PACER+: On-Demand Pedestrian Animation Controller in Driving Scenarios

Jingbo Wang* Zhengyi Luo* Ye Yuan Yixuan Li Bo Dai

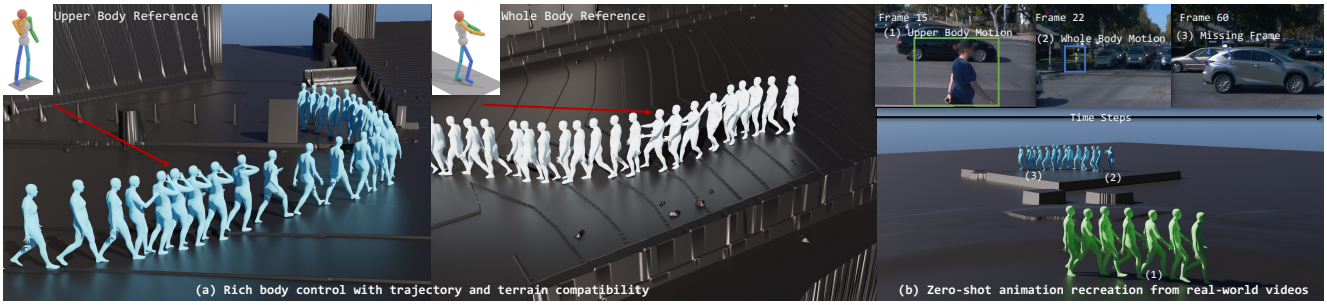

Figure 1. We showcase the effectiveness of our proposed framework in synthetic and real-world driving scenarios. Our framework excels at generating physically realistic animations that adhere to provided trajectories while offering extensive control over the upper and full body movements. Additionally, our framework demonstrates the remarkable ability to recreate pedestrian animations with occlusions from real-world videos in a *zero-shot* manner. These inherent capabilities make our framework a robust and versatile approach for **on-demand** pedestrian animation in driving scenarios.

Abstract

We address the challenge of content diversity and controllability in pedestrian simulation for driving scenarios. Recent pedestrian animation frameworks have a significant limitation wherein they primarily focus on either following trajectory [46] or the content of the reference video [57], consequently overlooking the potential diversity of human motion within such scenarios. This limitation restricts the ability to generate pedestrian behaviors that exhibit a wider range of variations and realistic motions and therefore restricts its usage to provide rich motion content for other components in the driving simulation system, e.g.*, suddenly changed motion to which the autonomous vehicle should respond. In our approach, we strive to surpass the limitation by showcasing diverse human motions obtained from various sources, such as generated human motions, in addition to following the given trajectory. The fundamental contribution of our framework lies in combining the motion tracking task with trajectory following, which enables the tracking of specific motion parts (*e.g.*, upper body) while simultaneously following the given trajectory by a single policy. This way, we significantly enhance both the diversity of simulated human motion within the given scenario and the controllability of the content, including language-based control. Our framework facilitates the generation of a wide range of human motions, contributing to greater realism and adaptability in pedestrian simulations for driving scenarios.*

1. Introduction

Autonomous vehicle (AV) simulation systems have gained increasing attention, given their potential to help develop safe and adaptable self-driving algorithms. One of its crucial functionalities is creating realistic and diverse pedestrian animations to train self-driving algorithms to react to a diverse array of human behaviors. It can be crucial for the safety of AV since subtle changes in pedestrians' moving directions or gestures could entail large changes in vehicle behaviors. However, despite the promising results [5, 45, 52] of the background scene and vehicle motion creation in the current simulation system, the performance of pedestrian animation still lags behind.

While state-of-the-art pedestrian animation methods often use keyframe animations authored by artists [34],they lack the proper reaction to the scene geometry due to the absence of the laws of physics. Recent physics simulation-based pedestrian animation method [46] can create pedestrian animations that are human-like, physically plausible, and conform to the geometry of the scene. Yet its animation is only controlled by 2D trajectories and is limited to basic locomotion such as walking and running, which makes it insufficient to reflect the natural diversity of pedestrian behaviors. Alternatively, pedestrian animation can also be obtained from video sequences via simulation-based motion

capture [57], which, however, adopts a per-video optimization strategy that is computationally intensive and thus cannot create *new* and *diverse* animation at a large scale or in an on-demand manner.

In this work, we propose PACER+, a simulation-based framework for generating diverse and natural pedestrian animation **on-demand**. Our framework offers richer *zero-shot* control beyond trajectory following and enables the creation of diverse animation in both manual and real-world scenarios, to meet the demand for more controllable generation. Specifically, PACER+ supports fine-grained control over different body parts while following the given trajectory, which is achieved by selectively tracking *specific body parts* instead of rigidly tracking the entire body [24, 27, 58]. This creates room for more life-like animation, such as walking while making a phone call, and simultaneously ensures smoothness of the motion, compatibility of the terrain, and adherence to the provided trajectory. Using our framework, a variety of pedestrian behaviors can be introduced into the simulation system from various sources, including motion generation models, pre-captured motions, and videos, as depicted in Figure 1. Moreover, for the demand of recreating real-world pedestrian animation into simulation environments, PACER+ can also demonstrate motion from the given video without re-training or fine-tuning, where the missing part will be infilled automatically, as shown in Figure 1.

