Physics-Guided Human Motion Capture with Pose Probability Modeling

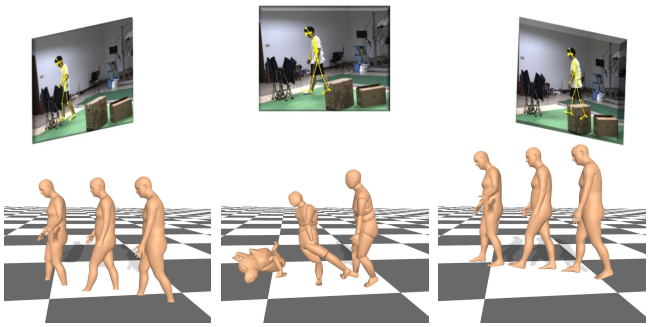
Jingyi Ju[1,2*], **Buzhen Huang**[1,2*], **Chen Zhu**[1,2], **Zhihao Li**[3] and **Yangang Wang**[1,2†]
[1]Southeast University
[2]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China
[3]Huawei Noah's Ark Lab
{jingyiju, hbz, yangangwang}@seu.edu.cn, 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.



(a) Kinematics-based (b) Physics-based (c) Ours

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,

*Equal contribution.

†Corresponding author. 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).

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 footskate. 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 Mocon [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-hypothesis 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

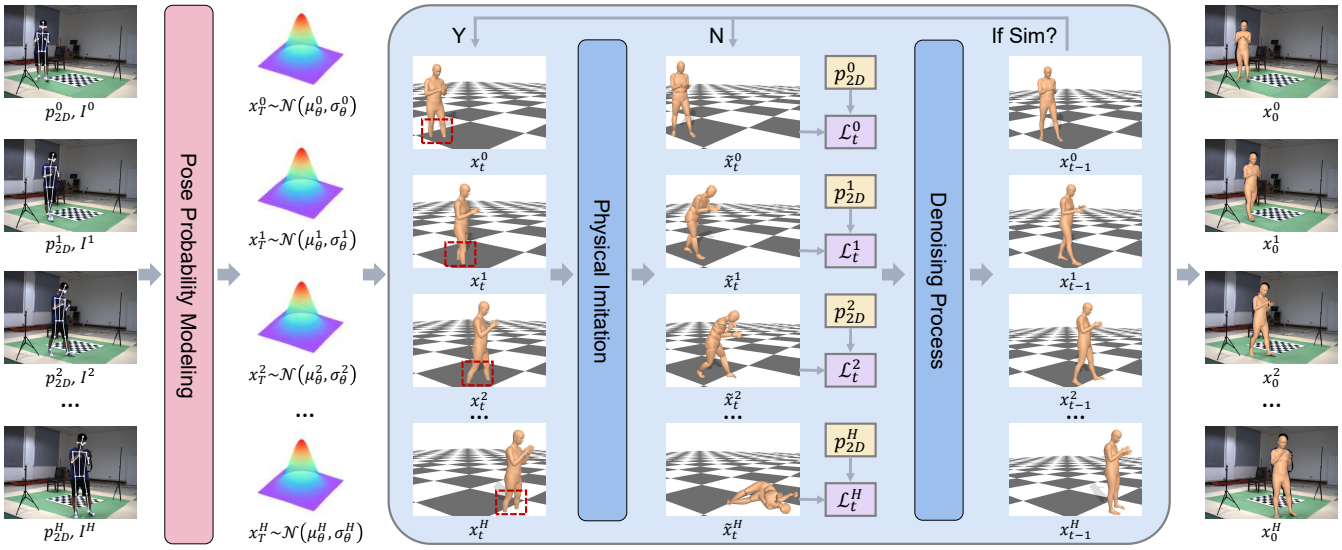


Figure 2: We formulate the physics-based motion capture as a reverse diffusion process. Given images and 2D poses estimated from off-the-shelf 2D pose detector, our method first regresses a series of gaussian distributions $\mathcal{N}(\mu_\theta, \sigma_\theta)$ from color images with a trained VAE encoder. We then sample a human motion from the encoded distributions and use it as the initial value for the diffusion model. To improve the physical plausibility and tracking success rate, we further propose a physical guidance that combines physics and 2D observations to guide the denoising. After several denoising steps, the physically plausible human motion can be obtained.

