two frames that can learn a finer control policy. Specifically, the state $s_t \triangleq (q_t, \nu_t, \dot{q}_{t+1}, \dot{\nu}_{t+1}, \ddot{q}_{t+2})$ and we apply a hand-craft feature extraction to obtain informative features $f_t$ as the input of $\pi_{\text{CRP}}$ from $s_t$. We formulate $f_t \triangleq (q_t, \nu_t, \dot{q}_{t+1}, \ddot{q}_{t+2}, \dot{\nu}_{t+1} - \dot{q}_t, \ddot{\nu}_{t+1} - \ddot{q}_t)$, where $q_t$ is the joint angle of the simulated humanoid, $\nu_t$ is the joint velocity computed by finite difference, $\dot{q}_t$ and $\ddot{q}_t$ are the 3d joint position in local head coordinates and global coordinates, $\dot{\nu}_{t+1}$ is the euler angle of reference pose in time $t+1$, $\ddot{\nu}_{t+1}$ the joint velocity of the reference pose in time $t+1$, $\dot{\nu}_{t+2}$ and $\ddot{\nu}_{t+2}$ are the 3d joint position of the reference pose in local head coordinates and global coordinates.

**Action.** Following prior works [26, 49, 50], we utilize proportional derivative (PD) controller [39] at each non-root joint to generate torques. Under this setting, the action $a_t$ is the target joint angles $u_t$ of the PD controllers. The joint torques $\tau_t$ are specifically calculated by:

\[
\tau_t = k_p \circ (u_t - q_t) - k_d \circ \dot{q}_t,
\]

where $k_p$ and $k_d$ are manually defined parameters and $\circ$ is element-wise multiplication. The PD controllers act like damped springs that drive joints to target angles $u_t$, while $k_p$ and $k_d$ are the stiffness and damping of the springs. Prior works [26, 49] also allowed the policy to apply external residual forces $\pi_{\text{CRP}}(a_t|s_t)$ to joints to improve the motion imitation performance. $S_t$ is a predefined residual force scale and $k_e \in [0, 1]$ is a weighting parameter learned by policy.

In section 3.4.2, we will introduce our curriculum residual force learning that can dynamically adjust the residual force scale $S_t$ in RL training. Overall, our action is $a_t \triangleq (u_t, \pi_{\text{CRP}}(a_t|s_t)).$
Reward. Following [26, 49], we implement the same reward function that encourages the policy to mimic the reference pose.

3.4.2 Curriculum Learning for Residual Forces. The residual forces $r_t$ proposed by [49] aims at utilizing additional forces to enable finer control of the humanoid character. Recent works [49, 50] used it to perform humanoid motions with extra interaction such as sitting on the chair even there is no chair in the physics simulator. However, large scale residual forces will make humanoids perform some unnatural motions or overly energetic movements. Meanwhile, removing the residual forces will cause the low converge speed of RL and a bad imitation performance. To balance the trade-off between the motion naturalness and the imitation performance, we propose a curriculum learning strategy that gradually adjusts the residual force scale in the training of reinforcement learning.

Starting from a large residual forces, we gradually decrease the residual force scale in RL training. Suppose $i$ is the training iteration of PPO RL algorithm, $S_r$ is residual force scale which is adjusted by:

$$S_r = \begin{cases} 
S_r^{\text{init}}, & \text{if } 0 \leq i \leq i_{\text{start}} \\
S_r^{\text{init}} - \alpha \cdot (i - i_{\text{start}}), & \text{if } i_{\text{start}} \leq i \leq i_{\text{end}} \\
S_r^{\text{init}}, & \text{if } i \geq i_{\text{end}} 
\end{cases}$$

(9)

where $i_{\text{start}}$ and $i_{\text{end}}$ are the start and end iterations of the curriculum, and $\alpha$ is a hyper-parameter indicating the changing rate of $S_r$. We empirically set $S_r^{\text{init}} = 220$, $i_{\text{end}} = 1300$, $i_{\text{start}} = 100$, $\alpha = 0.1$ and train RL for 1500 iterations. By introducing curriculum learning for residual forces, the policy can quickly learn to imitate the reference motion precisely with the help of large residual forces and learn to maintain the imitation performance while getting rid of large residual forces.

