Physics-Guided Human Motion Capture with Pose Probability Modeling

Jingyi Ju\textsuperscript{1,2*}, Buzhen Huang\textsuperscript{1,2*}, Chen Zhu\textsuperscript{1,2}, Zhihao Li\textsuperscript{3} and Yangang Wang\textsuperscript{1,2†}

\textsuperscript{1}Southeast University
\textsuperscript{2}Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China
\textsuperscript{3}Huawei Noah’s Ark Lab

{jingyiju, hbz, yangangwang}@seu.edu.cn, zc1213856@163.com, zhihao.li@huawei.com

Abstract
Incorporating physics in human motion capture to avoid artifacts like floating, foot sliding, and ground penetration is a promising direction. Existing solutions always adopt kinematic results as reference motions, and the physics is treated as a post-processing module. However, due to the depth ambiguity, monocular motion capture inevitably suffers from noises, and the noisy reference often leads to failure for physics-based tracking. To address the obstacles, our key-idea is to employ physics as denoising guidance in the reverse diffusion process to reconstruct physically plausible human motion from a modeled pose probability distribution. Specifically, we first train a latent gaussian model that encodes the uncertainty of 2D-to-3D lifting to facilitate reverse diffusion. Then, a physics module is constructed to track the motion sampled from the distribution. The discrepancies between the tracked motion and image observation are used to provide explicit guidance for the reverse diffusion model to refine the motion. With several iterations, the physics-based tracking and kinematic denoising promote each other to generate a physically plausible human motion. Experimental results show that our method outperforms previous physics-based methods in both joint accuracy and success rate. More information can be found at https://github.com/Me-Ditto/Physics-Guided-Mocap.

1 Introduction

Human motion capture is a fundamental task in sports broadcasting, human behavior understanding, and virtual reality, which require the accurate perception of human pose, position, and contact. Previous kinematics-based works [Kocabas et al., 2020; Arnab et al., 2019; Kanazawa et al., 2019; Rempe et al., 2021] neglect the physical laws when exploring human motion capture from monocular videos and images. As shown in Fig. 1 (a), even state-of-the-art monocular kinematics-based motion capture suffers from artifacts (e.g., floating, foot sliding, and ground penetration) due to the occlusion and depth ambiguity.

Figure 1: The kinematics-based approaches (a) suffer from artifacts, and dynamics-based works (b) encounter tracking failure, while our method (c) can reconstruct physically plausible human motion with a high success rate.

To tackle this problem, recent works introduce physical laws in human motion capture. The optimization-based framework [Huang et al., 2022a; Shimada et al., 2020] relieves artifacts by solving a highly-complex formulation. Others rely on Reinforcement Learning (RL) [Yuan et al., 2021; Luo et al., 2021] with non-differentiable physics simulators to obtain a physically plausible human motion. However, these methods all use a physical character to track the kinematic motion, and the physics is treated as a post-processing module. The noises in the kinematic motion always lead to tracking failure and thus result in a low success rate.

To address these limitations, our key idea is to employ physics as denoising guidance in a reverse diffusion process to reconstruct physically plausible human motion from modeled pose probability distributions. Thus, physics can guide the denoising process to progressively improve the motion quality. Nonetheless, its implementation still faces several technical obstacles. First, since highly-precise simulators are non-differentiable, the physics module cannot be incorporated into the network to provide explicit gradients to optimize human motions. In the generative task, PhysDiff [Yuan et al., 2022] directly uses a tracked motion as the input of the reverse diffusion step to avoid artifacts. However, this method also requires relatively high-quality reference motion. In addition, the strategy cannot be directly
applied in motion capture since it does not consider 2D observations. Second, the denoising models [Ho et al., 2020; Song et al., 2020a] always start from the standard gaussian distribution to generate a sample [Li et al., 2022b; Ramesh et al., 2022; Zhang et al., 2022; Yuan et al., 2022], which ignores the prior knowledge from image observations and may require thousands of denoising steps to reconstruct a satisfactory motion.