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 $\mathcal{P}_{2D} = \left\{p_{2D}^h \in \mathbb{R}^{J\times 3}\right\}_{h=1}^{H}$, 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 $x^{1:H} = \left\{x^h\right\}_{h=1}^{H}$, where x denotes the parameters of human pose θ in 6D representation [Zhou *et al.*, 2019]. We also regress SMPL shape β and translation τ. 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 β. We construct convex hulls to approximate mesh [Luo *et al.*, 2021], which can be simulated in MuJoCo [Todorov *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 $\pi(a^h \mid s^h, x^{h+1})$, we can sample an action a^h according to the current state s^h and reference pose 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 ϵ on the data $x_t^{1:H} \sim q(x_t^{1:H})$ at timestep t, which is formulated as:

$$q(x_t^{1:H} \mid \bar{x}_0^{1:H}) = \sqrt{\hat{\alpha}_t}\bar{x}_0^{1:H} + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathrm{I}), \quad (1)$$

where α_t is a manually designed constant hyper-parameter, and $\hat{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, and $\bar{x}_0^{1:H}$ 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:

$$q(x_{t-1}^{1:H} \mid x_t^{1:H}, c) = \mathcal{N}(x_{t-1}^{1:H}; \mu_\alpha(x_t^{1:H}, c), \tilde{\beta}_t \mathrm{I}), \quad (2)$$

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

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 $\mathcal{F}^{1:H}$ with a backbone network, and then fed the features into a

VAE model to output 3D motion $x^{1:H}$, translation τ and body shape β. Since we adopt a Gate Recurrent Unit (GRU) for the VAE encoder, the correlated gaussian distributions $\left\{\mathcal{N}\left(\mu_\theta\left(\mathcal{F}^h\right), \sigma_\theta\left(\mathcal{F}^h\right)\right)\right\}_{h=1}^{H}$ 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:

$$\begin{aligned} &\log p_\theta\left(x^{1:H}\right) \geq \mathbb{E}_{q_\phi}\left[\log p_\theta\left(x^{1:H} \mid \mathbf{z}\right)\right] \\ &-D_{\mathrm{KL}}\left(q_\phi\left(\mathbf{z} \mid \mathcal{F}^{1:H}\right) \| p_\theta(\mathbf{z})\right). \end{aligned} \tag{3}$$

The overall loss function is:

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{motion}} + \mathcal{L}_{\mathrm{shape}} + \mathcal{L}_{\mathrm{joint}} + \mathcal{L}_{\mathrm{reproj}} + \mathcal{L}_{\mathrm{kl}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{motion}}$ and $\mathcal{L}_{\mathrm{shape}}$ are used to supervise the estimated human motion and shape:

$$\mathcal{L}_{\mathrm{motion}} = \sum_{h=1}^{H} \left\| x^h - \bar{x}^h \right\|^2, \tag{5}$$

$$\mathcal{L}_{\mathrm{shape}} = \left\| \beta - \bar{\beta} \right\|^2, \tag{6}$$

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

$$\mathcal{L}_{\mathrm{joint}} = \sum_{h=1}^{H} \left\| J_{3D}^h - \bar{J}_{3D}^h \right\|^2, \tag{7}$$

where $\bar{J}_{3D}^h$ and J_{3D}^h 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:

$$\mathcal{L}_{\mathrm{reproj}} = \frac{1}{H} \sum_{h=1}^{H} \left\| \Pi\left(J_{3D}^h\right) - p_{2D}^h \right\|_2^2, \tag{8}$$

where Π denotes the projection operation.

$$\mathcal{L}_{kl} = KL(q_\phi(\mathbf{z} \mid \mathcal{F}^{1:H}) \| \mathcal{N}(0, \mathrm{I})), \tag{9}$$

