A Plug-and-Play Physical Motion Restoration Approach for In-the-Wild High-Difficulty Motions

Supplementary Material

MCM	TTA	Mask	Prior	WA-MJE ↓	RTE ↓	MPJPE ↓	MPS ↑
				155.517	7.279	116.183	0.732
✓				149.081	7.292	113.580	0.742
✓	✓			129.736	5.049	98.246	0.765
✓	✓	✓		131.872	5.213	**93.896**	**0.806**
✓	✓	✓	✓	**123.365**	**4.850**	94.037	**0.806**

Table 1. **Additional Ablation Study on AIST++.**

Datasets	Method	SR ↑	E_{g_mpjpe} ↓	E_{mpjpe} ↓	E_{pa_mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
AIST++	UHC	47.34%	147.5	67.83	49.42	5.59	7.76
	PHC+	69.85%	93.17	51.92	46.6	3.05	4.16
	PTM	**94.43%**	**62.28**	**31.68**	**29.73**	**2.59**	**3.41**
H36M	UHC	23.6%	133.14	67.4	52.91	14.9	17.2
	PHC+	92.9%	50.31	33.34	30.34	3.74	5.52
	PTM	**98.84%**	**44.73**	**30.82**	**24.65**	**2.06**	**3.12**
kungfu	UHC	42.91%	86.23	48.91	39.73	12.11	9.57
	PHC+	76.41%	84.86	47.98	39.43	5.54	7.89
	PTM	**98.16%**	**72.13**	**33.45**	**26.12**	**3.95**	**4.23**

Table 2. **Physical transfer ability.**

1. Additional Experimental Results

Additional Ablation Study for Motion Restoration. To further exemplify the effectiveness of the various parts of our approach in more general scenarios, we performed ablation experiments on the AIST++ dataset. As shown in Table 1, all parts of our method show beneficial effects. The test-time adaptation strategy and the prior motion enhanced various metrics, especially the globally relevant WA-MPJPE and RTE. Meanwhile, the role of Mask guidance in the adaptation process is mainly manifested in the MPJPE and MPS metrics in the camera coordinate system. In addition, although there are relatively few difficult motions in the AIST++ dataset, our MCM still reflects a positive role in motion restoration.

Physical transfer ability (Motion tracking ability). We conduct experiments on three datasets, AIST++, H36M, and kungfu, to verify the effectiveness of PTM on the motion tracking task; the results are presented in Table 2. On all three datasets, our PTM outperforms current mainstream reinforcement learning methods. The improvement in metrics is attributed to our proposed pre-training and adaptation design tailored for complex motions and the ability of motion prior to mitigate the forgetting phenomenon and overfit issues. Previous studies have indicated that motion imitation can suffer from a rapid loss of earlier knowledge when attempting to imitate newer motions [7]. In our PTM, the motion prior and the pre-trained controller serve as a cornerstone for learning new actions, allowing the optimization of specific data samples with respect to human motion patterns it learned before. We aim for the model to proactively explore solutions rather than merely reproducing answers it has memorized.

2. Additional Visualization

Flaw Motion Cases. In Figure 1, we provide additional cases of flawed motion shown in the advanced video motion capture method GVHMR [11], flawed motion usually happens from the rapid and extreme movement of high-difficulty motions and the blurred frames. Our method successfully corrected these flawed motions and converted them to physical realism motions.

Physical Restoration Visualization. In Figure 2, we select high-difficulty in-the-wild motions (taekwondo and rhythmic gymnastics) and illustrate a comparison before and after our restoration. As shown in the figure, our method effectively eliminates issues of ground penetration and floating in the original motions, successfully transferring the raw movements into physical space. Despite the original motions containing multiple continuous complex actions that are challenging to reproduce in physical space, our method effectively eliminates issues of ground penetration and floating in the original motions while successfully maintaining the original motion patterns. Notably, our method has never seen taekwondo or rhythmic gymnastics motions before, which underscores the strong generalization ability of our method.

