Learning Human Dynamics in Autonomous Driving Scenarios

Jingbo Wang[1,2] Ye Yuan[1] Zhengyi Luo[1,3] Kevin Xie[1,4] Dahua Lin[2] Umar Iqbal[1]
Sanja Fidler[1,4,5] Sameh Khamis[1]
[1]NVIDIA [2]The Chinese University of Hong Kong [3]Carnegie Mellon University
[4]University of Toronto [5]Vector Institute

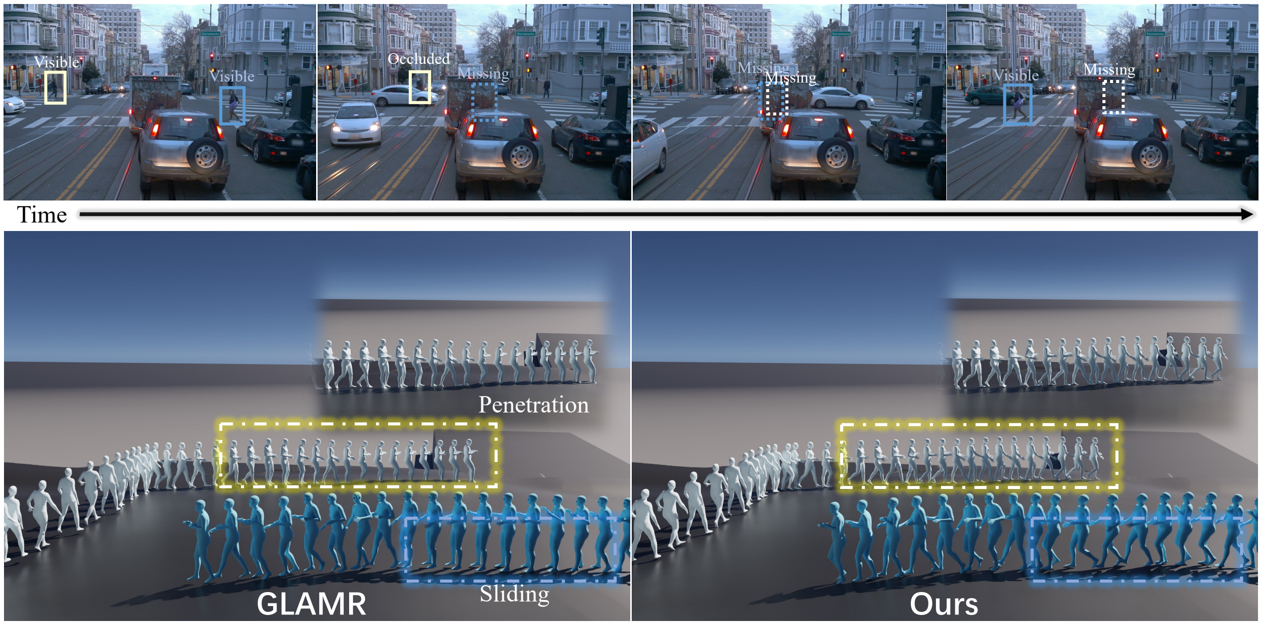

Figure 1: We compare our method against GLAMR [67], the state-of-the-art method for global human motion mesh recovery. The output of GLAMR (**left**) suffers from various physical implausibilities, such as floating, sliding, or terrain penetration. Our method (**right**) yields a clear improvement.

Abstract

Simulation has emerged as an indispensable tool for scaling and accelerating the development of self-driving systems. A critical aspect of this is simulating realistic and diverse human behavior and intent. In this work, we propose a holistic framework for learning physically plausible human dynamics from real driving scenarios, narrowing the gap between real and simulated human behavior in safety-critical applications. We show that state-of-the-art methods underperform in driving scenarios where video data is recorded from moving vehicles, and humans are frequently partially or fully occluded. Furthermore, existing methods often disregard the global scene where humans are situated, resulting in various motion artifacts like foot sliding, floating, or ground penetration. To address this challenge, we propose an approach that incorporates physics with a reinforcement learning-based motion controller to learn human dynamics for driving scenarios. Our framework can simulate physically plausible human dynamics that accurately match observed human motions and infill motions for occluded body parts, while improving the physical plausibility of the entire motion sequence. Experiments on the challenging Waymo Open Dataset show that our method outperforms state-of-the-art motion capture approaches significantly in recovering high-quality, physically plausible, and scene-aware human dynamics.

1. Introduction

Self-driving systems have come a long way in recent years. One major advancement that directly impacted these systems is the widespread adoption of simulation. While considerable attention was given to simulating traffic and other vehicles (*e.g.*, TrafficSim [59], GeoSim [3], STRIVE [51]), simulating pedestrian motion and behavior received less attention in the literature. In this work, we argue that understanding human behavior and intent is a critical step towards realistic simulation, which in turn is key to the safety of autonomous vehicles (AVs) in real-world settings. The first step towards this goal is capturing physically plausible human dynamics of entire motion sequences in driving scenarios.

Contrary to human motion in indoor scenes [12], motion captured in AV scenarios presents several challenges. These challenges include camera motion, long-term partial or full occlusions, scale variability, and complex interactions with the real-world environment. State-of-the-art (SOTA) approaches [69, 68, 70, 46, 34, 66, 35] for learning human dynamics in indoor scenarios tend to focus on enhancing physical plausibility for visible frames, neglecting several of these obstacles, such as long-term motion occlusions and the interplay between humans and terrains. As a result, it is challenging to directly apply these approaches to complex AV scenarios. Although recent works [67, 49] have demonstrated progress in infilling missing motions, there are significant limitations to using these methods for learning physically plausible human dynamics in AV scenarios, as shown in Figure 1. Firstly, these methods consider only physiologically inspired kinematics constraints such as joint limits, and do not model the physical plausibility of the pose in relation to the environment, such as when an object is floating. Secondly, these motion generation models are usually trained on indoor datasets, and the domain gap between indoor and driving scenarios renders them incapable of infilling missing motions in a plausible way.