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; Holmquist 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 $\mathcal{N}\left(\mu_\theta^{1:H}, \sigma_\theta^{1:H}\right)$ for the diffusion framework. The difference from the standard diffusion process Equ. (1) is that we sample the noise ϵ from the encoded distributions rather than the standard gaussian distribution. We then gradually add the sampled noises on the 3D motion $\bar{x}_0^{1:H}$ towards the uncertainty distribution $\mathcal{N}\left(\mu_\theta^{1:H}, \sigma_\theta^{1:H}\right)$.

$$q(x_t^{1:H} \mid \bar{x}_0^{1:H}) = \sqrt{\hat{\alpha}_t}\bar{x}_0^{1:H} + \sqrt{1-\hat{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}\left(\mu_\theta^{1:H}, \sigma_\theta^{1:H}\right). \tag{10}$$

In the reverse diffusion process, we train a network to denoise the noisy motion $x_T^{1:H}$ to the original data $x_0^{1:H}$. 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.

$$q\left(x_{t-1}^{1:H} \mid \tilde{x}_t^{1:H}, c_t\right) = \mathcal{N}\left(x_{t-1}^{1:H}; \mu_\alpha\left(\tilde{x}_t^{1:H}, c_t\right), \tilde{\beta}_t \mathrm{I}\right). \tag{11}$$

We follow Ramesh *et al*. [Ramesh *et al.*, 2022] to make the diffusion model to predict the target data $\check{x}_0^{1:H}$, and then construct the mean of the distribution $\mu_\alpha\left(\tilde{x}_t^{1:H}, c_t\right)$ in timestep $t-1$ according to $\check{x}_0^{1:H}$. 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.

$$\mathcal{L}_t^h = \frac{\partial \left\| \Pi\left(\tilde{J}_{3D}^h\right) - p_{2D}^h \right\|_2^2}{\partial \tilde{J}_{3D}^h}. \tag{12}$$

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 π 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:

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \mathcal{L}_{\mathrm{joint}} + \mathcal{L}_{\mathrm{reproj}}. \tag{13}$$

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{\bar{x}_0^{1:H} \sim q\left(\bar{x}_0^{1:H}\right), t \sim [1,T]} \left[\left\| \bar{x}_0^{1:H} - F\left(\tilde{x}_t^{1:H}, t, c_t\right) \right\|_2^2 \right], \tag{14}$$

where $F\left(\tilde{x}_t^{1:H}, t, c_t\right)$ is the output of the neural network. The functions $\mathcal{L}_{\mathrm{joint}}$ and $\mathcal{L}_{\mathrm{reproj}}$ are the same as Equ. (7) and Equ. (8).

Method	Human3.6M					3DOH			3DHP	
	MPJPE↓	PA-MPJPE↓	$e_s \downarrow$	$\sigma_s \downarrow$	$e_{f,z} \downarrow$	MPJPE↓	PA-MPJPE↓	$e_s \downarrow$	MPJPE↓	PCK↑
EgoPose [Yuan and Kitani, 2019]	130.3	79.2	-	-	-	-	-	-	-	-
PhysCap [Shimada *et al.*, 2020]	97.4	65.1	7.2	6.9	-	107.8	93.3	12.2	104.4	83.9
Gärtner *et al*. [Gärtner *et al.*, 2022b]	84.0	56.0	-	-	-	-	-	-	-	-
DiffPhy [Gärtner *et al.*, 2022a]	81.7	55.6	-	-	-	-	-	-	-	-
SamCon [Liu *et al.*, 2015]	78.4	63.2	4.0	4.3	20.4	102.4	95.4	9.7	-	-
NeuralPhysCap[Shimada *et al.*, 2021]	76.5	58.2	4.5	6.9	-	-	-	-	99.1	85.5
Neural MoCon [Huang *et al.*, 2022a]	72.5	54.6	3.8	2.4	14.4	93.4	86.7	9.2	-	-
PoseTriplet [Gong *et al.*, 2022b]	68.2	45.1	-	-	-	-	-	-	**79.5**	**89.1**
Xie *et al*. [Xie *et al.*, 2021]	68.1	-	4.0	**1.3**	18.9	-	-	-	-	-
SimPoE [Yuan *et al.*, 2021]	56.7	41.6	-	-	-	-	-	-	-	-
D&D [Li *et al.*, 2022a]	**52.5**	**35.5**	-	-	-	-	-	-	-	-
Ours	55.4	41.3	**3.5**	2.1	**12.2**	**79.3**	**72.8**	**8.9**	83.6	88.1