To fully utilize motion prior knowledge to improve the denoising efficiency, we first train a latent gaussian model based on Variational Autoencoder (VAE) [Kingma and Welling, 2013]. With the trained VAE encoder, the image features are mapped to a series of gaussian distributions to reflect the 3D probabilistic motion. The modeled probabilistic distributions can be used as good initial values to facilitate the denoising process. To alleviate the artifacts in the reconstructed motion, we further propose a physics module in the reverse diffusion model to provide implicit guidance for the denoising. Different from PhysDiff [Yuan et al., 2022], we utilize the discrepancies between the tracked motion and image observation to guide the reverse diffusion process in the next timestep. Specifically, we project the tracked joint positions to the 2D image plane and calculate the projection loss gradients. Then, we use the combination of the gradients and image features as a condition, and feed the tracked motion to the next reverse diffusion step. After several iterations, physics-based tracking and kinematic denoising can promote each other to obtain a physically plausible human motion. The main contributions of this paper are as follows:

- We construct a physical guidance to combine the physics and image observations for the reverse diffusion process to progressively promote a physically plausible human motion capture.
- We propose a VAE-based latent gaussian distribution to facilitate the reverse diffusion process with motion prior knowledge.
- We incorporate physics and kinematics in the same framework to improve joint accuracy and success rate for physics-based human motion capture.

2 Related Work

2.1 Kinematics-Based Motion Capture

Previous monocular kinematics-based motion capture leverages 3D pose estimation [Li et al., 2021; Li et al., 2022c; Song et al., 2020b] for each frame to construct the human motion. They cannot obtain the temporal information among frames, which leads to obvious jittering. Recent works [Kocabas et al., 2020; Luo et al., 2020] exploit the temporal context of human motion for better temporal consistency. These methods encounter global inconsistency since they can only produce a root-relative motion. Several approaches [Arnab et al., 2019; Xiang et al., 2019] adopt smooth priors over time to reduce jittering. However, the smooth constraints may result in over-smooth and foottskate. Since the aforementioned methods do not consider depth ambiguity and occlusion from monocular motion capture, recent diffusion-based works [Gong et al., 2022a; Choi et al., 2022; Holmquist and Wandt, 2022] model the uncertainty of 2D-3D lifting for 3D pose estimation. Although the aforementioned approaches achieve great performance on kinematic metrics, they still encounter artifacts since they ignore the physics laws.

2.2 Physics-Based Motion Capture

To relieve the artifacts of human motion capture, [Wei and Chai, 2010; Vondrak et al., 2012; Zell et al., 2017; Shimada et al., 2020; Shimada et al., 2021; Li et al., 2022d; Yi et al., 2022] adopt optimization to obtain the physical forces to induce the human motion, which results in high approximation errors since they do not utilize the highly-precise non-differentiable simulators. While Neural Motion [Huang et al., 2022a] generates a learned motion distribution for sampling-based motion control with supervision from a non-differentiable simulator. Others [Yuan et al., 2021; Luo et al., 2021; Yuan and Kitani, 2020; Yu et al., 2021; Peng et al., 2018] use Reinforcement Learning (RL) with non-differentiable simulators to obtain physically plausible human motion. However, all the approaches adopt a strong assumption that the reference motion is accurate. Luo et al. [Luo et al., 2022] adopts current character state and environmental cues to promote tracking when the reference pose is unreliable. The accuracy of the estimated motion relies on the 2D observations. Since all these approaches implement physics as a post-processing process, the character always fails to track due to the noises in the reference motion. In contrast, we propose a physics module in the reverse diffusion model to provide implicit guidance for the denoising. In this case, the physics-based tracking and kinematic denoising can promote each other for physically plausible human motion capture.