In this paper, we present a novel holistic framework to learn human dynamics in such challenging AV scenarios. Our proposed framework distinguishes itself from prior works by its ability to generate physically plausibility motions for long-term partially or fully missing body parts, thus addressing an important limitation in the SOTA approaches to learn human dynamics. Firstly, we integrate off-the-shelf motion capturing (*e.g*., KAMA [13]) and scene reconstruction (*e.g*., Possion Surface Reconstruction [17]) methods to recover observed motions of visible humans and recover the terrain mesh as well. Before learning human dynamics from the captured motions, we fix the missing terrains by the observed motion trajectories, as well as filter out frames with low-confidence estimation caused by partial occlusion, to guarantee the simulation framework is based on reasonable visual observation. Next, we track the captured motion on the reconstructed terrains by generating infilling motions for the missing frames while enforcing the physical plausibility of the captured motions (*e.g*. penetration free against the ground). In contrast to GLAMR [67], which infills a few missing frames together using the pretrained transformer-based model, our method generated the motions in a stepwise fashion using a local motion controller, similar to [28, 62]. This controller-style motion generation reduces foot sliding over long-term occlusions. We demonstrate the adaptability of our approach by showing that, even though it is trained on indoor motion data, it can generalize to in-the-wild driving scenarios by placing physical constraints on the human-scene interactions. Specifically, we first train the conditional variational autoencoder [18] (cVAE) as the local motion generator whose latent space is the action space of movement. We then train a high-level controller to sample this latent space to perform infilling. Although all the motion generation models and physics-aware imitators are trained on an indoor dataset with flat ground, our method can easily adapt these models to uneven terrains in driving scenarios. Finally, we use an additional joint optimization, based on the physics-based imitator and generated motion, to match the video evidence (*e.g*., 2D keypoints with high confidence) not utilized in the previous stage. In summary, our framework is capable of learning physically plausible human dynamics for entire motion sequences in driving scenarios through visual observations.

We summarize our contributions as follows: 1) We propose the first framework for learning physically plausible human dynamics in driving scenarios, which is capable of generating physically plausible motions for partially or fully missing body parts. 2) We adapt a reinforcement learning-based motion generation framework trained only on indoor motion data, but can generalize to in-the-wild driving scenarios, infilling physically-plausible motion for occluded frames. 3) We achieve a significant improvement in motion quality over our motion capture framework to learn human dynamics, especially on partially or fully occluded frames.

2. Related Works

Kinematics-based Motion Capture: In recent times, the research community has focused on advancing the task of estimating 3D human keypoints directly to capture human motion, leading to impressive outcomes [6, 38, 40, 61, 44, 39, 1, 23]. Alternatively, several works have proposed adopting parametric human models, such as SMPL and SMPL-X [32, 42], as templates to capture human physiological motions [43, 15, 60, 2, 11, 58]. While some methods encourage the estimated motion to match observations during training [15, 53, 8] or fit the SMPL body through post-optimization [14, 22, 57], others predict SMPL parameters based on accurate keypoints [5] and apply inverse kinematics to transform the keypoint skeleton into the parametric body space [13, 26] to enable better keypoint localization.

Recent works have also recognized the importance of exploiting temporal information to improve motion consistency [4, 19, 16, 33]. To this end, some methods, such as VIBE [20] and humor [49], employ a variational auto-

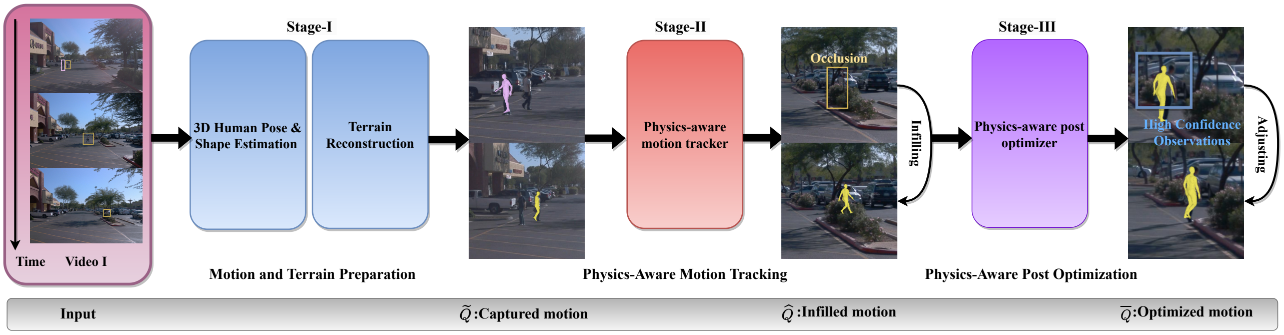
Stage-I
Stage-II
Stage-III
Time
Video I
3D Human Pose & Shape Estimation
Terrain Reconstruction
Physics-aware motion tracker
Occlusion
Infilling
Physics-aware post optimizer
High Confidence Observations
Adjusting
Motion and Terrain Preparation
Physics-Aware Motion Tracking
Physics-Aware Post Optimization
Input
$\widetilde{Q}$:Captured motion
$\widehat{Q}$:Infilled motion
$\overline{Q}$:Optimized motion

Figure 2: System Overview. Our approach processes each pedestrian mesh sequence in a stage-wise fashion. We first estimate motions for visible frames $\widetilde{Q}$ using an off-the-shelf motion capture method. We also reconstruct the ground terrain G in preparation for the physics-based stages (Details in Section 3.1). The physics-aware motion tracking (Section 3.2) infills the motion $\widehat{Q}$ for the occluded frames, as well as adapts the previously reconstructed motion to the reconstructed ground. In the last stage (Section 3.3), we optimize the entire motion $\widehat{Q}$ to closely match the evidence from a 2D keypoint-based system to produce the final motion $\overline{Q}$.

encoder (VAE) to learn motion priors on large-scale motion datasets like AMASS [36] to enhance the robustness of the system. Similarly, GLAMR [67] introduces a transformer-based cVAE model that leverages a data-driven approach to smooth out captured motions and infill occluded frames. Nevertheless, the lack of consideration for the dynamics attributes of human motions, particularly in complex AV scenarios, may induce undesirable physical artifacts, such as foot sliding and ground penetrations.