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

In this section, we first introduce metrics (Sec. 4.1) and datasets (Sec. 4.2) used in the experiments. Then, the qualitative and quantitative comparisons with state-of-the-art methods are conducted in Sec. 4.3. Finally, we ablate important components to verify their effectiveness in Sec. 4.4.

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 σ_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.

Method	SamCon	Neural MoCon	UHC	**Ours**
success rate	76.2%	83.4%	84.1%	89.6%

Table 2: The success rate on 3DOH dataset. Our method significantly outperforms other physics-based works in terms of success rate.

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.

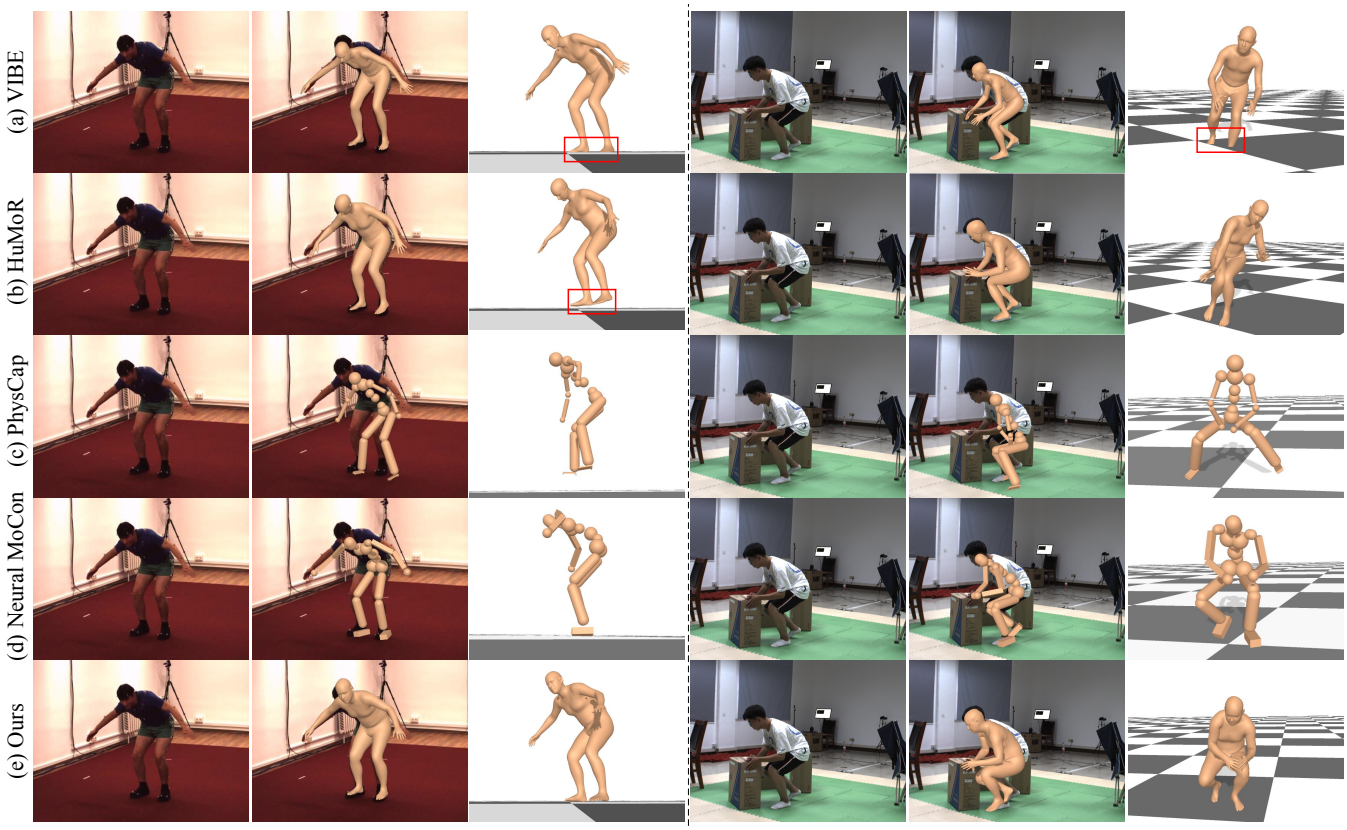


Figure 3: Qualitative comparison with other methods. VIBE and HuMoR are kinematics-based methods that utilize temporal information to predict 3D human motion, which results in severe artifacts. PhysCap and Neural MoCon incorporate physics in their framework. However, PhysCap does not consider the scene interactions and 2D observations in the physics-based tracking, and Neural MoCon cannot reconstruct accurate body contact. 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.

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

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

We conduct several ablation experiments in this section to verify the effectiveness of important components.