2.3 Multi-Hypothsis Estimation

Human motion capture from monocular videos and images is an ill-posed problem, for which directly regressing a determinate solution may be inaccurate [Kocabas et al., 2020; Arnab et al., 2019; Kanazawa et al., 2019]. Multi-hypothesis methods [Huang et al., 2022b; Li and Lee, 2019; Jahangiri and Yuille, 2017] are proposed to represent the uncertainty of 2D-3D lifting. Recently, Sharma et al. [Sharma et al., 2019] adopted a conditional VAE to predict 3D pose candidates conditioned on detected 2D poses. Wehrbein et al. [Wehrbein et al., 2021] employs normalizing flow to model the posterior distribution of 3D poses. However, it is difficult to select the best 3D pose from multi-hypothesis. Unlike these works, we propose a physics module to provide implicit guidance for the denoising process to promote a more accurate human motion.

3 Method

We aim to reconstruct the physically plausible 3D human motion from monocular videos. We first design a latent gaussian distribution encoded from image features to provide initial values to facilitate the reverse diffusion process (Sec. 3.2). To alleviate the artifacts in the reconstructed motion, we further propose a physics module to combine the physics and observed 2D poses to guide the denoising of reverse
diffusion (Sec. 3.3). With the constructed framework, the physics and kinematics can progressively promote each other to obtain physically plausible motions with a high success rate (Sec. 3.4).

3.1 Preliminaries

Motion Representation. The 2D poses with corresponding confidence detected by AlphaPose [Fang et al., 2017] from color images $I^{1:H}$ are defined $\mathcal{D}_{2D} = \{\mathcal{P}^{2D}_{j} \in \mathbb{R}^{J \times 3} \}_{h=1}^{H}$, where $H$ is the length of motion and $J$ is the number of joints. We adopt SMPL model [Loper et al., 2015] to represent the kinematic human motion $x^{1:H} = \{x^{h} \}_{h=1}^{H}$, where $x$ denotes the parameters of human pose $\theta$ in 6D representation [Zhou et al., 2019]. We also regress SMPL shape $\beta$ and translation $\tau$. The physics-based human motion is defined as $\tilde{x}^{1:H}$.

Physics-Based Tracking. We briefly introduce the physical imitation. The character in the physics engine is created based on the SMPL kinematic tree and estimated body shape $\beta$. We construct convex hulls to approximate mesh [Luo et al., 2012] and shares the same pose parameters with SMPL model. The physical imitation aims to control the character to track the reference motion $x^{1:H}$. With a trained policy $\pi(a^{h} \mid s^{h}, \tilde{x}^{h+1})$, we can sample an action $a^{h}$ according to the current state $s^{h}$ and reference pose $\tilde{x}^{h+1}$ to control the character to move to the next state $s^{h+1}$. The state $s^{h} = (\tilde{x}^{h}, \dot{\tilde{x}}^{h})$ contains the character’s current pose $\tilde{x}^{h}$ and joint velocity $\dot{\tilde{x}}^{h}$. Finally, the physically plausible human motion $\tilde{x}^{1:H}$ can be obtained from the simulated character state. More details can be found in the supplementary material.

Diffusion Model. The diffusion model [Tevet et al., 2022] can generate target data from a simple noise distribution under a condition. Specifically, the forward diffusion process gradually adds infinitesimal gaussian noise $\epsilon$ on the data $x^{1:H} \sim q(x^{1:H})$ at timestep $t$, which is formulated as:

$$q(x^{1:H} \mid \tilde{x}^{1:H}) = \sqrt{\alpha_{t}x^{1:H}} + \sqrt{1 - \alpha_{t}} \epsilon \sim \mathcal{N}(0, I),$$

(1)

where $\alpha_{t}$ is a manually designed constant hyper-parameter, and $\tilde{x}^{1:H}$ is the ground truth original data. The reverse diffusion process samples an initial input from the standard gaussian distribution and progressively denoises it to the target data under the guidance of condition $c$, which is defined as:

$$q(x^{1:H} \mid x^{1:H}_t, c) = \mathcal{N}(x^{1:H}_{t-1}; \mu_{t}(x^{1:H}_t, c), \tilde{\beta}_{t}I),$$

(2)

where $\mu_{t}(x^{1:H}_t, c)$ is the estimated mean by a neural network and $\tilde{\beta}_{t}$ is the variance which is calculated by the hyper-parameters $\beta_{t}$, $\tilde{\alpha}_{t}$ and $\tilde{\alpha}_{t-1}$.