4 EXPERIMENTS

In this section, we perform experimental evaluations of our Skeleton2Humanoid system on the motion realism and accuracy, then we analyze the properties of the test time adaptation stage, the skeleton to humanoid matching stage and the motion imitation stage based on RL individually.

Dataset. To demonstrate the effectiveness of our Skeleton2Humanoid system on the motion in-betweening task, we employ the following datasets in our experiments: (1) LaFAN1 [13], which is a widely used large scale dataset for evaluating motion in-betweening performance. LaFAN1 dataset contains various complex motions such as fighting, crawling and climbing obstacles. To be noticed, we only evaluate the test time adaptation stage of our Skeleton2Humanoid system on LaFAN1 dataset and we will explain it later. (2) Noobjects LaFAN1 is a subset of LaFAN1 dataset for evaluating the physical plausibility of our synthesized in-betweening motions. LaFAN1 dataset contains many human object interaction motions such as climbing obstacles and we can not construct objects in the physics simulator due to the absence of the ground truth of 3d objects, hence we remove these motions and only consider human motions performing on the ground. We evaluate our whole Skeleton2Humanoid system on this dataset.

Evaluation Metrics. We use the following metrics including physics-based and accuracy-based metrics for all following evaluations:

1. $\text{FP}$, which is the foot penetration that computes the mean penetration distance of feet joints and ground following [36].
2. $\text{FQ}$, which computes the frequency of the penetrations between the feet and ground penetrations.
3. $\text{JQ}$, which similarly computes the frequency of penetrations between all the joints and the ground.
4. $\text{SM}$, which represents motion smoothness that computes the average L2 distance between body joints within two adjacent frames.
5. $\text{MPJPE}$, which is the mean per joint position error for evaluating the mismatch between poses. (6) $\text{L2Q}$ and $\text{L2P}$, where $\text{L2Q}$ denotes the the average L2 distances of the global quaternions between the predicted motions and their ground truth. Similarly, $\text{L2P}$ denotes the average L2 distance of the global positions. We follow standard evaluations from [13] to compute them.

4.1 Evaluation of Skeleton2Humanoid System

Experiment Setting. We evaluate our whole Skeleton2Humanoid system on the Noobjects LaFAN1 dataset. We select the typical state-of-the-art motion in-betweening approach Harvey et al. [13] as a baseline and validate our approach based on it. The baseline method is trained with various transition lengths (5 frames to 30 frames) between past keyframes and a future keyframe and is evaluated when transition lengths are 5 frames, 15 frames and 30 frames.

Results and Discussion. The results on the Noobjects LaFAN1 dataset are summarized in table 1. First, it is evident that our Skeleton2Humanoid outperforms the baseline in all physics-based metrics and achieves comparable performance in accuracy-based metrics. In particular, Skeleton2Humanoid achieve significantly lower FP, FQ, JQ and SM, which demonstrates that utilizing the simulated humanoid character can effectively reduce the physical artifacts such as the penetration problem between human bodies and the ground and can produce physically-plausible motions. Second, we can observe that our test time adaptation stage (TTA) outperforms the baseline in both physics-based metrics and accuracy-based metrics, demonstrating that our TTA can optimize the motion in-betweening model to produce not only plausible but also accurate motions. Third, we find that applying the analytical inverse kinematics and the curriculum residual force control policy (CRP) cause worse L2P. The reason is that our efficient analytical inverse kinematics strategy can not perfectly matching the skeleton and humanoid even its matching error is already really low (MPJPE: 11.29mm) as discussed in section 4.2.2, and our state-of-the-art motion imitation policy CRP as shown in section 4.2.3 still struggle to imitate various complex in-betweening motions perfectly. Such a phenomenon is also observed in [26] when they utilize RL to imitate human motions for ego pose estimation. We emphasize that current motion in-betweening method should not only consider the accuracy of synthesized motions but also consider the physical plausibility.