Dynamics-based Motion Capture: To ameliorate the physical artifacts associated with capturing human motion sequences, recent works have sought to leverage the physics attributes of human dynamics. These methods can be broadly classified into two categories. The first category entails post-optimization during test time [50, 56, 63, 7] on both trajectories and body poses by leveraging physics forces and human motion dynamics equations to optimize physical metrics or characteristics. The second category involves reinforcement learning (RL) and motion imitation [69, 68, 70, 46, 34, 66, 35], with the focus on capturing human motion first, followed by utilizing RL and carefully-designed policies to imitate the captured motions in a simulator environment, enforcing physical plausibility. Unfortunately, training RL-based models on each motion sequence are time-consuming, necessitating the development of regression-based approaches to directly estimate physical attributes and then update the captured motions to reduce the computational time. Several recent works [55, 71, 25] have taken this approach. However, many of these works assume that the character is walking on flat ground or without long-term occlusions and missing frames. Differently, this work aims to tackle these challenges of learning human dynamics in complex AV scenarios.

Occlusion-aware Motion Capture: Most of the existing works for human motion capturing assume the target human is fully visible in the image, and thus limit the robustness for the strong and long-term occluded human motions. Several recent works [49, 72, 9, 52, 21] try to address this problem but still can not handle the long-term occlusions, especially when the person of interest is obstructed completely. GLAMR [67] is the first work to solve this problem by infilling the missing motions using a transformer-based cVAE model. While this is a promising direction with many successful applications, the domain gap between their training dataset and the captured outdoor motions is large enough to limit its applicability in a complex outdoor environment.

Motion Control: Our framework relies on a key component known as the motion controller, a long-standing research topic in computer graphics and robotics [65, 24, 41]. Recent character control works follow a reinforcement learning-based pipeline [31, 29, 47, 45, 30, 69] to generate physical motions using either reference motion or motion prediction [10, 28, 62]. Similar to [62], our work employs a motion controller, in conjunction with a physics module, to generate physically plausible motions using a cVAE model. However, we diverge from past research by using the motion controller to recover human motion sequences in a complex outdoor environment, from being solely trained on indoor datasets. Our approach is the first to use motion controllers for this task of motion recovery, generating a more physically plausible representation of human behavior in complex driving environments.

3. Method

Our framework takes as input a monocular video sequence $I = (I_1, ..., I_M)$ with M frames captured by a fast-moving vehicle camera. Our goal is to obtain physically plausible human dynamics $\{Q^i\}_{i=1}^N$ for both visible and occluded frames of N entire motion sequences in the world coordinate system. Each person's motion Q is defined as (T, R, Θ) comprising the root translation $T = (\tau_s, ..., \tau_e)$, root rotation $R = (r_s, ..., r_e)$, and body motions $\Theta = (\theta_s, ..., \theta_e)$ from the first frame s to the last frame e. We

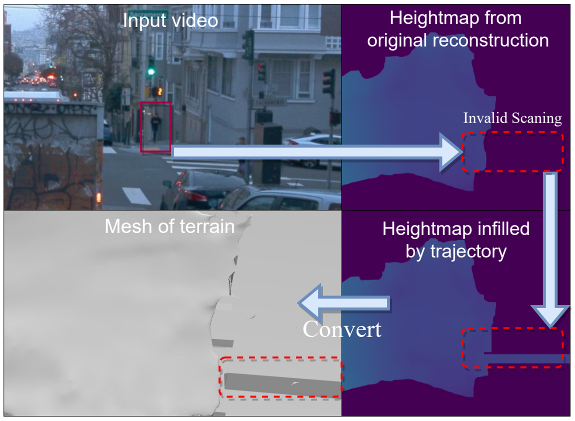
Input video
Heightmap from original reconstruction
Invalid Scaning
Mesh of terrain
Heightmap infilled by trajectory
Convert

Figure 3: Terrain Preparation. As shown in this figure, we first convert the reconstructed mesh to height map and extend this height map without valid LIDAR scanning for motions. At last, we convert the processed height map to mesh for the following physics-aware stages.

employ the SMPL [32] model's definition for the root translation $\tau_t \in R^3$, root orientation $r_t \in R^3$, and body pose $\theta_t \in R^{23\times3}$. To simulate the human dynamics captured from the real world, *e.g*. avoiding human-scene collisions, we also follow the definition of SMPL to estimate the body shape parameter $b_i \in R^{10}$ for each motion sequence i for meshes of characters in the simulator.

Our holistic framework (Figure 2) comprises three main stages. In **Stage-I** (Section 3.1), we prepare the initial observation, such as motion $\widetilde{Q}$ in visible frames and reconstructed terrain of the ground G, for the following physics-aware steps. In **Stage-II** (Section 3.2), we employ a novel physics-aware tracking framework to address occlusion issues with $\widetilde{Q}$ and adapt visible motions to reconstructed terrains with plausibly physical attributes, resulting in the occlusion-free motion $\widehat{Q}$. In **Stage-III** (Section 3.3), we apply our physics-based optimization to the generated motion $\widehat{Q}$ to ensure consistency between the generated motions and observations. More details as in the following sections.

3.1. Stage I: Motion and Terrain Preparation

In our physics-aware framework, we initiate physics-based simulation by computing initial kinematic human motions and reconstructing terrains. However, captured motions often have occlusions by cars and pedestrians in AV scenarios, leading to low-quality poses unsuitable for physics-based reasoning. Thus, we filter out occluded frames using 2D pose confidence scores and leverage physics-based priors to infill occluded frames later.

We also emphasize terrain reconstruction, an essential component of our framework. We use Poisson Surface Reconstruction [17] to reconstruct the terrain mesh from point clouds obtained from LIDAR on vehicles. However, point cloud densities are subject to occlusions and camera motion, leading to holes and uneven surfaces on the mesh. To obtain a well-formed mesh for simulation, we use a two-step approach of converting the reconstructed mesh to a height map and then infilling and expanding the height map to cover the entire range of human motions captured, as shown in Figure 3. By doing so, we significantly enhance the mesh quality, enabling accurate physics-based simulations of human agents interacting with the environment. We perform these two steps on all motions and terrains utilized in our experiments to ensure reliable and accurate simulations.