3.2 Probabilistic Modeling of 3D Human Motion

In this work, we formulate the motion capture as a reverse diffusion process conditioned on image observations. The previous diffusion models [Choi et al., 2022; Gong et al., 2022a] always start from the standard gaussian distribution, which requires thousands of reverse diffusion timesteps to denoise the input to produce a desired result. Although a recent work [Gong et al., 2022a] separately models the initial distribution for each pose coordinate, it neglects the correlation among different joints. To improve efficiency, we leverage prior knowledge from image features to provide initial values for the reverse diffusion process.
Inspired by the previous works [Huang et al., 2022c; Huang et al., 2021], we first extract image features \( F^{1:H} \) with a backbone network, and then fed the features into a VAE model to output 3D motion \( x^{1:H} \), translation \( \tau \) and body shape \( \beta \). Since we adopt a Gate Recurrent Unit (GRU) for the VAE encoder, the correlated gaussian distributions \( \{ N(\mu_0(F^h), \sigma_0(F^h)) \}_{h=1}^H \) in the latent space can describe the motion prior knowledge. During the training phase, we maximize the Evidence Lower Bound (ELBO) to train the model:

\[
\log p_\theta(x^{1:H}) \geq \mathbb{E}_{q_\phi(z | F^{1:H})} \left[ \log p_\theta(x^{1:H} | z) \right] - D_{KL}(q_\phi(z | F^{1:H}) || p_\theta(z)).
\]

(3)

The overall loss function is:

\[
\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{motion}} + \mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{reproj}} + \mathcal{L}_{\text{kl}},
\]

(4)

where \( \mathcal{L}_{\text{motion}} \) and \( \mathcal{L}_{\text{shape}} \) are used to supervise the estimated human motion and shape:

\[
\mathcal{L}_{\text{motion}} = \sum_{h=1}^H \left\| x^h - \bar{x}^h \right\|^2,
\]

(5)

\[
\mathcal{L}_{\text{shape}} = \left\| \beta - \bar{\beta} \right\|^2,
\]

(6)

where \( \bar{x}^h \) and \( \bar{\beta} \) are ground truth motion and shape.

\[
\mathcal{L}_{\text{joint}} = \sum_{h=1}^H \left\| J^h_{3D} - \bar{J}^h_{3D} \right\|^2,
\]

(7)

where \( \bar{J}^h_{3D} \) and \( J^h_{3D} \) are the ground truth and predicted 3D joint positions generated from the corresponding motion. We also follow CLIFF [Li et al., 2022c] to supervise the predicted joints in the original camera coordinates with detected 2D poses:

\[
\mathcal{L}_{\text{reproj}} = \frac{1}{H} \sum_{h=1}^H \left\| \Pi(J^h_{3D}) - \bar{p}^h_{2D} \right\|^2,
\]

(8)

where \( \Pi \) denotes the projection operation.

\[
\mathcal{L}_{\text{kl}} = KL(q_\phi(z | F^{1:H}) || N(0, I)),
\]

(9)

which is used to push the output of the encoder to approximate the gaussian distribution.

After the training, we freeze the network parameters of the VAE encoder and use it to generate specific gaussian distributions from image features for the reverse diffusion process.

### 3.3 Physics-Guided Motion Diffusion

Although the human motion sampled from the encoded distributions can match the image observations, it is still physically implausible due to the occlusion and depth ambiguity. Thus, we incorporate physics to refine the kinematic motion. Previous works [Yuan et al., 2021; Shimada et al., 2020] directly use the physics module as a post-processing to track the kinematic motion. However, the noises in kinematic motions always result in tracking failure. To address the obstacle, we use physics to guide the reverse diffusion process to denoise the motion.

