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

Jingbo Wang[1*] Zhengyi Luo[2*] Ye Yuan[3] Yixuan Li[4] Bo Dai[1]
[1]Shanghai AI Lab [2]Carnegie Mellon University [3]NVIDIA [4]The Chinese University of Hong Kong
* Donates Equal Contribution

1. Implementation Details

Humanoid State. In our policy, the state of the humanoid, denoted as s_t^p, comprises joint positions $j_t \in \mathbb{R}^{24\times3}$, rotations $q_t \in \mathbb{R}^{24\times6}$, linear velocities $v_t \in \mathbb{R}^{24\times3}$, and angular velocities $\omega_t \in \mathbb{R}^{24\times3}$. These components are normalized with respect to the agent's heading and root position in our simulator. The rotation q_t is represented using the 6-degree-of-freedom rotation representation [7].

Figure 1. Our simulated motion near a moving car allows for interactive actions, such as waving, when the pedestrian stops or turns around. This enhances the realism of the simulated motion.

Network Architecture. In this study, we adopt the network structure from PACER [2], which separates the policy network into a task feature processor and an action encoder. For the task processor, we employ a convolutional neural network (CNN) to process the terrain map, following the approach outlined in [2]. Additionally, we utilize an MLP network to encode the trajectory, reference motion, and tracking mask. Subsequently, we concatenate the humanoid state with the output of the task processor to form the input of the action network. The action network consists of MLP layers with ReLU activations, comprising two layers with 2048 and 1024 units, respectively. The output of this action network, indicated as $a_t \in \mathbb{R}^{23\times3}$, corresponds to the PD target of the joints on the SMPL [1] human body, excluding the root joint.

2. Additional Results.

The trajectory following results are presented in Table 1. We mainly compare our method with the approaches proposed in [2] and [4]. We evaluate these different methods on the synthetic terrain and trajectories similar to training these policies. The metric is the average deviation of the character from trajectories. As depicted in the table, PACER achieves the most favorable results. Our method demonstrates performance comparable to that of PACER in terms of trajectory following. However, it is worth noting that the kinematics policy employed in [4] exhibits limitations when dealing with complex terrains without fine-tuning.

Table 1. Comparison of trajectory following task with PACER [2] and [4]. Our method achieves a comparable result with PACER and a significantly better result than [4].

Method	PACER [2]	Wang et.al [4]	Ours
Error ↓	**0.118**	0.164	0.123

Additionally, as shown in Figure 1, when dealing with a simulated car, we can manually determine the trajectory and movements of the pedestrian to enable more grounded reactions. For example, we can incorporate actions such as waving in the car or altering the direction of movement. In this work, we focus mainly on the "on-demand" prospect on increasing the capabilities of simulators, which can pave the way for a more antonymous response in intelligent agents.

3. System Details.

Figure 2 showcases the user interface (UI) we have developed, which allows users to control the animation within the specified driving scenario. Through this system, users have the capability to modify the trajectory, motion content, body parts, and the starting time of the motion content seamlessly in a zero-shot manner. Further details regarding this system can be found in our supplementary video.

4. Further Discussion.

In the realm of integrating language-based motion generation models with physics simulators, recent work has emerged to address this area. Physdiff [6] stands as a notable contribution, employing a whole-body imitator within the motion diffusion model's denoising process to achieve