3.2. Stage II: Physics-Aware Motion Tracking

In this section, we will simulate physics-plausible human motion based on the processed observations from the first stage. In this section, we simulate physically plausible human motion based on the observations processed in the first stage. Our simulation aims to track the entire motion sequence on the reconstructed terrains, focusing on two aspects: 1) infilling physics-plausible human motions $\{\widehat{Q}_{t_1+1}, ...\widehat{Q}_{t_2-1}\}$ for occluded/missing frames between two visible frames $(\widetilde{Q}_{t_1}, \widetilde{Q}_{t_2})$; 2) ensuring that the infilled entire human motion sequence $\widehat{Q}$ can walk on the terrain with correct foot contact.

While previous state-of-the-art [67] attempts to infill missing frames using a transformer-based cVAE, as shown in Figure 1, low-quality motions persist due to the domain gap between indoor motion data and outdoor pedestrian data in busy city scenes. To address this issue, we design a hierarchical control framework that is capable of generating motions in different scenarios. This framework mainly consists of a high-level latent-space controller and a lower-level generative motion transition model. Specifically, our framework begins with a low-level generative motion transition model that can provide plausible human motion based on sparse input. Inspired by [28, 62], we develop a controllable latent space that will serve as the action space of our high-level motion controller. Our high-level motion controller can then sample latent codes based on terrain, input video observations, and past states. In contrast to [67], our framework can generate plausible motions for occluded frames by fine-tuning our controller on each video. We then apply a physics-aware humanoid controller that imitates the generated motion in a physics simulator on the reconstructed terrain, to ensure the accurate physics attributes of the entire generated motion.

3.2.1 Framework

Overview: Within our physics-aware motion tracking framework, we identify three critical components: the motion model π_M, the physics-aware motion imitator π_D, and the high-level controller π_C, as depicted in Figure 4. The motion model π_M employs a generative transition model to

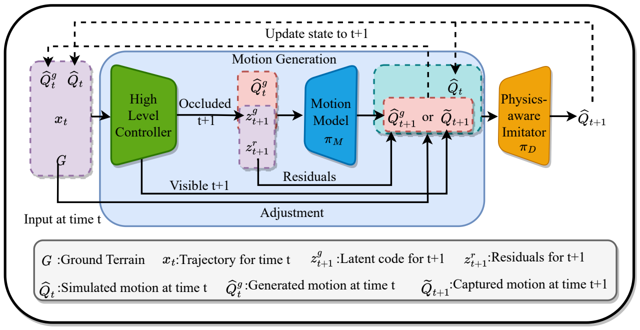
Update state to t+1
Motion Generation
$\widehat{Q}_t^g$ $\widehat{Q}_t$
x_t
G
High Level Controller
Occluded t+1
$\widehat{Q}_t^g$
z_{t+1}^g
z_{t+1}^r
Motion Model π_M
$\widehat{Q}_t$
$\widehat{Q}_{t+1}^g$ or $\widetilde{Q}_{t+1}$
Physics-aware Imitator π_D
$\widehat{Q}_{t+1}$
Residuals
Visible t+1
Adjustment
Input at time t
G :Ground Terrain
x_t:Trajectory for time t
z_{t+1}^g :Latent code for t+1
z_{t+1}^r:Residuals for t+1
$\widehat{Q}_t$:Simulated motion at time t
$\widehat{Q}_t^g$:Generated motion at time t
$\widetilde{Q}_{t+1}$:Captured motion at time t+1

Figure 4: Motion Tracking Framework. At time step t, our high-level controller predicts the latent code for the motion model π_M and the residuals for the occluded motion at $t+1$. Next, the physics-aware imitator π_D updates this prediction to ensure the physical plausibility of the generated motion. If the initial motion is visible at $t+1$, our high-level controller will directly use this captured motion rather than predict the latent code and generate motions by π_M.

compute the next pose using the previous pose and a latent code. Meanwhile, the physics-aware motion imitator π_D takes either an occluded frame pose generated by π_M or a visible frame pose directly observed as input, thereby controlling a virtual character to mimic the motion in a physics simulator. For visible frames, we circumvent the use of π_M and provide the observed pose directly to π_D. When dealing with occluded frames, we call upon the high-level controller π_C to generate the latent code required by π_M in order to control the motion of the virtual character and infill coherent motions between visible frames. To accomplish this infilling, we train π_C to address the motion tracking task, which involves two objectives. The first goal requires driving the virtual character to reach the same position as $\widetilde{Q}_{t_2}$ at time t_2, starting from $\widetilde{Q}_{t_1}$, while ensuring the character attains a similar pose to $\widetilde{Q}_{t_2}$. The second objective entails maintaining compatibility with the underlying terrain G.

Motion Transition Model: We follow one of the state-of-the-art local motion transition models [49] as π_M and use a pre-trained model from the official implementation. To infill occluded motions between two visible frames from $\widetilde{Q}_{t_1}$ to $\widetilde{Q}_{t_2}$, the initial observation of model is $\widetilde{S}_{t_1} = (\widetilde{\tau}_{t_1}, \widetilde{r}_{t_1}, \widetilde{\theta}_{t_1}, \widetilde{j}_{t_1}, \widetilde{\dot{\tau}}_{t_1}, \widetilde{\dot{r}}_{t_1}, \widetilde{\dot{j}}_{t_1})$, corresponding to the root translation, root orientation, body pose, joint position, the velocity of translation, velocity of rotation, and the velocity of joints respectively. The motion $\widehat{S}^g_{t_1+1} = (\widehat{\tau}^g_{t_1+1}, \widehat{r}^g_{t_1+1}, \widehat{\theta}^g_{t_1+1}, \widehat{j}^g_{t_1+1}, \widehat{\dot{\tau}}^g_{t_1+1}, \widehat{\dot{r}}^g_{t_1+1}, \widehat{\dot{j}}^g_{t_1+1})$ for the occluded frame t_1+1 can be generated by sampling latent code z_{t_1}. With this model, we can generate $t_2 - t_1 - 1$ step motions step by step for motion infilling.