Physical Restoration for Motion Generated by Text2motion Method. In Figure 3, although the state-of-the-art motion generation method Momask still generates motion with physical inaccuracies such as mold penetration and levitation, our method successfully repairs the physical realism of the generated results while maintaining motion quality comparable to that of the original generated results.

Visualization of different motion types. In Figures 4 and 5, we demonstrate the performance of our method on various types of challenging motions, including Taekwondo, Chinese kungfu, breakdancing, rhythmic gymnastics, and ballet. The visualized results show that our method can effectively perform physics-based motion restoration for a diverse range of high-difficulty movements, producing high-fidelity motions to the original video. Notably, our training dataset does not include such a wide variety of complex motions, nor does it rely on any expensive 3D annotations, which further proves the effectiveness and broad applicability of our approach.

Additional SOTA Comparison. In Figure 6, we pro-

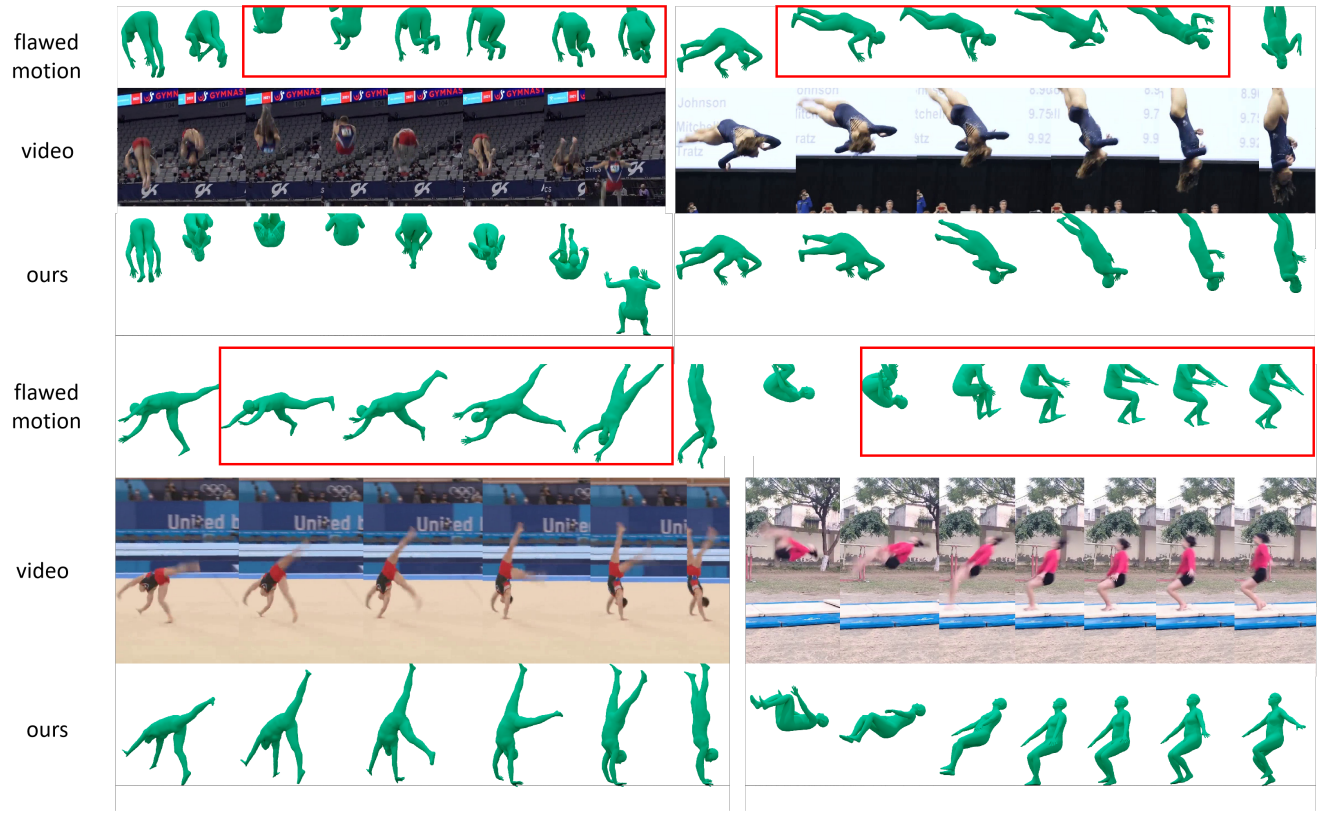

Figure 1. **Additional cases of flawed motion.**

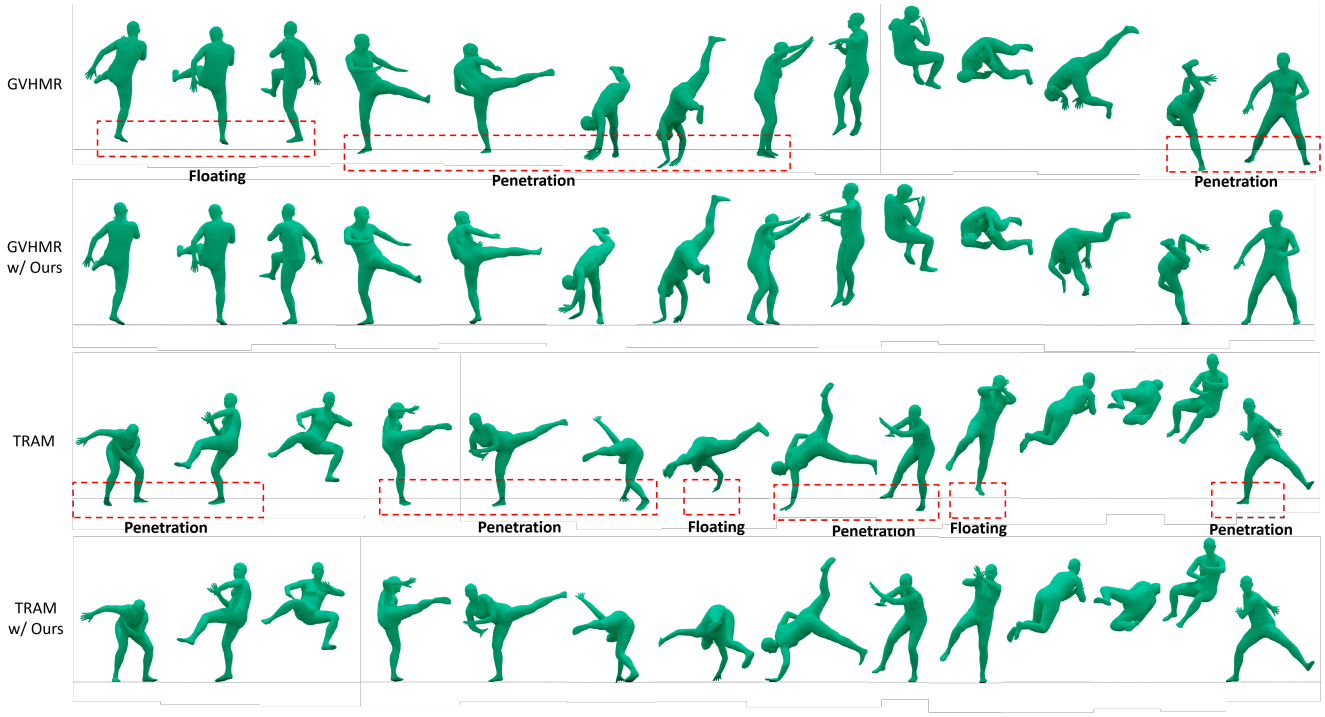

Figure 2. **Visualization of the high-difficulty motions before and after our physical restoration .**