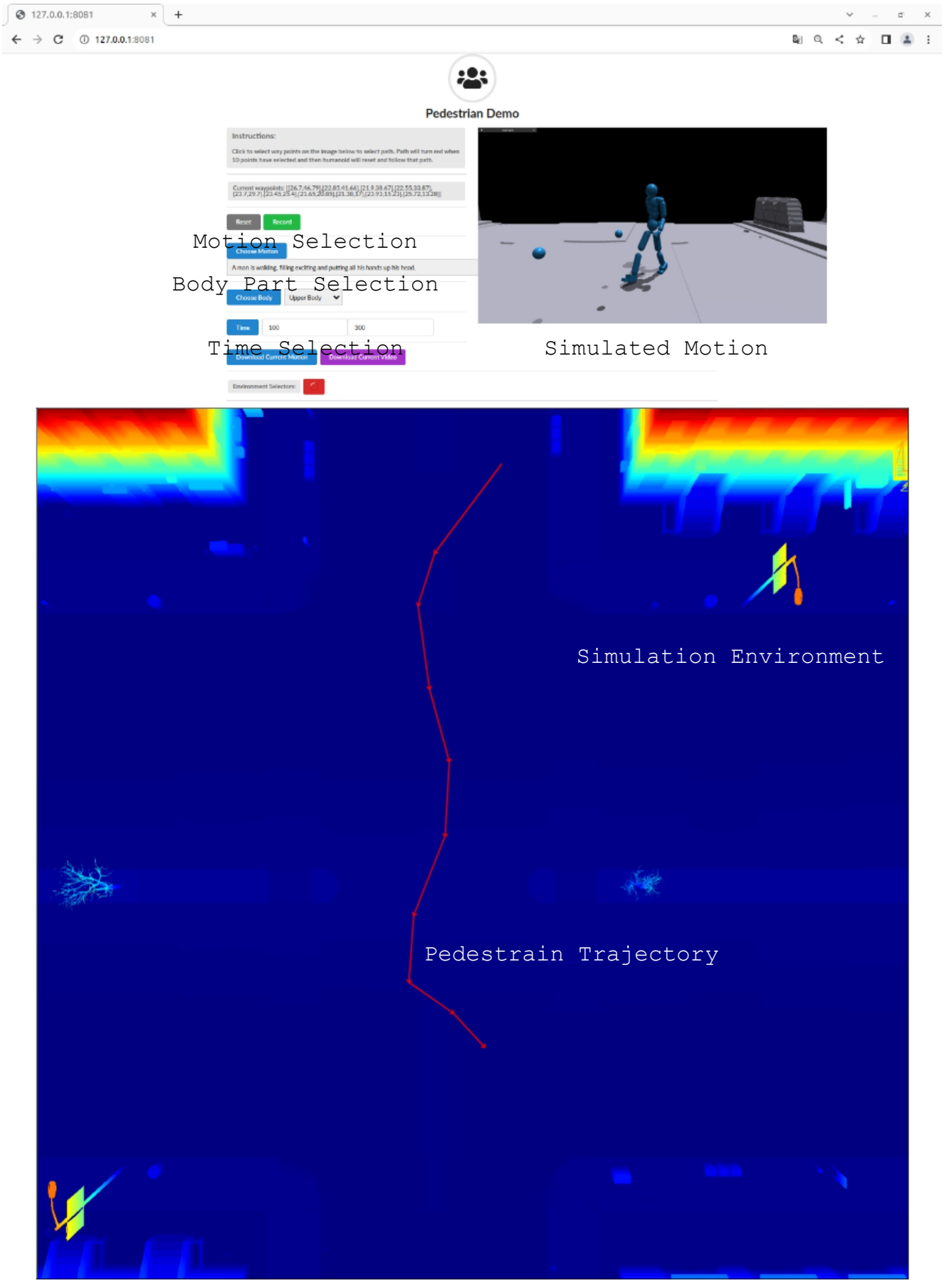

Figure 2. User interface (UI) of our on-demand animation controller. Our system is capable of changing trajectory, environment, motion content, and body parts for pedestrian animation in driving scenarios.

physically plausible animations. On the other hand, InsActor [3] and MoConVQ [5] rely on model-based character simulation. However, these approaches overlook the influence of terrain in pedestrian animation and lack body part control. Consequently, these methods are limited to controlling the entire body joints based on language instructions solely on flat-ground scenarios. In contrast, our framework offers enhanced flexibility in control while also facilitating cooperation with real-world motions on various terrains. Additionally, with the advances in motion content synthesis by language-based motion models, our framework has the potential to generate superior animations in driving scenarios.

References

[1] 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. 1

[2] 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] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, Xiao Ma, Liang Pan, and Ziwei Liu. Insactor: Instruction-driven physics-based characters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[4] 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

[5] Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *arXiv preprint arXiv:2310.10198*, 2023. 3

[6] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[7] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 1