The key insight behind PACER+ lies in the synergy between motion imitation and trajectory following tasks. While the lower-body motion is often influenced by the trajectory and terrain, the upper-body motion has the flexibility to encompass a diverse range of motions. Therefore, we establish a synergistic relationship between motion imitation and trajectory following tasks through a joint training scheme. In this scheme, a single policy is employed to track partial body motion and follow trajectories simultaneously in a physically plausible way. To achieve this, we introduce a per-joint spatial-temporal mask that indicates the presence of a reference motion for the policy to track. During training, we randomly select time steps and joints to insert as the reference motion into the trajectory following task. This encourages the policy to concurrently track the trajectory and imitate the reference motion, enabling generalizable trajectory and motion tracking.

Our contributions can be summarized as follows: (1) We propose a unified physics-based pedestrian animation framework, named PACER+, which can control a simulated pedestrian to follow the 2D trajectory and *specific body parts* reference motion at the same time *on-demand*. (2) Our framework supports the generation of diverse pedestrian behaviors from various sources, including generative models, pre-captured motions, and videos, in any given driving scenario, such as manually built or real scanned environments. (3) Notably, our framework achieves the *zero-shot* recreation of real-world pedestrian animations into simulation environments, where the missing part will be infilled automatically.

2. Related Works

Controllable Character Animation. Controllable character animation has been a longstanding research topic in computer graphics and robotics [17, 33, 65]. Previous research in controllable character animation has often focused on integrating high-level tasks, such as trajectory following or goal-reaching, with low-level control of body joints, involving joint positions or angles. By combining these two levels of control, researchers aimed to achieve controllable animation that adheres to specific tasks or objectives. Recent methods have explored primarily two main approaches: (1) kinematics-based [8, 23] methods and (2) physics-based method [40, 41, 59, 64]. These works primarily aim to achieve predefined tasks with plausible human motions.

More recently, researchers have begun to extend the range of motion content while still adhering to given tasks. For instance, PADL [12] and CALM [54] introduce language-based and example-based control to generate diverse motions for the given tasks. Some recent works [3, 61] introduce spatial composition to expand the range of skills for more complex tasks. Based on the success of ControlNet [72], AdaptNet [62] incorporates a similar design choice into its policy network to generate diverse human motions on complex terrains.

The key distinction between our work and these existing approaches lies in our focus on zero-shot fine-grained control for character animation, specifically for following given tasks. Once trained, our method does not require additional policy network training for new skills [61, 62]. Furthermore, our control framework enables flexible yet fine-grained control over the given character, including the location of upper body joints of specific examples, which has not been fully addressed in previous style-based controlling works [3, 12, 54]. Moreover, our approach supports motion content from various sources, such as videos, motion capture data, or even motions generated by other methods. This capability enhances the versatility and adaptability of our framework, allowing for on-demand pedestrian animation. Users can generate desired character behaviors by leveraging the flexibility of incorporating diverse motion sources.

Physics-based Humanoid Motion Tracking. Using a deep neural network [24, 26, 27, 38, 58, 68] to track kinematics human motions in physics simulation achieves promising results in recent years. To achieve a better success rate, previous works introduce residual force [68] and Mixture-of-Experts network structures [26, 27, 58]. However, unlike tracking all upper bodies, our framework allows for selective tracking of *specific body parts* within the upper body

and following the given trajectory.

Physics-based Human Motion Capture. In recent years, the research community has developed various framework to recover human 3D poses [1, 6, 16, 30–32, 37, 56] and motions [2, 9–11, 13, 15, 19, 21, 36, 50, 51, 53] from images and videos [14, 44, 70]. To ameliorate the physical artifacts (*e.g.* foot slidings) associated with the captured motion, recent work has sought to take advantage of the physical attributes of human dynamics. These methods can be broadly classified into three categories: (1) post-optimization based methods during test time [7, 43, 48, 60], (2) reinforcement learning (RL) based methods[24, 25, 39, 57, 66–69] with motion imitation, and (3) physics-aware models [20, 49, 71] to adjust global trajectories.

Our framework is also capable of capturing physically plausible human motion via tracking high-confidence keypoints [57, 70]. However, the main objective of our paper is to achieve zero-shot reproduction of pedestrian motions in real-world driving scenarios. In contrast to existing approaches, our framework does not involve additional optimization for infilling missing frames and low-confidence motions for the captured motion in real-world driving scenarios while tracking high-confidence motions. After we reproduce these real-world scenarios, our framework is also capable of argument these environments with additional virtual pedestrians or editing infilled frames.

3. Methodology

In this paper, we mainly focus on building up on-demand control of pedestrian animation, which encompasses two main aspects: (1) trajectory following on terrains, which determines the desired path of the simulated pedestrian in complex environments, and (2) motion content control, which specifies the desired actions and gestures exhibited by the pedestrian (*e.g.*, making a phone call or waving a hand) while adhering to the provided trajectory and terrain.