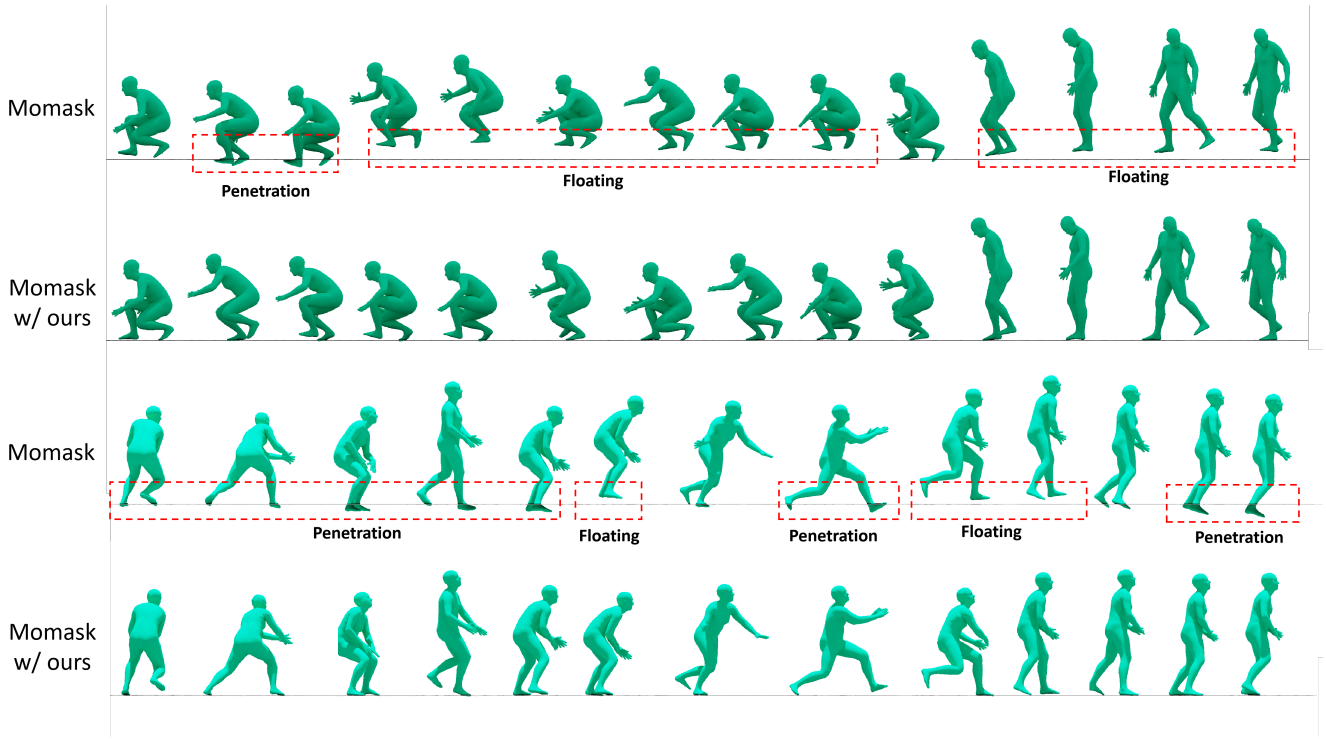

Figure 3. **Physical restoration for motion generated by text2motion method.**

vide additional visual comparisons with state-of-the-art methods (gymnastics and yoga). The performance of each method on different types of motions is consistent with the visual comparisons presented in the main text. Due to limitations in generalization, PHC+ [7, 8] struggles to simulate high-difficulty movements, resulting in falls during the simulation. PhysPT [14], due to its simplified physical constraints, avoids falling during the restoration process, but its physical repair ability is limited when facing high-difficulty motions. It still suffers from issues related to physical realism and struggles to maintain the original motion patterns of the reference motions. In the first row of Figure 6, the ground in the motion generated by GVHMR appears somewhat unusual as we calculate it with the mean operation. This is because we need to display the entire motion in the visualization. We do not use this approach in other visualizations and metric calculations. Instead, we calibrate the ground based on the foot position in the first frame.