Unlike previous diffusion-based pose estimation [Choi et al., 2022; Holmqvist and Wandt, 2022], we start from the modeled probabilistic distributions, which can provide prior knowledge to improve the efficiency of the denoising process. To train the diffusion model, we first use the trained VAE encoder to produce gaussian distributions \( N(\mu_0(F^h), \sigma_0(F^h)) \) for the diffusion framework. The difference from the standard diffusion process Eq. (1) is that we sample the noise \( \epsilon \) from the encoded distributions rather than the standard gaussian distribution. We then gradually add the sampled noises on the 3D motion \( \bar{x}^{1:H}_0 \) towards the uncertainty distribution \( N(\mu^{1:H}_0, \sigma^{1:H}_0) \).

\[
q(x^{1:H}_t | \bar{x}^{1:H}_0, c_t) = N(x^{1:H}_{t-1}, \mu_0(\bar{x}^{1:H}_0, c_t), \bar{\beta}_t 1).
\]

(10)

In the reverse diffusion process, we train a network to denoise the noisy motion \( x^{1:H}_0 \) to the original data \( x^{1:H}_0 \). Since the distributions have encoded prior knowledge, the noisy motion only contains a few artifacts and is still close to the real motion. Although directly applying the physics-based tracking on the noisy reference motion may fail, the discrepancies between the tracked motion and 2D poses can provide implicit guidance for the next reverse diffusion step to optimize the kinematic motion. Thus, we combine the projection loss gradient and image features as a condition, and feed the tracked motion to the network to predict the distributions of the next step.

\[
q(x^{1:H}_t | \bar{x}^{1:H}_0, c_t) = N(x^{1:H}_{t}, \mu_0(\bar{x}^{1:H}_0, c_t), \bar{\beta}_t 1).
\]

(11)

We follow Ramesh et al. [Ramesh et al., 2022] to make the diffusion model to predict the target data \( x^{1:H}_0 \), and then construct the mean of the distribution \( \mu_0(\bar{x}^{1:H}_0, c_t) \) in timestep \( t - 1 \) according to \( \bar{x}^{1:H}_0 \). The condition \( c_t \) is a concatenated vector of image features and projection loss gradient information. The image features vector is encoded from extracted image features with a linear layer, and the gradient vector records the gradient of each frame.

\[
\mathcal{L}_t = \frac{\partial}{\partial J^h_{3D}} \left\| \Pi(J^h_{3D}) - \bar{p}^h_{2D} \right\|^2.
\]

(12)

We mask the gradient for the frames that are not successfully tracked. We also fill the gradient factor with 0 when the physical guidance is not executed.

**Training Procedure.** For the physics-based tracking, we first follow [Luo et al., 2021] to train a policy \( \pi \) on motion capture datasets by maximizing the expected return, and the parameters of the trained policy are fixed in the diffusion model training. To train the diffusion model, the following loss function is adopted:

\[
\mathcal{L} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{reproj}}.
\]

(13)

\[
\mathcal{L}_{\text{diff}} = \mathbb{E}_{z^{1:H}_0 \sim q(z^{1:H}_0), t \in [1, T]} \left[ \left\| \bar{x}^{1:H}_0 - F(\bar{x}^{1:H}_t, t, c_t) \right\|^2 \right],
\]

(14)

where \( F(\bar{x}^{1:H}_t, t, c_t) \) is the output of the neural network. The functions \( \mathcal{L}_{\text{joint}} \) and \( \mathcal{L}_{\text{reproj}} \) are the same as Eq. (7) and Eq. (8).
Table 1: Our method can achieve competitive performance in terms of joint accuracy. NeuralPhysCap, PoseTriplet, and D&D output human motion from the kinematics-based network with implicit physical laws, which may also contain artifacts.

3.4 Reverse Diffusion for Motion Capture
Once the networks are trained, we can construct a physics-based human motion capture framework. We first extract image features with a backbone network, and then predict 2D poses for each frame. Then, the latent gaussian distributions can be generated from the extracted image features with the trained VAE encoder. A noisy initial motion is sampled from the distributions for the reverse diffusion process. By applying the physics-based tracking, we can combine the physics information with the image features to guide the denoising. Finally, the physically plausible human motion can be obtained after several iterations.