To achieve our objective, our framework builds upon PACER [46] and investigates the synergy between motion imitation and trajectory following tasks. In the context of pedestrian animation in driving scenarios, the lower body motion is typically influenced by the trajectory and terrain, while the upper body motion can leverage rich semantic information specific to pedestrians. This grants the upper body the freedom to track a diverse range of possible motions. To attain fine-grained control over different body parts we introduce a per-joint spatial-temporal mask rather than tracking all body parts throughout the sequence. This mask indicates the presence of a reference motion that the policy should track. Using this tracking task, our framework enables diverse pedestrian behaviors at specific time steps and locations in a *zero-shot* manner. This means that we can generate a wide range of motion behaviors without the need for additional training or optimization. Our framework also seamlessly integrates generative human motion models, motion capture sequences, and videos into the simulation system.

Our framework is designed not only for manually synthetic scenarios but also for simulating pedestrians from real-world videos, as demonstrated in [57]. To enable accurate tracking of various parts of pedestrian motion in real-world videos, we expand the spatial-temporal mask to cover whole-body joints instead of solely the upper body. This enhancement allows our framework to track high-confidence motion obtained from pose estimation methods, particularly in real-world captured driving scenarios. By incorporating this capability, our framework becomes more versatile and applicable, showcasing its potential for realistic synthesis and tracking of pedestrian motion in real-world settings. This feature ensures smooth continuity and accuracy in the animation when integrating real-world data into the simulated environments while preserving the motion characteristics observed in real-world scenarios.

In Section 3.1 to Section 3.3, we provide detailed insights into our controller. Subsequently, in Section 3.4, we delve into the integration of different motion content and scenarios within our framework. We discuss how our controller seamlessly adapts to various types of motion content, including generative models, motion capture sequences, and videos. Furthermore, we explore the applicability of our framework to different scenarios, allowing the generation of diverse pedestrian behaviors in specific contexts.

3.1. Physics-aware Character Control

In this section, we first present the formulation of our pedestrian animation controller. In the following Section 3.2 and Section 3.3, we will introduce the details of the tasks in this framework.

Formulation. We follow the general framework of goal-conditioned RL, as shown in Figure 2. The objective of our controller encompasses two aspects: (1) faithfully following the given trajectory $\mathcal{P}$ on terrain $\mathcal{G}$, and (2) imitating the specified motion content $\hat{\mathcal{Q}} = \hat{q}_{1:t}$ provided by our content module within the designated time range $\{t_s : t_s + t\}$ along the trajectory.

Similar to prior works [27, 40, 46], we formulate our character control as a Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$, comprising states, actions, transition dynamics, reward function, and the discount factor. The state $s_t \in \mathcal{S}$ and the transition dynamics $\mathcal{T}$ are determined by the underlying physics simulator, while the action $a_t \in \mathcal{A}$ is computed by our policy network. The reward $r_t \in R$ relates to the given trajectory and motion-tracking task. The objective of our policy is to maximize the accumulated discounted reward $\sum_{t=0}^{T} \gamma^t r_t$, where γ represents the discount factor. To accomplish this,

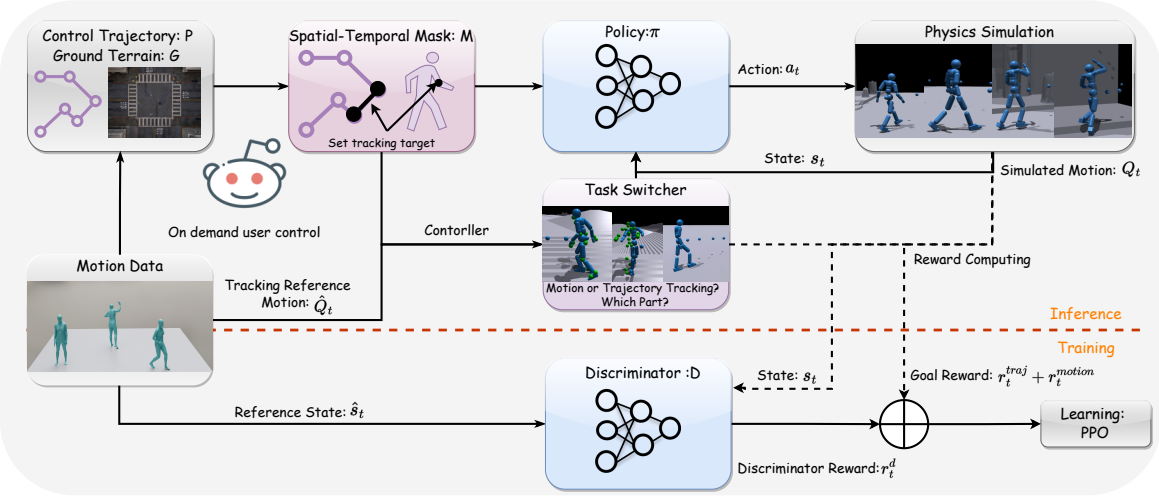

Figure 2. Framework of PACER+. Our framework follows the goal-conditioned reinforcement learning with Adversarial Motion Prior. To enable fine-grained control of specific body parts, we introduce an additional spatial-temporal mask to the motion-tracking task. This mask indicates the presence of a reference motion that the policy should track. By focusing on this tracking task, our framework enables the demonstration of diverse pedestrian behaviors at specific time steps and locations in a *zero-shot* manner.

we employ the widely adopted proximal policy optimization (PPO) algorithm [47].