We also provide qualitative results in Fig. 6, we can see that the imitated humanoid motions can accurately synthesize joint contact with the ground and without penetration. Our state-of-the-art RL policy is also able to control the humanoid to perform complex motions such as crawling. We provide more qualitative results in supplementary material for demonstrating the physical plausibility of our synthesized humanoid motions.
Figure 6: Qualitative Results of motions synthesized by our Skeleton2Humanoid system. The top row shows the skeleton motions (blue) after our test time adaptation stage (TTA), the middle row shows the matched humanoid motions (red) by our analytical inverse kinematics method (IK), the bottom row shows our final imitated humanoid motions (pink) by our curriculum residual force control policy $\pi_{\text{CRP}}$. Ground penetration is highlighted with boxes.

|        | length | FP      | FQ      | JQ      | SM      | L2P     |
|--------|--------|---------|---------|---------|---------|---------|
|        | 5      | 15      | 30      | 5       | 15      | 30      | 5        | 15      | 30      |
| baseline[13] | -0.47  | -0.43   | -1.43   | 1.04%   | 4.12%   | 8.37%   | 0.68%    | 1.25%   | 2.08%   | 20.8 | 20.8 | 20.2 | 0.19 | 0.55 | 1.16 |
| baseline w/ TTA | -0.04  | -0.29   | -1.13   | 0.89%   | 2.91%   | 6.57%   | 0.07%    | 1.08%   | 1.79%   | 20.6 | 20.6 | 19.9 | 0.16 | 0.53 | 1.13 |
| baseline w/ TTA,IK | -0.24  | -0.63   | -1.73   | 6.28%   | 7.70%   | 11.3%   | 1.49%    | 1.80%   | 2.53%   | 20.6 | 20.6 | 20.0 | 0.23 | 0.56 | 1.16 |
| baseline w/ TTA,IK,CRP | -0.06  | -0.04   | -0.08   | 1.48%   | 0.73%   | 0.80%   | 0.35%    | 0.16%   | 0.16%   | 20.5 | 20.5 | 19.8 | 0.34 | 0.68 | 1.27 |

Table 1: Evaluation of our Skeleton2Humanoid i.e. baseline w/ TTA,IK,CRP on Noobjects LaFAN1 dataset. “TTA”, “IK” and “CRP” represent the test time adaptation, analytical inverse kinematics and curriculum residual force control policy respectively.

4.2 Evaluation of Each Stage

In this section, we concretely analyze the effect of each stage in our Skeleton2Humanoid system.

4.2.1 Evaluation of Test Time Motion In-betweening Network Adaptation Stage. We evaluate our test time adaptation stage on the test set of the whole LaFAN1 dataset [13]. Since no domain generalization approach or test time adaptation approach has been proposed for motion synthesis tasks, we compare against the state-of-the-art motion in-betweening approaches including Kaufmann et al. [17], duan et al. [7] and our implemented baseline Harvey et al. [13]. We select L2Q, L2P, FP, FQ, JQ, SM as evaluation metrics.

Results and discussion The results are shown in Table 3 and Table 2. First, As shown in Table 2, we observe that our test time adaptation stage (TTA) outperforms the baseline approach regarding all physics-based metrics. It shows that TTA can help reduce the physical artifacts and produce more physically-plausible motions for the humanoid character to imitate. Second, we can see that with our test time adaptation stage, our adapted motion in-betweening network significantly outperforms comparison approaches regarding all accuracy-based metrics in Table 3. We also achieve a new state-of-the-art performance in the L2Q and the L2P metrics. The ablation (w/o ablation, i.e. w/o smooth loss) shows that the motion smooth loss is also useful for test time adaptation.

4.2.2 Evaluation of Skeleton to Humanoid Matching Stage. To evaluate our analytical inverse kinematics (IK) method, we report the matching performance on the synthesized in-betweening motions (transition lengths: 30 frames) by TTA on the test set of Noobject LaFAN1 dataset by computing the mean per joint position error (MPJPE) and inference time. Our analytical inverse kinematics method achieves excellent performance in terms of precision (MPJPE: 11.29mm) and speed (0.122s/sequence). It is worth noting that even our method precisely transfers the skeleton motions to humanoid motions, it still causes a slight performance degradation problem in the L2P metric as shown in table 1 (baseline w/ TTA, IK). This shows that current state-of-the-art motion in-between methods can already predict transition motions with very high accuracy, hence a slight difference in poses will sensitively influence the prediction accuracy.