4 Experiments
4.1 Metrics
We report the Mean Per Joint Position Error (MPJPE) and the MPJPE after aligning the prediction with ground truth using Procrustes Analysis (PA-MPJPE) to evaluate joint accuracy. We use the 3D extension of the Percentage of Correct keypoints (PCK) at the threshold of 150mm to evaluate the 3D joint position accuracy. We follow PhysCap [Shimada et al., 2020] to measure motion jitter error by $e_s$, which is the deviation of joint velocity between predicted output and ground truth. The $e_s$ and its standard deviation $\sigma_s$ are adopted to assess the motion smoothness. To evaluate foot contact, we adopt $e_{f,z}$ proposed in [Xie et al., 2021], which is the foot position error on the z-axis.

4.2 Datasets
Human3.6M [Ionescu et al., 2013] is an indoor dataset for human motion capture. The videos are captured at 50Hz which includes 7 subjects. Following previous works [Yuan et al., 2021; Shimada et al., 2020], we use 2 subjects (S9, S11) for evaluation, and the others are used for training. We convert the dataset to 30Hz to reduce redundancy.

3DOH [Zhang et al., 2020] is the first dataset to handle the object occluded human body estimation, which contains 3D motions in occluded scenarios. We use the sequence 0013, 0027, 0029 to evaluate our method in occlusion cases.

MPI-INF-3DHP [Mehta et al., 2017] is a single-person 3D pose dataset. Following previous works [Gong et al., 2022b; Shimada et al., 2021], we use its test set to demonstrate the generalization of our method.

4.3 Comparison with State-of-the-Art Methods
We compared our method with state-of-the-art dynamics-based human motion capture approaches on Human3.6M dataset, and their average errors are shown in Tab. 1. All approaches in Tab. 1 are dynamics-based methods. EgoPose, SimPoE, and our method rely on RL policy to control the character. When the kinematic motion is inaccurate, EgoPose and SimPoE may fail to track and require re-initialization. In contrast, our method can progressively denoise the artifacts in the reference motion during the reverse diffusion process and promote successful tracking. Thus, our method outperforms these two techniques. In addition, we found that D&D can achieve the best performance in terms of MPJPE and PA-MPJPE. However, D&D uses a kinematics-based network to implicitly learn the physical laws, from which the output motion may still contain artifacts. We also use other metrics like motion smoothness and foot contact error to measure the physical plausibility. The results show that our method achieves state-of-the-art on most of the metrics. We further conduct a qualitative comparison with both kinematics-based and physics-based methods in Fig. 3. VIBE [Kocabas et al., 2020] and HuMoR [Rempe et al., 2021] are recent works that exploit temporal information to obtain kinematic motions. The results show that the reconstructed motion in 3D scenes cannot get accurate contact. Besides, PhysCap is an optimization-based method that can obtain almost physically plausible human motion. However, it also does not consider the scene interactions and cannot utilize 2D observations in physics-based tracking. Thus, the results may hover above the floor and deviate from 2D poses.
Figure 3: Qualitative comparison with other methods. The results show that our method can achieve physically plausible and accurate human motion from monocular videos.

We further conduct experiments on 3DOH dataset. Since the dataset contains a lot of object occlusions, it is more difficult to reconstruct an accurate motion. In Fig. 3, VIBE and HuMoR cannot get satisfactory results due to the depth ambiguity and occlusions. Although Neural MoCon outputs more precise joint positions, it uses a skeletal character and cannot reconstruct accurate body contact. PhysCap, SamCon, and Neural MoCon formulate the physics-based motion capture as a trajectory optimization, and the tracking results strongly depend on the quality of the estimated reference motion. In contrast, our method can adjust both the kinematic and physical motion in the reverse diffusion process. The quantitative results in Tab. 1 show that our method significantly outperforms these baseline methods in all metrics.