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

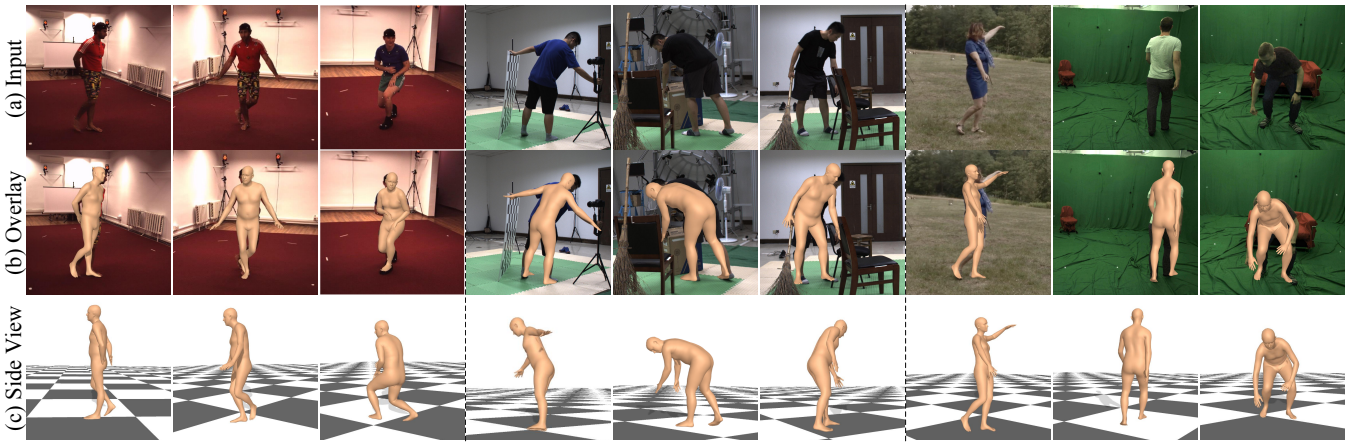


Figure 4: Qualitative results on Human3.6M, 3DOH, and 3DHP dataset. Our method can produce physically plausible human motion with accurate contact in different scenes.

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

Table 4: Ablation studies. We study the physical guidance with different denoising steps. The physical guidance can significantly improve joint accuracy and success rate. 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 motion 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.

Method	PA-MPJPE↓	e_s ↓	success↑
phys (s=3) w/o guidance	44.7	5.9	73.9%
phys (s=3) w/ guidance	41.3	3.5	90.3%

Table 5: Ablation study on the physical guidance. w/o guidance means that we remove the projection loss gradients in the condition. The experiment describes a comparison between the strategy in PhysDiff and our method. With the gradients, the 2D observations can also be considered in the denoising, and thus promote a more accurate motion capture.

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.

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

The authors would like to thank Yuan Yang for the helpful discussion about physics-based imitation.

References

[Arnab *et al.*, 2019] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019.