State and Actions. In our framework, the state $s_t \triangleq (s_t^p, s_t^g)$ consists of humanoid proprioception [27] s_t^p and the goal state s_t^g. The goal state s_t^g consists of two components, as the goal for trajectory following s_t^{traj}, and the goal for motion tracking s_t^{motion}. We will present the details of these two components in the following sections. We use a proportional derivative (PD) controller at each degree of freedom (DoF) of the humanoid to control pedestrian animation.

Adversarial Motion Prior. Similar to the previous state-of-the-arts [27, 40, 41, 46], we learn our optimal control policy with Adversarial Motion Prior (AMP). AMP employs a motion discriminator to encourage the policy to generate motions that align with the movement patterns observed in a dataset of human-recorded motion clips. Specifically, AMP uses a discriminator to compute a style reward, which is added to the task reward: $r_t = 0.5r_t^{amp} + 0.5(r_t^{traj} + r_t^{motion})$. We will illustrate the details of the task reward in the following sections.

3.2. Trajectory following on terrains

Trajectory Following State. In the trajectory following task, the humanoid a local height map $\mathcal{G}$ and the trajectory $\mathcal{P}$ to follow. The 3D trajectory input is defined as $\mathcal{P}_t^{traj} = \{\hat{p}_t, \hat{p}_{t+\rho}, \cdots, \hat{p}_{t+N\rho}\}$, where ρ is the sampling rate of the trajectory, and N is the number of steps in the future. $\hat{p}_{t+\rho}$ is the relative xy value between the position of path $\mathcal{P}$ at time step $t+\rho$ and the root position of simulated character at time step t. In practice, we set ρ as 0.5 seconds and N as 10. For the height map of the ground terrain $\mathcal{G}^t$, we render a 32x32 square centered at the root of the humanoid and render the local height map as input $\mathcal{G}^t$. Therefore, the goal state of the trajectory following task can be defined as $s_t^{traj} \triangleq (\mathcal{P}_t^{traj}, \mathcal{G}_t)$.

Trajectory Following Reward and Early Termination. Trajectory following task reward is defined as xy distance between the position of trajectory $\hat{p}_t^{xy}$ and the root position of the simulated character r_t^{xy} at time step t, formulated as $r_t^{traj} = e^{-2||\hat{p}_t^{xy} - r_t^{xy}||}$. To better follow the trajectory, we introduce an early termination mechanism to this task while training the policy network. Specifically, we terminate the trajectory following task if the distance between the position of trajectory $\hat{p}_t^{xy}$ and the root position of the simulated character r_t^{xy} at time step t is larger than a threshold τ. We set τ at 0.5 meters in our experiments.

3.3. On-demand Motion Tracking

Masked Motion Tracking. In contrast to previous works [24, 27, 58], our motion tracking tasks deviate in that we require the policy network to track specific motion parts within a given time range while following trajectories. To facilitate this, we introduce a spatial-temporal mask to the tracking tasks, denoted as $\mathcal{M}1 : T = \{m_1, m_2, \cdots, m_T\}$, where $m_t = \{m_t^1, \cdots, m_t^J\}$ is a set of binary masks indicating whether the motion tracking task j is required at

time step t. By employing this observation mask, we can define the state of the motion-tracking task and the reward function as follows.

Motion Tracking State. The motion content of our track task $\hat{q}_{t+1}$ for the frame $t+1$ consists of joint position $\hat{p}_{t+1}$, joint rotation $\hat{\theta}_{t+1}$, joint velocity $\hat{v}_{t+1}$, and rotation velocity $\hat{\omega}_{t+1}$, similar to the rotation-based imitation of PHC [27]. In our simulation stage, we can only set the motion demonstration tasks for some specific frames, rather than tracking all frames as [27]. In general, for frames without motion demonstration tasks at time step t_1, we directly set the mask as 0 to indicate that motion tracking tasks are not required at these time steps. For the tracking target, we can directly set it as the same value as the state of the simulated character. For the frame t_2 with target motion, we can set the mask $m_{t_2} = \{m_{t_2}^1, \cdots m_{t_2}^J\}$ with 1 for the joints that should be tracked and 0 for the ignored joints. We also set the target motion as the same value as the state of the simulated character for the ignored joints. Therefore, the state of motion content demonstration at time step t can be defined as $S_d \triangleq (\hat{\theta}_{t+1} - \theta_t, \hat{p}_{t+1} - p_t, \hat{v}_{t+1} - v_t, \hat{\omega}_{t+1} - \omega_t, \hat{\theta}_{t+1}, \hat{p}_{t+1}, m_{t+1})$.

Demonstration Reward and Early Termination. The reward of our motion demonstration is mainly related to the motion tracking error between the simulated character and the target motion. Therefore, we can define the reward as $r_t^{motion} = w_{jp}e^{-100||\hat{p}_t - p_t||\circ m_t} + w_{jr}e^{-10||\hat{q}_t - q_t||\circ m_t} + w_{jv}e^{-0.1||\hat{v}_t - v_t||\circ m_t} + w_{rv}e^{-0.1||\hat{\omega}_t - \omega_t||\circ m_t}$. The mask m_t helps us to ignore the joints that should not be demonstrated in the simulation process.