Physics-Based Motion Imitator: Our physics-based motion imitator π_D adapts the output motions of the generation module to the reconstructed terrains G. Following [45, 69, 70], this model drives a simulated humanoid to imitate the target pose, producing a physically valid motion sequence through the simulation process. As shown in Figure 4, the input of this imitator is the motion $\widehat{S}^g_{t+1}$ of the motion generation model, as well as the motion simulated in the previous time step $\widehat{S}_t$. In practice, we adjust the height of $\widehat{S}^g_{t+1}$ to remove floating and penetration to the reconstructed terrain G. Our π_D predicts the target joint angle as a_{t+1} for the physics simulator. Similar to [69, 70], we use the proportional derivative controllers (PD) on each non-root joint to produce joint torques $\widehat{\mathscr{T}}_{t+1}$ and obtain the physics-plausible motion $\widehat{S}_{t+1} = (\widehat{\tau}_{t+1}, \widehat{r}_{t+1}, \widehat{\theta}_{t+1}, \widehat{j}_{t+1}, \widehat{\dot{\tau}}_{t+1}, \widehat{\dot{r}}_{t+1}, \widehat{\dot{j}}_{t+1})$ by the simulator upon the reconstructed terrain G.

High-level Controller: The goal of our high-level controller is to compute latent codes that can drive π_M from time step t to $t+1$. For the visible frame at $t+1$, the motion $\widetilde{G}_{t+1}$ is adapted to the reconstructed terrain G by first adjusting the height, and the high-level controller directly uses the adjusted captured motion as the imitation target for π_D, rather than predicting latent codes for π_M. For the occluded frame, our high-level controller samples specific latent code z_{t+1} for π_M. Basically, to ensure reaching the same position of $\widetilde{Q}_{t_2}$ at t_2 from $\widetilde{Q}_{t_1}$, we formulate this reaching problem as a trajectory following task, which can guide the character to reach $\widetilde{Q}_{t_2}$ at t_2 step-by-step and drive the character to a pose similar to $\widetilde{Q}_{t_2}$. We directly interpolate the root translation of $(\widetilde{\tau}_{t_1}, \widetilde{\tau}_{t_2})$ as the following trajectory $\{x_t\}$ for these missing frames. For each time step $t \in (t_1, t_2)$, the input of the high-level controller consists of the generated motion state $\widehat{S}^g_t$, the simulated motion $\widehat{S}_t$, the future trajectories x, the reconstructed terrain G, and the target motion $\widetilde{Q}_{t_2}$. In addition to the latent code z_{t+1}, our high-level controller additionally predicts z^r_{t+1} as residuals of the motion generated at $t+1$, since the motion generation model π_M is only trained on indoor motions on flat ground and may have trouble producing motions on uneven terrains. In practice, z^r_{t+1} consists of root translation, root orientation, and character body pose. Thus, for the missing frame, the motion at $t+1$ can be generated as follows:

$$z^g_{t+1}, z^r_{t+1} = \pi_C(\widehat{S}^g_t, \widehat{S}_t, x, G, \widetilde{Q}_{t_2}) \tag{1}$$

$$\widehat{S}^g_{t+1} = \pi_M(\widehat{S}^g_t, z^g_{t+1}) + z^r_{t+1}, \widehat{a}_{t+1} = \pi_D(\widehat{S}_t, \widehat{S}^g_{t+1}) \tag{2}$$

$$\widehat{S}_{t+1} = \mathrm{Sim}(\widehat{S}_t, \widehat{\mathscr{T}}_{t+1}) = \mathrm{Sim}(\widehat{S}_t, \mathrm{PD}(\widehat{S}_t, \widehat{a}_{t+1})) \tag{3}$$

3.2.2 Training Strategy

We follow [28] to train this high-level controller using the standard reinforcement learning algorithm [54]. Notice that

π_M is pre-trained and frozen during this process. The reward for training is as follows:

$$r = w_p \cdot r_p + w_i \cdot r_i. \tag{4}$$

The reward for the trajectory following task is as

$$r_p = \exp(-\alpha_p(\|\widehat{r}_p^{xy} - x_p^{xy}\|)), \tag{5}$$

where $\widehat{r}_p^{xy}$ and x_p^{xy} are the xy coordinates of the translation of physics state $\widehat{S}_t$ and the interpolated trajectory at time step t. The infilling reward encourages the motion model to generate a similar motion as $\widetilde{S}_{t_2}$ when the character is near to t_2:

$$r_i = \gamma_i \cdot \exp(-\alpha_i(\|\widehat{S}_t - \widetilde{S}_{t_2}\|)). \tag{6}$$

The weights (w_p, w_i) and (α_t, α_i) can be adjusted to fit different scenarios. The γ_i is equal to 1 if $t \in (t_2 - 15, t_2)$. Otherwise, we set this γ_i as 0 and thus the task is only trajectory following.

Training such a high-level controller from scratch for different scenarios is time-consuming (more than 12 hours on a single V100 GPU by IsaacGym [37] for 1000 iterations). To mitigate the time cost and obtain a more robust controller for different environments (*e.g.*, trajectories and terrains), we propose a pre-training and fine-tuning strategy for our high-level controller. We generate diverse uneven terrains and driving trajectories meant to simulate a range of scenarios in the simulator, followed by training the high-level controller on these synthetic environments for our designated tasks. For infilling ending motion, we sample from various motions in the AMASS [36] while performing motion-matching tasks. Leveraging the pre-trained high-level controller on synthetic data, the convergence of fine-tuning on real data is significantly faster, as illustrated in Figure 6. We present additional information about our pre-training strategy in our supplementary materials.

3.3. Stage III: Physics-aware Motion Optimization

After we obtain the physics plausible and occlusion-free human motions $\widehat{Q}$ in Section 3.2, the motion may not align with the image evidence such as 2D keypoints perfectly. Additionally, in Section 3.1, we have filtered out the whole body for some partially occluded motions, even if they have several high-confidence estimated keypoints. For these frames, our physics-aware motion tracking in Section 3.2 always obtains motions by motion generation and thus causes misalignment to the video observation.