To evaluate the generalization of our method, we use 3DHP dataset as a benchmark. The results in Fig. 4 show that our method can obtain accurate motion with precise contact on different scenes. We also compared our method with PhysCap, NeuralPhysCap, and PoseTriplet on 3DHP. The results in Tab. 1 show that our method gets more accurate joint positions than PhysCap and NeuralPhysCap, but inferior to PoseTriplet. The reason is that PoseTriplet trains the kinematics-based network assisted by physics but discards the physics module in the pose estimation. Thus, the results from the trained network may also contain artifacts.

<table>
<thead>
<tr>
<th>Method</th>
<th>step = 1</th>
<th>step = 5</th>
<th>step = 10</th>
<th>step = 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard w/o tracking</td>
<td>180.3</td>
<td>68.3</td>
<td>43.5</td>
<td>41.3</td>
</tr>
<tr>
<td>VAE w/o tracking</td>
<td>55.5</td>
<td>40.2</td>
<td>40.1</td>
<td>39.8</td>
</tr>
<tr>
<td>standard + phys (s=3)</td>
<td>–</td>
<td>77.6</td>
<td>47.1</td>
<td>43.7</td>
</tr>
<tr>
<td>VAE + phys (s=3)</td>
<td>–</td>
<td>41.3</td>
<td>42.1</td>
<td>42.8</td>
</tr>
</tbody>
</table>

Table 3: Ablation on the latent gaussian distribution on Human3.6M dataset. The standard gaussian distribution requires more denoising steps to obtain a satisfactory motion, while the latent gaussian distribution can directly sample a plausible motion. The numbers are PA-MPJPE in mm.

4.4 Ablation Study

**Latent Gaussian Distribution.** The conventional diffusion model requires a lot of denoising steps to reconstruct a satisfactory motion, which is inefficient for the motion capture task. To facilitate the reverse diffusion process, we propose a VAE-based latent gaussian distribution to employ the motion prior knowledge for a good initial value. In Tab. 3, we compared the proposed latent distribution with the standard gaussian distribution in our motion capture framework. The results show that we can directly sample plausible motions from the encoded distributions. Although the sampled motion may contain a lot of artifacts due to the depth ambiguity, it still has a relatively high joint accuracy. In contrast, we require more denoising steps to reconstruct a motion when we
Figure 4: Qualitative results on Human3.6M, 3DOH, and 3DHP dataset. Our method can produce physically plausible human motion with accurate contact in different scenes.

<table>
<thead>
<tr>
<th>Method</th>
<th>PA-MPJPE</th>
<th>ϵₜ</th>
<th>success</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAE w/o tracking</td>
<td>58.6</td>
<td>19.8</td>
<td>–</td>
</tr>
<tr>
<td>VAE w/ tracking</td>
<td>66.9</td>
<td>13.3</td>
<td>36.5%</td>
</tr>
<tr>
<td>VAE w/o tracking + denoise (T=1)</td>
<td>55.5</td>
<td>20.0</td>
<td>–</td>
</tr>
<tr>
<td>VAE w/o tracking + denoise (T=3)</td>
<td>46.2</td>
<td>14.3</td>
<td>–</td>
</tr>
<tr>
<td>VAE w/o tracking + denoise (T=5)</td>
<td>40.2</td>
<td>16.1</td>
<td>–</td>
</tr>
<tr>
<td>VAE + phys (s=1) + denoise (T=5)</td>
<td>43.7</td>
<td>9.9</td>
<td>66.7%</td>
</tr>
<tr>
<td>VAE + phys (s=2) + denoise (T=5)</td>
<td>41.6</td>
<td>4.7</td>
<td>83.4%</td>
</tr>
<tr>
<td>VAE + phys (s=3) + denoise (T=5)</td>
<td>41.3</td>
<td>3.5</td>
<td>90.3%</td>
</tr>
<tr>
<td>VAE + phys (s=5) + denoise (T=5)</td>
<td>41.4</td>
<td>3.1</td>
<td>90.7%</td>
</tr>
<tr>
<td>VAE + phys (s=3) + denoise (T=7)</td>
<td>41.1</td>
<td>3.4</td>
<td>90.9%</td>
</tr>
<tr>
<td>VAE + phys (s=3) + denoise (T=10)</td>
<td>42.1</td>
<td>3.3</td>
<td>90.0%</td>
</tr>
</tbody>
</table>