For effective training of our motion tracking task using the Adversarial Motion Primitives (AMP) approach, we made two critical design choices: incorporating additional motion sequences and implementing early termination. To address mode collapse [40, 46], we trained AMP with a smaller dataset of approximately 200 sequences, as discussed previously [46]. While this selection ensures naturalness in generated motions, it limits generalization to unseen motion in the motion-tracking task. Including supplementary motion sequences as references in motion tracking introduces diverse motion content and has the potential to enhance tracking performance. However, training the motion tracking task with an additional dataset poses a challenge in jointly learning AMP alongside the smaller dataset as contrasting motion styles are introduced. *Supplementary videos* visually depict the challenges faced by the policy network in accurately learning motion-tracking outcomes.

To overcome this limitation, we incorporate an early termination mechanism during training. Specifically, we terminate the motion demonstration task if the largest distance between the joint positions of the reference poses $\hat{p}_t^{xy}$ and the simulated character p_t^{xy} at time step t exceeds a threshold τ. In our experiments, we set τ to 0.3 meters. We use more than 10,000 motion sequences from the AMASS dataset [28] to train our motion demonstration task for practical implementation.

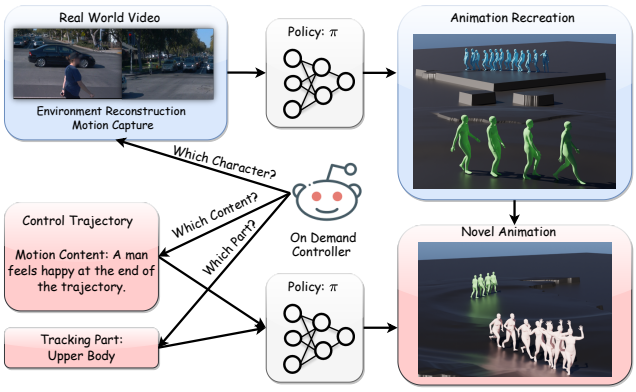

Figure 3. Our framework presents an on-demand control system tailored for real-world videos. Beginning with the pre-processing in [57], our policy network can track high-confidence motions and effectively fill in missing parts without additional fine-tuning. Moreover, our framework offers the novel functionality of introducing customized animations into real-world scenarios with flexible control options.

3.4. System Overview

Finally, we outline the training process for our policy using the combination of these tasks. Subsequently, we introduce the methodology for controlling pedestrian animation on-demand using our framework in both manually synthetic and real-world scenarios.

Training Procedure. In our framework, the policy network undergoes training through a combined approach of trajectory-following and motion-tracking tasks. Initially, each training environment involves joint training of the policy with both tasks. During this step, binary masks for the reference motion are randomly generated at each time step, and early termination is applied to the motion-tracking task. The reference motions are sampled randomly from the AMASS dataset [28]. Subsequently, we train the trajectory following task using randomly generated synthetic trajectories [46, 57]. In this stage, all joints within the spatial-temporal mask are assigned a value of 0. This ensures that the policy focuses solely on learning to follow the generated trajectories without considering motion tracking. By employing this combined training approach, we enable the policy network to acquire proficiency in trajectory-following and motion-tracking tasks, enhancing our framework's overall performance and adaptability.

Manually Synthetic Scenarios. In manually synthetic scenarios, our framework offers flexibility in manually setting the trajectory while generating the desired motion content. First, we identify the specific body part from the motion content obtained from other references. Then, we set the motion tracking task's mask to indicate the desired motion for demonstration. During this process, we align the reference motion's location and orientation with the trajectory to facilitate accurate tracking. This alignment ensures that the generated motion content precisely follows the specified trajectory, allowing for diverse and customizable pedestrian animations. Our experimental results will present further details and insights on this approach. By employing this methodology, our framework empowers the generation of tailored motion content for manually synthetic scenarios, enabling greater control and realism in the animation process.

Real-world Scenarios. In our real-world scenarios, we adopt the definition of high-confidence frames, as described in prior works [57, 70], using 2D keypoint detection. We track the entire body motion for these high-confidence frames to maintain optimal motion content. Conversely, in low-confidence frames, we assign a value of *1* only to keypoints with high-confidence estimation scores in the spatial-temporal mask. This approach enables motion capture even when half of the body is occluded without requiring additional optimization steps. Additionally, we can apply the same process as in manually synthetic scenarios to introduce additional content from other sources into real-world scenarios using our unified policy. Figure 3 illustrates this capability. By following this approach, we enhance the quality and realism of animation in real-world videos, leveraging the flexibility of our framework.

4. Experiments

Dataset. In our experiments, we utilized motion data from various sources. We employed motion from the AMASS dataset [28] for motion tracking evaluation. To enhance the diversity of demonstrated motion, we collaborated with off-the-shelf language-based motion generation models [4, 55]. Additionally, we utilized NIKI [21], a state-of-the-art human motion capture approach, to capture motions from videos and recreate real-world scenarios. Regarding the simulation environment in our framework, it encompasses two aspects: (1) manually synthetic scenarios built using Unreal Engine following the MatrixCity framework [22], and (2) real-world scenarios reconstructed from scanned point cloud data in the Waymo Open Dataset [73]. Following the methodology described in [57], we resampled human motion captured from videos to 30 fps to match the simulation environment. To evaluate the performance of our framework effectively, we selected motion sequences with a trajectory length of more than 3 meters.