3. Motion Inbetweening

Motion Representation. In the Mask-conditioned Motion Correction Module (MCM), we adopt a simpler representation that ensures convenient conversion to the SMPL [6] format. Given a human motion sequence $\boldsymbol{x} \in \mathbb{R}^{N \times D}$, the segmented human mask $\boldsymbol{m} \in \mathbb{R}^{N \times w \times h}$, and a keyframe signal (which identifies frames requiring inbetweening), this

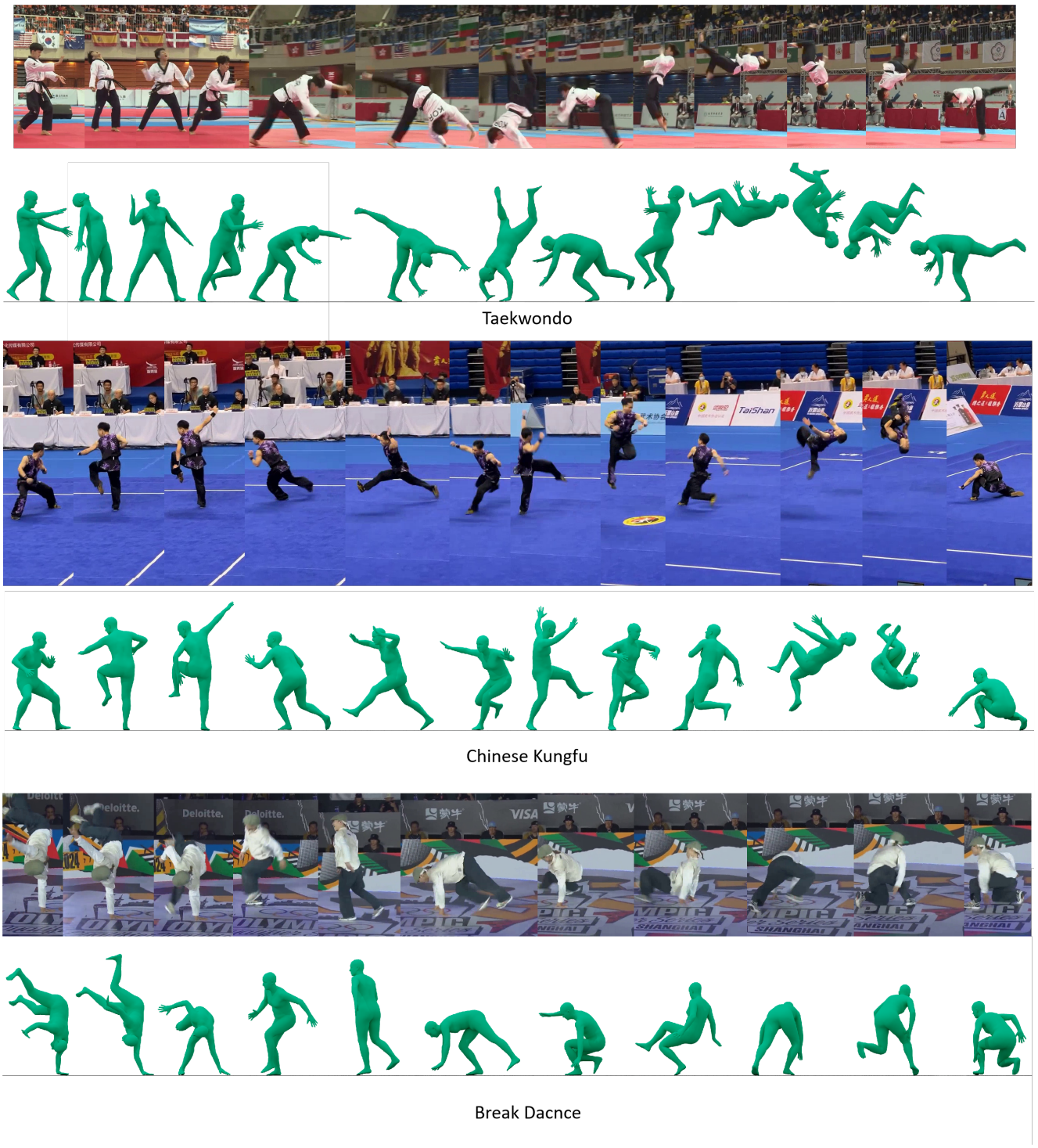

Figure 4. **Visualization of the in-the-wild and high-difficulty motions restored from our method (1/2).**