Table 4: Ablation studies. We study the physical guidance with different denoising steps. The tracking denotes applying physics-based tracking, and phys means physical guidance. T denotes the number of timesteps used in the reverse diffusion process, and s is the number of timesteps in which physical guidance is applied. VAE denotes that the reverse diffusion process starts from the latent gaussian distributions.

Use the standard gaussian as the initial distribution. In the first several steps, the sampled results are pure noises. The framework needs more than 10 steps to reconstruct the motion.

Physical Guidance. We study the physical guidance in this section. The conventional kinematics-based motion capture predicts the 3D motion from pure image features, which cannot avoid the artifacts due to the depth ambiguity. In Tab. 4, although the PA-MPJPE of the results sampled from the latent distribution is 58.6, it contains a lot of jitters. We can find that the diffusion model with only kinematics cannot remove the artifacts. Existing physics-based capture frameworks use physics as post-processing after the kinematics-based prediction. To demonstrate its weakness with a fair setting, we directly add physics-based tracking on the sampled motion. This strategy can alleviate the artifacts, but it has a low success rate since the noises in the kinematic motion always result in tracking failure. We further use the denoising framework to refine the kinematic motion. The results in Tab. 4 show that the kinematics-based diffusion model can improve the joint accuracy, but does not prevent the artifacts. By incorporating the physical guidance, the success rate can be significantly improved. In addition, the results in PA-MPJPE also demonstrate that the physics also provides the correct direction to enhance the motion capture accuracy. To compare the strategy adopted in PhysDiff, we remove the projection loss gradient condition, and use the tracked motion for the input of the next denoising step in Tab. 5. We found that the gradients can provide implicit guidance for the diffusion model to optimize the kinematic motion, and thus improve the success rate in the physics-based tracking.

Denoising Step. We also study the impact of the denoising step. In Tab. 4, we found that the joint accuracy can be significantly improved with the physical guidance. The performance increases at first with more times of physical guidance and then becomes stable. In addition, the gains are also declining with more than 7 denoising steps.

Table 5: Ablation study on the physical guidance. w/o guidance means that we remove the projection loss gradients in the condition.

<table>
<thead>
<tr>
<th>Method</th>
<th>PA-MPJPE</th>
<th>ϵₜ</th>
<th>success</th>
</tr>
</thead>
<tbody>
<tr>
<td>phys (s=3) w/o guidance</td>
<td>44.7</td>
<td>5.9</td>
<td>73.9%</td>
</tr>
<tr>
<td>phys (s=3) w/ guidance</td>
<td>41.3</td>
<td>3.5</td>
<td>90.3%</td>
</tr>
</tbody>
</table>

5 Conclusion

In this work, we formulate the physically plausible human motion capture as a reverse diffusion process. Latent gaussian distributions are built based on VAE to utilize the motion prior knowledge to facilitate the reverse diffusion. To employ physics information, we further construct a physics module to combine the physics and image observations to guide the denoising. With the physical guidance, physics and kinematics can promote each other to progressively reconstruct a high-quality human motion. Experimental results on several human motion capture datasets demonstrate that our method can achieve competitive performance with a higher success rate.
Acknowledgments
This work was supported in part by the National Natural Science Foundation of China (No. 62076061) and the Natural Science Foundation of Jiangsu Province (No. BK20220127). The authors would like to thank Yuan Yang for the helpful discussion about physics-based imitation.

Contribution Statement
Jingyi Ju and Buzhen Huang contribute equally to this work. The corresponding author is Yangang Wang.

References