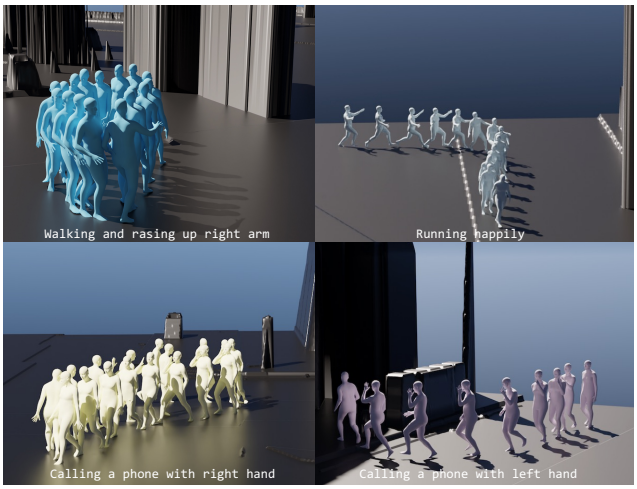

Figure 4. Results on manually synthetic terrains. Our framework enables the synthesis of animations by combining a given trajectory with motion content generated by language-based motion generation models [4, 55].

Metrics. To evaluate our framework, we employed a range of kinematics-based and physics-based metrics. We use motion Fréchet Inception Distance (FID) [18, 42] and diversity metric [4, 55] to evaluate the quality and diversity of synthesized animations. To evaluate tracking accuracy, we employed the Mean Per-Joint Position Error (E_{mpjpe}) and Global Mean Per-Joint Position Error (E_{gmpjpe}) metrics, between the simulated character and the reference motion in root space and global space. Regarding the physical attributes of the animation, we evaluated food sliding (FS) and foot penetration (FL) metrics for animation synthesis, following the methodologies outlined in [20, 69]. Motion jitter is computed by the velocity (Vel) and acceleration (Accel) between the physics character and the reference motion. The units for these metrics are measured in millimeters (mm), except for Accel, which is measured in $mm/frame^2$.

Implementation Details. We followed the capsule model of the SMPL robot as the simulation target, as described in [26, 27, 46]. Our policy network was trained on a single NVIDIA A100 GPU, which took approximately three days to converge. Once trained, the composite policy runs at a frame rate exceeding 30 FPS. The physics simulation is performed in NVIDIA's Isaac Gym [29]. The control policy operates at 30 Hz, while the simulation runs at 60 Hz. In our evaluation, we did not consider body shape variation and used the mean body shape of the SMPL.

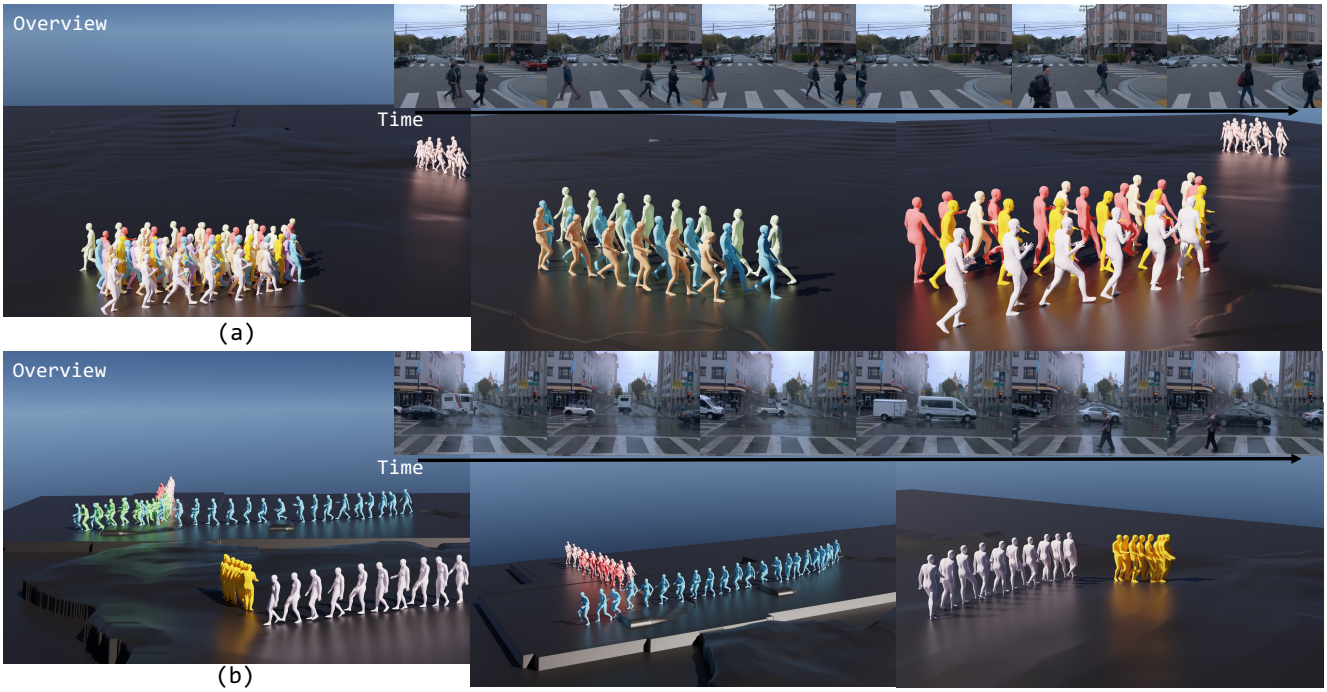

Figure 5. Zero-shot animation recreation of real-world pedestrians. Our framework is capable of simulating pedestrian animation following the motion content of real-world videos.