4.2.3 Evaluation of Motion Imitation based on RL Stage. Now we detailedly study the performance of our proposed curriculum residual force control policy $\pi_{\text{CRP}}$, we evaluate the motion imitation performance on the synthesized humanoid motions with transition
Table 2: Evaluation of test time adaptation stage in terms of physics-based metrics on LaFAN1 dataset. "TTA" represents the test time adaptation.

| Method           | FP 5 | FP 15 | FP 30 | FQ 5 | FQ 15 | FQ 30 | JQ 5 | JQ 15 | JQ 30 | SM 5 | SM 15 | SM 30 |
|------------------|------|-------|-------|------|-------|-------|------|-------|-------|------|-------|-------|
| baseline [13]    | -0.039 | -0.351 | -1.186 | 0.884% | 3.49% | 7.04% | 0.538% | 1.02% | 1.71% | 20.7 | 20.9 | 20.6  |
| baseline w/ TTA  | -0.036 | -0.235 | -0.930 | 0.775% | 2.42% | 5.52% | 0.527% | 0.861% | 1.46% | 20.5 | 20.5 | 20.0  |

Table 3: Evaluation of test time adaptation stage in terms of accuracy-based metrics on LaFAN1 dataset. "TTA" represents our test time adaptation. "ablation" represents applying test time adaptation without motion smoothness loss.

| Method                  | L2Q | L2P |
|-------------------------|-----|-----|
|                          | 5   | 15  | 30  | 5   | 15  | 30  |
| Zero-Vel                | 0.56 | 1.10 | 1.51 | 1.52 | 3.69 | 6.60 |
| Interp                  | 0.22 | 0.62 | 0.98 | 0.37 | 1.25 | 2.32 |
| Kaufmann et al. [17]    | 0.49 | 0.60 | 0.78 | 0.84 | 1.07 | 1.53 |
| duan et al. [7]         | 0.18 | 0.37 | **0.61** | 0.23 | 0.56 | **1.06** |
| baseline [13]           | 0.17 | 0.42 | 0.69 | 0.23 | 0.65 | 1.28 |
| baseline [13]"          | 0.138 | 0.373 | 0.667 | 0.183 | 0.548 | 1.167 |
| baseline w/ ablation    | 0.134 | 0.367 | 0.662 | 0.162 | 0.527 | 1.142 |
| baseline w/ TTA         | **0.133** | **0.367** | 0.661 | **0.162** | **0.526** | 1.140 |

Table 4: Evaluation of our curriculum residual force control policy \( \pi_{\text{CRP}} \) on Noobjects LaFAN1 dataset. "CurForce" represents applying curriculum learning for residual forces. "TwoFrameFeats" represents utilizing state information from target poses of next two frames.

| Method               | L2P (\( \downarrow \)) | MPJPE (\( \downarrow \)) |
|----------------------|------------------------|--------------------------|
| UHC [50]             | 1.307                  | 31.68mm                  |
| CRP w/o TwoFrameFeats| 1.281                  | 29.16mm                  |
| CRP w/o CurForce     | 1.292                  | 29.95mm                  |
| CRP (Ours)           | 1.272                  | **27.63mm**              |

Figure 7: Visualization of the reference motions (red humanoids) and imitated motions (pink humanoids) by our curriculum residual force control policy \( \pi_{\text{CRP}} \).

extracts informative features from more frames to learn control that advances the character to the next pose.

We also provide qualitative results in Fig. 7, we can see that our curriculum residual force control policy can imitate many complex motions accurately. More qualitative results can be seen in supplementary material.

5 CONCLUSION

In this paper, we proposed a system "Skeleton2Humanoid" which performs physics-oriented motion correction at test time by regularizing synthesized skeleton motions in a physics simulator. Concretely, our system consists of three sequential stages: (I) test time motion synthesis network adaptation, (II) skeleton to humanoid matching and (III) motion imitation based on reinforcement learning (RL). We verified our system on the motion-in-betweening task and showed our method can significantly improve the physical plausibility of synthesized motions.

ACKNOWLEDGMENTS

This work was supported by NSFC 62176159 and 62101326, China Postdoctoral Science Foundation (2022M712090), Natural Science Foundation of Shanghai 21ZR1432200 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102.
[50] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. 2021. SimPoE: Simulated Character Control for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7159–7169.

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

[52] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, et al. 2020. Generative tweening: Long-term inbetweening of 3d human motions. arXiv preprint arXiv:2005.06891 (2020).