To close these gaps, in this section, we propose a physics-aware motion optimization method to further optimize $\widehat{Q}$. This stage mainly consists of two components. The first is a new physics-aware imitator π_K similar as π_D in Section 3.2, to maintain physics plausibility while matching video evidence. Additionally, we introduce residual parameters $\{\delta R_i\}_{i=1}^{T}$ for the target motion $\widehat{Q}$ of the imitator at each time step. During optimization, the imitation target $\widehat{Q}_{t+1}$ is adjusted by δR_{t+1} and thus encourages the imitator to predict a consistent motion $\overline{a}_{t+1}$ with video observation. In practice, we introduce these residual parameters to the root orientation and body pose of characters in the simulator. After training, the updated target motion is defined as:

$$\widehat{r}_{t+1}^{u} = \widehat{r}_{t+1} + \delta R_{t+1}^{r}, \widehat{\theta}_{t+1}^{u} = \widehat{\theta}_{t+1} + \delta R_{t+1}^{\theta}, \tag{7}$$

$$\widehat{S}_{t_1+1}^{u} = (\widehat{\tau}_{t+1}, \widehat{r}_{t+1}^{u}, \widehat{\theta}_{t+1}^{u}, \widehat{j}_{t+1}, \widehat{\dot{\tau}}_{t+1}, \widehat{\dot{r}}_{t+1}, \widehat{\dot{j}}_{t+1}). \tag{8}$$

Based on this adapted target motion, we first use the imitator π_K to predict the target joint angle as $\overline{a}_{t+1}$ for the physics simulator and produce joint torques $\overline{\mathscr{T}}_{t+1}$ to obtain the final result $\overline{Q}_{t+1}$, similar as π_D.

During optimizing, we fine-tune this pre-trained imitator and these residual parameters for each motion sequence. To encourage consistency between observation and output motion, we use the following reward function:

$$r_{proj} = \exp(-\alpha_p \sum(\|\Pi(\overline{j}_t) - \widetilde{j}_t^{2D}\| \times \widetilde{c}_t)), \tag{9}$$

where Π is the projection function from world to image space, $\widetilde{j}_t^{2D}$ is the estimated 2D pose at time step t, as well as $\widetilde{c}_t$ is the corresponding confidence score. Besides, we still use the similar reward function r_{im} as [69, 70] used for motion imitation to encourage correct physics attributes for the motions with residuals. The final reward is as follows:

$$r = r_{proj} \cdot w_p + r_{im} \cdot w_{im}. \tag{10}$$

where (w_p, w_{im}) and (α_p, α_{im}) can be adjusted to fit different scenarios.

4. Experiments

Dataset: We conduct our experiments on the Waymo Open Dataset [73], currently the largest dataset recorded for autonomous driving scenarios featuring point clouds representing the surrounding terrain and sparsely annotated 3D keypoints. To evaluate the effectiveness of our proposed method, we curate a subset of 20 sequences from the training and validation sets, each containing various scenarios, and incorporating more than 30 different pedestrian 3D keypoint annotations, as well as high-quality terrain reconstructions. The Waymo Open Dataset captures data at a 10Hz frame rate using a camera installed on the vehicle. As we require a 30Hz frame rate in our physics simulator, we upsample the captured motions using linear interpolation.

Metric: We use both *kinematics-based* and *physics-based* metrics for evaluation. Firstly, to quantify the generated motions from our framework, we report the motion FID, a standard metric for evaluating motion fidelity [27, 48]. Besides, to demonstrate the reconstruction accuracy, we use

Table 1: Baseline Comparison. We compare against several different baselines on the following metrics. GLAMR* means use the same physics-aware imitator as our framework after GLAMR. Our method achieves the significantly better result on FID, PAMPJPE on frames with occlusion (Occ), and **physics-based** metrics (GP, FS, FL).

Method	FID ↓ (All)	PA-MPJPE ↓ (Occ)	PA-MPJPE ↓ (All)	GP ↓ (All)	FS ↓ (All)	FL ↓ (All)	Accel (All)
KAMA	4.62	97.52	75.06	82.20	49.27	78.89	-
GLAMR	4.17	91.74	74.34	81.42	46.33	78.23	154.14
GLAMR*	4.28	93.44	76.34	17.34	38.34	15.38	138.44
Ours	**1.96**	**86.02**	**74.22**	**12.62**	**7.44**	**13.25**	105.48

Table 2: Ablation studies on the physics-aware imitator. Although we have adapted the generated motion to the ground during training, the physics-aware motion imitator still can improve the motion quality on these physics attributes. Our* means the result without post-optimization.

Method	GP ↓	FS ↓	FL ↓	Accel
Without π_D	16.82	10.84	23.02	114.48
Ours*	**12.92**	**7.05**	**14.08**	108.59

the Procrustes-aligned mean per-joint position error (PA-MPJPE) and 2D localization error (2D-LE) for the generated human dynamics motion. For the physical attributes, we follow the metrics to measure jitter, foot sliding, ground penetration, and floating as in [70, 25]. The jitter is estimated by computing the acceleration (Accel). The foot sliding (FS) is estimated by finding the body mesh vertices that contact the ground in two adjacent frames and then computing their moving distance. The ground penetration (GP) is computed as the average distance between the ground and the mesh vertices below the ground. For floating (FL), we compute the distance between the ground and the nearest vertex of mesh to the ground. The unit for these metrics are millimeters (mm), except for Accel (mm/frame^2).

4.1. Implementation Details

In Section 3.2 and Section 3.3, we follow the methodology outlined in RFC [69] to train our physics-aware motion imitator. This imitator is trained on the flat ground using the AMASS [36] dataset and IsaacGym [37]. We model the SMPL agent in this simulator following [35]. In Section 3.2, we use the official motion generation model put forth by HuMOR [49] as our motion model π_M. Regarding our high-level controller, we pre-train this model on synthetic environments for 5000 iterations before fine-tuning it in real-world environments. For motion infilling, we optimize the result for 2000 iterations with our physics-aware motion optimization. Finally, to ensure the accuracy of our optimization, we leverage VITPose [64] to extract precise 2D keypoints in the last stage of our framework.