Table 1. Comparison of Motion Quality and Diversity between Our framework and PACER. FID and Diversity metrics were used for trajectories with normal speed, while *l*-FID and *l*-Diversity are employed for animations under low speed.

Method	FID ↓	Diversity ↑	*l*-FID ↓	*l*-Diversity ↑
PACER [46]	7.97	1.29	8.84	1.24
Ours	**6.74**	**1.67**	**7.62**	**1.36**

Table 2. Motion tracking quality of our method between different body parts by introducing spatial-temporal mask to the corresponding region. We compare with [57] for whole-body tracking because this method can not only track specific regions, *e.g.*, upper-body.

Metric	Wang [57]	Whole	Upper	Left Arm	Right Arm
E_{mpjpe} ↓	80.29	**72.10**	77.87	78.75	79.52
E_{gmpjpe} ↓	137.48	**123.88**	128.15	128.84	133.92

4.1. Evaluation.

In this section, our framework is primarily compared with PCAER [46], the state-of-the-art controllable pedestrian animation approach. The comparison focuses on motion quality and evaluation metrics for specific tasks, such as trajectory following and motion tracking. By comparing our framework with PCAER, we aim to demonstrate the advances and improvements in these areas.

Motion Quality and Diversity. We conducted a comparative analysis between our framework and PACER, focusing on motion quality and diversity. For this evaluation, we randomly synthesized 1000 different trajectories on the synthetic terrain, which were used to train both PCAER and our policy. The motion content of our framework is synthesized by off-the-shelf approaches [4, 55]. Table 1 presents the results of this comparison. Our method achieves a lower FID and demonstrates better diversity compared to PACER. These findings indicate the superior ability of our framework to generate diverse and contextually relevant pedestrian animations. Additionally, we consider the issue of synthesizing animations at low speeds, which can often result in unnatural motion, as presented in PACER. Specifically, we compare our framework with PACER under trajectories with low speeds (speed $< 1m/s$). As shown in Table 1, our framework consistently achieves better motion realism and diversity in these low-speed scenarios. Overall, our framework surpasses PACER in motion quality and diversity, showcasing its advancements and improvements in realistic and diverse pedestrian animation synthesis.

Motion Tracking. To evaluate the motion tracking performance of our framework, we utilize motion content from two sources: the AMASS dataset [28] and synthesized motion generated by state-of-the-art motion generation models [4, 55]. We randomly select 1000 sequences from AMASS to assess the tracking performance, providing diverse and real-world motion content. Additionally, we generate 200 synthesized motion sequences using Chat-

Table 3. We present the results of our method on real-world scenarios and compare with [57]

Method	$E_{mpjpe}\downarrow$	$E_{gmpjpe}\downarrow$	FS $\downarrow$	FL $\downarrow$	Vel $\downarrow$	Acc $\downarrow$
Motion [21]	×	×	45.32	54.21	×	×
Wang [57]	89.42	137.84	7.87	14.21	8.21	7.42
Ours	**77.67**	**127.84**	**7.68**	**12.12**	**7.42**	**6.43**

Table 4. Ablation studies on motion tracking and spatial-temporal mask. Our design choice achieves better results on both motion tracking and motion quality.

Tracking	Mask	FID	$E_{mpjpe}\downarrow$	$E_{gmpjpe}\downarrow$
×	×	7.97	215.55	254.17
√	×	7.07	79.57	132.24
√	√	**6.74**	**77.87**	**128.15**

GPT [35] with 20 distinct prompts for driving scenarios, resulting in 10 sequences per prompt. The evaluation focuses on synthetic terrains and trajectories, with a comprehensive assessment of whole-body, upper-body, and left/right arm tracking. The tracking results are presented in Table 2, allowing us to analyze and quantify the effectiveness of our framework in different tracking scenarios and body parts. Furthermore, we compare our method with [57] for whole-body tracking, demonstrating superior *zero-shot* tracking results on terrains.

4.2. Results on Real-world Scenarios

In real-world scenarios, we evaluate our framework using the NIKI [21] to obtain joint rotations of the human body. Following the evaluation methodology outlined in [57], we use the ground truth trajectory and 2D bounding box to assess our framework's performance. To evaluate the confidence of the estimated results, we employ ViTPose [63] to extract confidence scores for each body joint. During the inference process, we selectively track body parts with high-confidence joint estimations, ensuring a fair comparison by refraining from additional fine-tuning or optimization, as stated in [57]. This unbiased evaluation allows for a comparison of our framework's performance. Our method demonstrates improvements in the physics attributes of the motion content, as presented in Table 3. Moreover, it achieves better E_{mpjpe} and E_{gmpjpe} results, indicating improved matching to high-confidence parts and the given trajectory compared to [57]. Through these evaluation techniques, we showcase the results of our framework in real-world scenarios, highlighting its performance and effectiveness in practical settings.