module restore the mismatched motion frames. The human motion $\boldsymbol{x}$ is derived from the reference motion, where we compute a 135-dimensional motion representation based on SMPL parameters, which includes 3-dim translation, 6-dim

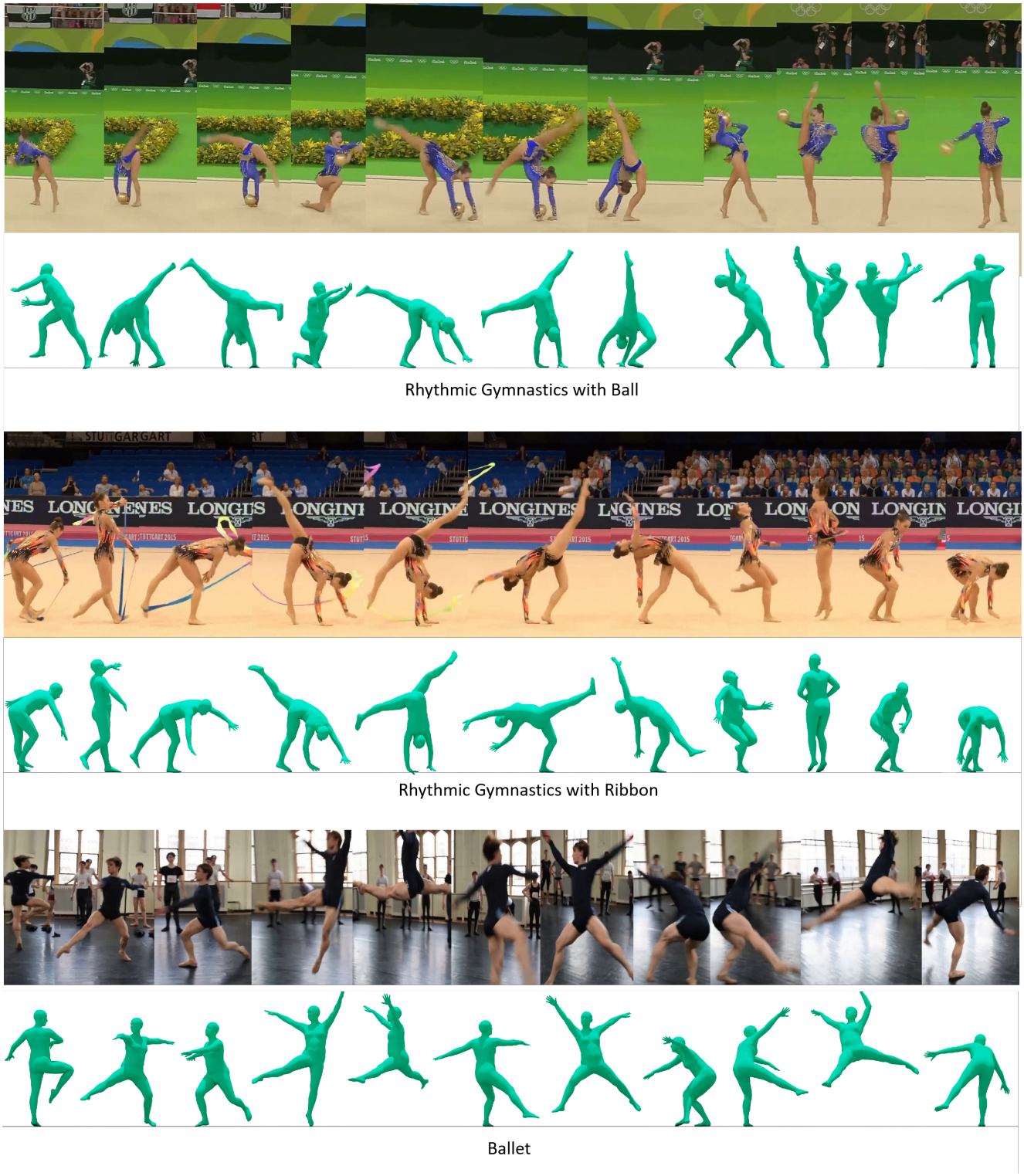

Figure 5. **Visualization of the in-the-wild and high-difficulty motions restored from our method (2/2).**

root rotation, and 21 body joint rotations. The observation signal is generated from the mismatch detection algorithm.

Motion Inbetween with Diffusion Models. Diffusion models have demonstrated impressive performance in gen-