4.2. Evaluation

Table 3: Ablation studies on post optimization. We compare the infilling results with and without the post-optimization after physics-aware motion tracking.

Method	2D-LE ↓ (All)	PA-MPJPE ↓ (Occ)	PA-MPJPE ↓ (All)	GP ↓ (All)	FS ↓ (All)	FL ↓ (All)	Accel (All)
Without Opt	26.41	89.84	75.48	12.92	**7.05**	14.08	108.59
With Opt	**17.37**	**86.02**	**74.22**	**12.62**	7.44	**13.25**	105.48

Baselines: We conduct our method based on KAMA [13], a strong human motion captured method for visible motions. To evaluate the quality of the generated motion by our method, we compare it with the state-of-the-art method [67], which is designed to generate missing parts for the entire motion sequence. As [67], we also compare with methods using linear interpolation for the infilling of the missing frames. For fair comparisons, we adapt all the methods and experiments by using the ground-truth camera extrinsic and intrinsic parameters provided by the Waymo Open Dataset.

Quantitative Results: We show our results in Table 1 to show the quality of generated motions. To begin with, we evaluate the quality of the motion sequences generated by our method and compare them with our baselines using the FID metric. Our approach produces the best performance with substantial improvements in this regard. Furthermore, our method exhibits the best result in the PA-MPJPE metric, particularly a notable reduction in the frames with occlusion. Regarding physics-based metrics, our method continues to outperform all baselines. However, we still encounter minor foot penetration issues while quantifying the physics attributes due to errors in the transformation between the height map and mesh, as well as in the character's SMPL model within the simulator. Despite utilizing GLAMR and the physics imitator cooperatively to facilitate better physics attribute outcomes, the infilling capability of GLAMR still influences the performance of kinematics metrics, FL, and Accel. This points to the considerable complexity involved in designing our controller-based physics-aware framework to address this problem. In summary, our approach possesses the potential to achieve high-quality human dynamics learning based on visual observations.

Qualitative Results: We demonstrate the qualitative results in Figure 5. We compare against the GLAMR [67], which is the state-of-the-art method to infill captured motions. We find that our method achieves better results on the physical attributes of human motion. More comparison videos are in our **supplementary material**.

4.3. Ablation Studies

Pre-training for High-level Controller: Firstly, we compare the performance of our method without the pre-training step on synthetic terrains and trajectories. As shown in Figure 6, the high-level controller based on the pre-trained model converges faster significantly. More comparisons of this pre-training are in our supplementary materials.

Physics-based Motion Tracker: Then, we compare with the motion without our physics-aware imitator in Table 2. Although we adapt generated motion to the ground during

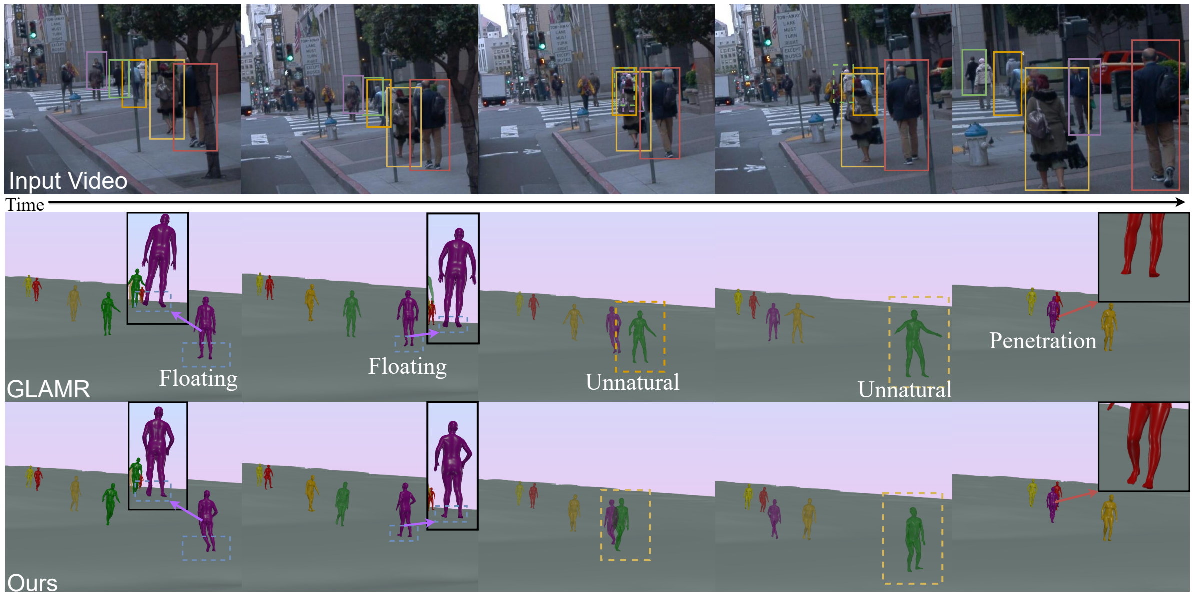

Figure 5: Qualitative Comparison. We compare our method with the state-of-the-art method [67] for global human motion mesh recovery in this figure. Under this comparison, our method mitigates the artifacts on physics attributes of the whole motion sequence significantly.

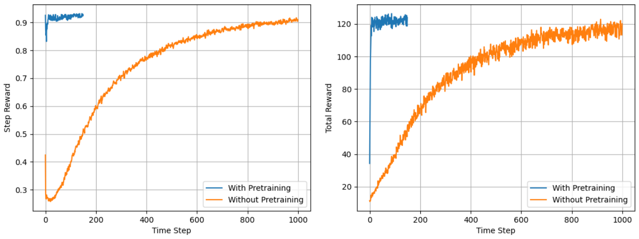

Figure 6: Convergence Results. We compare the time cost of training the high-level controller on the Waymo Open Dataset with and without the pre-training on synthetic data. We show that the high-level controller converges much faster and achieves a better reward with the pre-training.

training, the physics-aware imitator still improves the performance on physics attributes of generated motion. More qualitative comparisons of different design choices are in our supplementary materials.