[Choi *et al.*, 2022] Jeongjun Choi, Dongseok Shim, and H Jin Kim. Diffupose: Monocular 3d human pose estimation via denoising diffusion probabilistic model. *arXiv preprint arXiv:2212.02796*, 2022.

[Fang *et al.*, 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

[Gärtner *et al.*, 2022a] Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *CVPR*, 2022.

[Gärtner *et al.*, 2022b] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *CVPR*, 2022.

[Gong *et al.*, 2022a] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. *arXiv preprint arXiv:2211.16940*, 2022.

[Gong *et al.*, 2022b] Kehong Gong, Bingbing Li, Jianfeng Zhang, Tao Wang, Jing Huang, Michael Bi Mi, Jiashi Feng, and Xinchao Wang. Posetriplet: Co-evolving 3d human pose estimation, imitation, and hallucination under self-supervision. In *CVPR*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 2020.

[Holmquist and Wandt, 2022] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. *arXiv preprint arXiv:2211.16487*, 2022.

[Huang *et al.*, 2021] Buzhen Huang, Yuan Shu, Tianshu Zhang, and Yangang Wang. Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In *3DV*, 2021.

[Huang *et al.*, 2022a] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *CVPR*, 2022.

[Huang *et al.*, 2022b] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *TPAMI*, 2022.

[Huang *et al.*, 2022c] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2uv: Single-shot multiperson mesh recovery with deep uv prior. *TIP*, 2022.

[Ionescu *et al.*, 2013] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013.

[Jahangiri and Yuille, 2017] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *ICCV Workshops*, 2017.

[Kanazawa *et al.*, 2019] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kocabas *et al.*, 2020] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[Li and Lee, 2019] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, 2019.

[Li *et al.*, 2021] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021.

[Li *et al.*, 2022a] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D &d: Learning human dynamics from dynamic camera. In *ECCV*, 2022.

[Li *et al.*, 2022b] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022.

[Li *et al.*, 2022c] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022.

[Li *et al.*, 2022d] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of human–object interactions from internet videos. *IJCV*, 2022.

[Liu *et al.*, 2015] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. In *Computer Graphics Forum*, 2015.

[Loper *et al.*, 2015] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 2015.

[Luo *et al.*, 2020] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020.

[Luo *et al.*, 2021] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *NIPS*, 2021.

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

[Mehta *et al.*, 2017] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *TOG*, 2017.

[Peng *et al.*, 2018] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *TOG*, 2018.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[Rempe *et al.*, 2021] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021.

[Sharma *et al.*, 2019] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *ICCV*, 2019.

[Shimada *et al.*, 2020] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *TOG*, 2020.

[Shimada *et al.*, 2021] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *TOG*, 2021.

[Song *et al.*, 2020a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Song *et al.*, 2020b] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020.

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

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012.

[Vondrak *et al.*, 2012] Marek Vondrak, Leonid Sigal, Jessica Hodgins, and Odest Jenkins. Video-based 3d motion capture through biped control. *TOG*, 2012.

[Wehrbein *et al.*, 2021] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *ICCV*, 2021.

[Wei and Chai, 2010] Xiaolin Wei and Jinxiang Chai. Videomocap: Modeling physically realistic human motion from monocular video sequences. In *SIGGRAPH*. 2010.

[Xiang *et al.*, 2019] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019.

[Xie *et al.*, 2021] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *ICCV*, 2021.

[Yi *et al.*, 2022] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *CVPR*, 2022.

[Yu *et al.*, 2021] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *TOG*, 2021.

[Yuan and Kitani, 2019] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *ICCV*, 2019.

[Yuan and Kitani, 2020] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. *NIPS*, 2020.

[Yuan *et al.*, 2021] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *CVPR*, 2021.

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

[Zell *et al.*, 2017] Petrissa Zell, Bastian Wandt, and Bodo Rosenhahn. Joint 3d human motion capture and physical analysis from monocular videos. In *CVPR Workshops*, 2017.

[Zhang *et al.*, 2020] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020.

[Zhang *et al.*, 2022] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[Zhou *et al.*, 2019] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.