4.3. Ablation Study

We performed our ablation study at Table 4 to assess the effectiveness of motion tracking and the spatial-temporal mask in our framework. The study focused on upper body tracking and trajectory following. When motion tracking is not included, our framework resembles PACER and can not follow the content of the given motion sequences. Consequently, the motion quality was inferior to our framework. However, upon introducing the motion tracking task, combined with the spatial-temporal mask, the policy exhibited improved motion tracking and enhanced realism quality. The results of the ablation study highlight the significance of motion tracking and the spatial-temporal mask, underscoring their contributions to the effectiveness and quality of our framework.

4.4. Qualitative Results

Figure 4 showcases the synthesized animations on artificial terrains. All presented results adhere to the control of the given trajectory and upper body motion content. Our framework enables the synthesis of diverse and natural human animations, surpassing the limitations of conventional walking and running actions [46]. Furthermore, Figure 5 illustrates the *zero-shot* results of animation recreation. These examples highlight the capability of our framework to recreate animations in real-world scenarios. We refer viewers to our supplementary video for a more comprehensive presentation, including different tracking parts and collaborations with various motion sources. The synthesized animations on synthetic terrains and the animation recreation results demonstrate the effectiveness and versatility of our framework in generating diverse and natural human animations.

5. Conclusion and Limitation

Conclusion: In this paper, we introduce a novel framework for on-demand synthesis of diverse and natural pedestrian animation in driving scenarios. Our framework surpasses traditional trajectory control methods by enabling zero-shot generation of diverse motion using a range of motion content sources. To achieve this, we propose a joint tracking framework where a single policy is trained to simultaneously track the trajectory and imitate selected joints, such as upper-body joints. During training, we incorporate a spatial-temporal mask to guide the policy network in tracking specific joints within a designated time range. Our framework empowers comprehensive control over pedestrian animation in both manual and synthetic scenarios, offering a versatile tool for animation generation.

Limitations and Future Works: Our current approach uses pre-trained motion generation models for motion content and relies on user-provided trajectories, without explicitly considering the semantic relationship between pedestrians and the environment. In future work, we aim to investigate generating motion content directly through the policy network while incorporating semantic guidance.

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. 3

[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. 3

[3] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. *ACM Transactions On Graphics (TOG)*, 2023. 2

[4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 6, 7

[5] 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

[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. 3

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

[8] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 2

[9] 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. 3

[10] 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.

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

[12] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. *ACM Transactions on Graphics (TOG)*, 2022. 2

[13] 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. 3

[14] 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. 3

[15] 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. 3

[16] 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. 3

[17] Yoonsang Lee, Sungeun Kim, and Jehee Lee. Data-driven biped control. In *ACM SIGGRAPH 2010 papers*, pages 1–8. 2010. 2

[18] 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

[19] 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. 3

[20] 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, 6

[21] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. NIKI: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 6, 8

[22] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. 2023. 6

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

[24] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. *NIPS*, 2021. 2, 3, 4

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

[26] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. *arXiv preprint arXiv:2310.04582*, 2023. 2, 6

[27] Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 5, 6

[28] 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. 5, 6, 7

[29] 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

[30] 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. 3

[31] 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.

[32] 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. 3

[33] 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. 2

[34] NVIDIA. Drive sim. https://developer.nvidia.com/drive/simulation, 2020. 1

[35] OpenAI. Chatgpt. https://openai.com/chatgpt, 2020. 8

[36] 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. 3

[37] 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. 3

[38] 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. 2

[39] 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. 3

[40] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 2021. 2, 3, 4, 5

[41] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 2022. 2, 4

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

[43] 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

[44] 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. 3

[45] 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. 1

[46] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3, 4, 5, 6, 7, 8

[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4

[48] 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), 2020. 3

[49] 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), 2021. 3

[50] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 3

[51] 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. 3

[52] 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

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

[54] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. *ACM Transactions on Graphics (TOG)*, 2023. 2

[55] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 6, 7

[56] 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. 3

[57] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. 1, 2, 3, 5, 6, 7, 8

[58] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 39(4):33–1, 2020. 2, 4

[59] 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

[60] 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

[61] Pei Xu, Xiumin Shang, Victor Zordan, and Ioannis Karamouzas. Composite motion learning with task control. *ACM Transactions on Graphics (TOG)*, 2023. 2

[62] Pei Xu, Kaixiang Xie, Sheldon Andrews, Paul G Kry, Michael Neff, Morgan McGuire, Ioannis Karamouzas, and Victor Zordan. Adaptnet: Policy adaptation for physics-based character control. *ACM Transactions on Graphics (TOG)*, 2023. 2

[63] 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. 8

[64] Heyuan Yao, Zhenhua Song, Baoquan Chen, and Libin Liu. Controlvae: Model-based learning of generative controllers for physics-based characters. *ACM Transactions on Graphics (TOG)*, 2022. 2

[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. 2

[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. 3

[67] 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.

[68] 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

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

[70] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 3, 6

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

[72] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[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