Physics-based Post Optimization: Finally, we conduct a comparison of our method's performance without the post-optimization step outlined in Section 3.3. As indicated in Table 3, the inclusion of post-time optimization notably enhances the accuracy of matching 2D/3D observations and improves some physical attributes of the entire motion sequence. In addition, we demonstrate the effects of this stage by comparing the results with and without post-optimization in Figure 7.

5. Conclusion and Limitations

This paper presents a holistic framework for learning physically-plausible human dynamics motion of entire sequences in AV scenarios. Our approach stands out from prior work by generating motion sequences not only for visible frames but also for frames with occlusions or missing data. The generated motion sequences are constrained by physics and thus are suitable for downstream simulation tasks in AV scenarios. Our methodology begins with processing reconstructed terrains using LIDAR and recovering visible motions with off-the-shelf components. We then use a reinforcement learning-based motion controller within a physically-constrained environment to infill the motions. Finally, we propose a physics-aware post-optimization stage that utilizes keypoint observations to optimize the entire motion sequence. Our approach outperforms previous methods, particularly in terms of entire sequence motion quality and physical attributes, in several challenging AV scenarios. Although our results meet the necessary quality for simulation tasks, a limitation we aim to address in the future is the ability to optimize body shapes estimated by MoCap with our physics-aware framework.

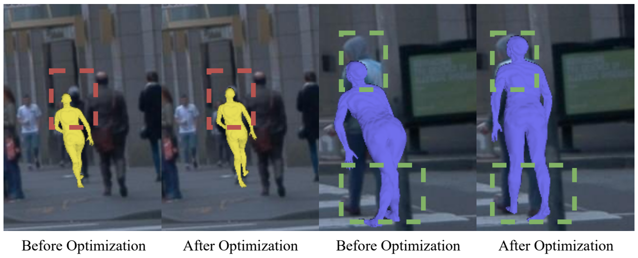
Before Optimization After Optimization Before Optimization After Optimization

Figure 7: Effect of physics-aware post-optimization. We demonstrate the effect of our physics-aware post-optimization in this figure. Following this optimization, generated motions from the previous stages end up better matched and aligned with the video evidence.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 2

[3] Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchen Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, and Raquel Urtasun. Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7230–7240, 2021. 1

[4] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021. 2

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2

[6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018. 2

[7] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *ICCV*, 2021. 3

[8] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. *arXiv preprint arXiv:2211.10705*, 2022. 2

[9] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020. 3

[10] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 3

[11] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017. 2

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[13] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *2021 International Conference on 3D Vision (3DV)*, 2021. 2, 7

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. 2020. 2

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 2

[16] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 2

[17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, 2006. 2, 4

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 2

[21] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 3

[22] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[23] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2

[24] Yoonsang Lee, Sungeun Kim, and Jehee Lee. Data-driven biped control. In *ACM SIGGRAPH 2010 papers*, pages 1–8. 2010. 3

[25] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *ECCV*, 2022. 3, 7

[26] 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. 2

[27] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv*, 2020. 6

[28] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4), 2020. 2, 3, 4, 5

[29] Libin Liu and Jessica Hodgins. Learning to schedule control fragments for physics-based characters using deep q-learning. *ACM Transactions on Graphics (TOG)*, 36(3):29, 2017. 3

[30] Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[31] Libin Liu, Michiel Van De Panne, and KangKang Yin. Guided learning of control graphs for physics-based characters. *ACM Transactions on Graphics (TOG)*, 35(3):29, 2016. 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 4

[33] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3d human motion estimation via motion compression and refinement. In *ACCV*, 2020. 2

[34] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *NIPS*, 2021. 2, 3

[35] Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 2, 3, 7

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3, 6, 7

[37] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. 6, 7

[38] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 2

[39] 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. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017. 2

[40] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *CVPR*, 2019. 2

[41] Uldarico Muico, Yongjoon Lee, Jovan Popović, and Zoran Popović. Contact-aware nonlinear control of dynamic characters. In *ACM SIGGRAPH 2009 papers*, pages 1–9. 2009. 3

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[43] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 2

[44] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2

[45] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3, 5

[46] Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. Mcp: Learning composable hierarchical control with multiplicative compositional policies. In *Advances in Neural Information Processing Systems*, pages 3681–3692, 2019. 2, 3

[47] Xue Bin Peng and Michiel van de Panne. Learning locomotion skills using deeprl: Does the choice of action space matter? In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–13, 2017. 3

[48] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, pages 10985–10995, 2021. 6

[49] 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. 2, 3, 5, 7

[50] Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[51] Davis Rempe, Jonah Philion, Leonidas J. Guibas, Sanja Fidler, and Or Litany. Generating useful accident-prone driving scenarios via a learned traffic prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[52] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *ECCV*, 2020. 3

[53] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCVW*, 2021. 2

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 5

[55] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), aug 2021. 3

[56] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics*, 39(6), dec 2020. 3

[57] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 2

[58] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the*

IEEE International Conference on Computer Vision, pages 5349–5358, 2019. 2

[59] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10400–10409, 2021. 1

[60] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. 2017. 2

[61] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, 2020. 2

[62] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 2, 3, 4

[63] 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. 3

[64] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 7

[65] KangKang Yin, Kevin Loken, and Michiel Van de Panne. Simbicon: Simple biped locomotion control. *ACM Transactions on Graphics (TOG)*, 26(3):105–es, 2007. 3

[66] Ri Yu, Hwangpil Park, and Jehee Lee. Human dynamics from monocular video with dynamic camera movements. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021. 2, 3

[67] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 1, 2, 3, 4, 7, 8

[68] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10082–10092, 2019. 2, 3

[69] Ye Yuan and Kris Kitani. Residual force control for agile human behavior imitation and extended motion synthesis. In *Advances in Neural Information Processing Systems*, 2020. 2, 3, 5, 6, 7

[70] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *CVPR*, 2021. 2, 3, 5, 6, 7

[71] Petrissa Zell, Bodo Rosenhahn, and Bastian Wandt. Weakly-supervised learning of human dynamics. In *ECCV*, 2020. 3

[72] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7376–7385, 2020. 3

[73] Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *CVPR*, 2022. 6