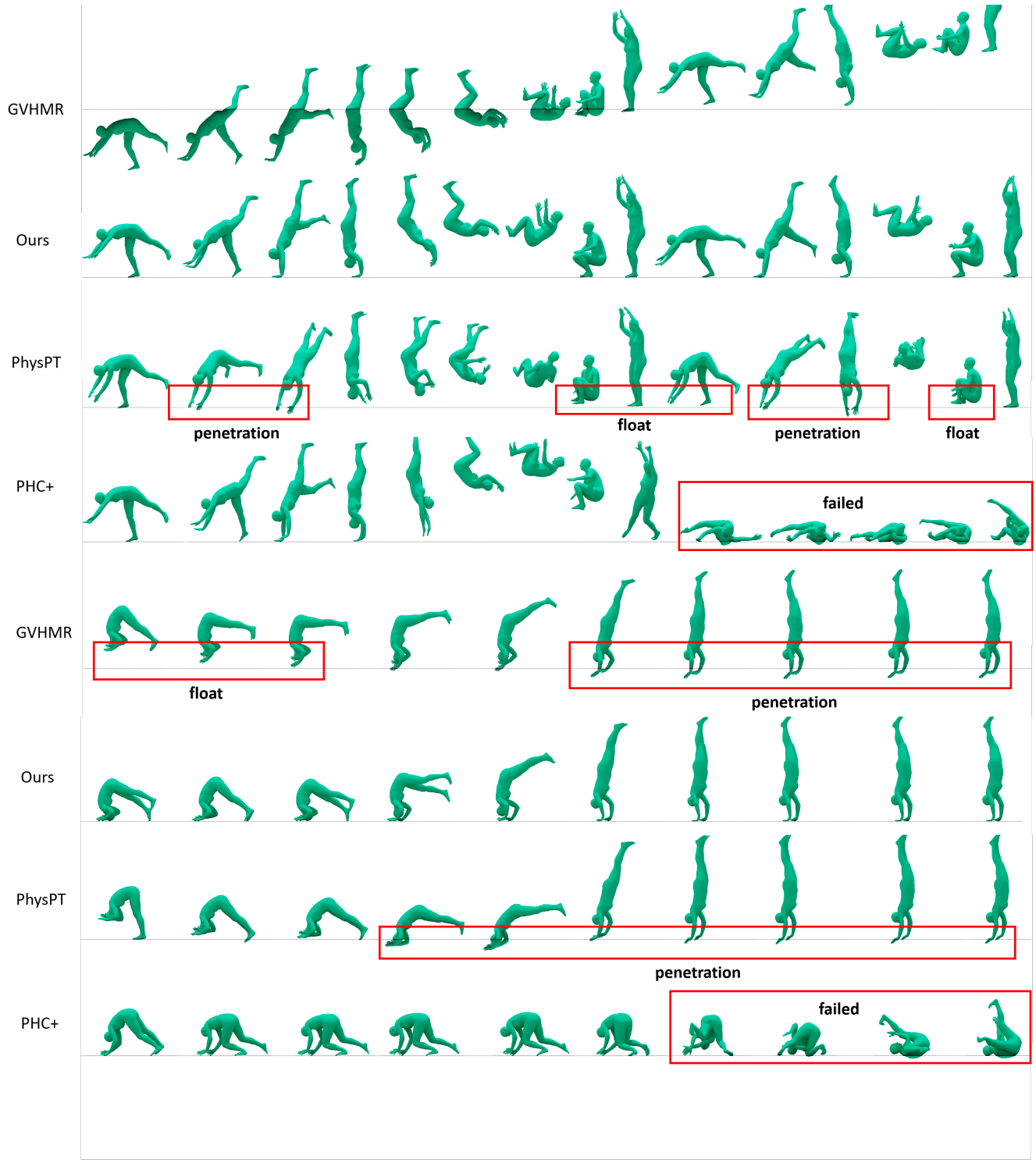

Figure 6. **Additional comparison with SOTA techniques.**

erative modeling, and recently, diffusion probabilistic models have been introduced into 3D human motion generation, framing motion generation as a sequence generation problem. Motion inbetween is a subtask of 3D human motion generation. Unlike regular motion generation, inbetween does not directly reconstruct the entire motion from noise

but provides a partial motion as a condition for the diffusion model. The goal of inbetween is to generate the missing part of the motion. Typically, motion is generated under certain conditions (such as text, music, or images), and these conditions can also be used for the inbetween task. In this work, we use the human body mask obtained from segmentation as a condition (essentially a simplified version of the image, leveraging the powerful priors of current segmentation models to alleviate the burden on the inbetween model). Guided by the human mask $\boldsymbol{m}$, the diffusion model uses a denoising network to learn the process of denoising Gaussian noise.

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{m}) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{m}), \Sigma_t) \quad (1)$$

where θ demonstrates the parameters of the denoising network and the untrained time-dependent covariance is set according to the variance schedule. Following current motion generation methods, we use sample-estimation reparameterization and directly predict the clean sample estimate from noise, rather than the mean estimate. The training goal of the diffusion model can be defined as:

$$\mathcal{L} := \mathbb{E}_{(\boldsymbol{x}_0, \boldsymbol{m})\sim q(\boldsymbol{x}_0, \boldsymbol{m}), t\sim[1,T]} \left[\|\boldsymbol{x}_0 - G_\theta(\boldsymbol{x}_t, t, \boldsymbol{m})\|^2\right]. \quad (2)$$

Auxiliary Losses In addition to the basic reconstruction loss, we introduce several auxiliary losses to enhance training stability. Since we adopt a rotation-based motion representation, we need to recover the global position of the human body before computing the other losses. We incorporate a loss on joint positions.

$$\mathcal{L}_{\text{joint}} = \frac{1}{J}\sum_{i=1}^{J} \|FK(\boldsymbol{x}^{(i)}) - FK(\hat{\boldsymbol{x}}^{(i)})\|_2^2 \quad (3)$$

where FK means the forward process for motion representation to get the absolute 3D position $\boldsymbol{p}$. $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{p}}$ is obtained from the model prediction. To enhance the smoothness of the generated motion, we introduce a velocity loss and an acceleration loss.

$$\mathcal{L}_{\text{vel}} = \|\boldsymbol{p}_{\text{vel}} - \hat{\boldsymbol{p}}_{\text{vel}}\|_2^2, \mathcal{L}_{\text{acc}} = \|\boldsymbol{p}_{\text{acc}} - \hat{\boldsymbol{p}}_{\text{acc}}\|_2^2, \quad (4)$$

Additionally, to improve the accuracy of root position generation, we introduce extra constraints on both the root translation and orientation.

$$\mathcal{L}_{\text{ro}} = \|\boldsymbol{x}^{ro} - \hat{\boldsymbol{x}}^{ro}\|_2^2 + \|\boldsymbol{x}^{ro}_{\text{vel}} - \hat{\boldsymbol{x}}^{ro}_{\text{ro}}\|_2^2 + \|\boldsymbol{x}^{ro}_{\text{acc}} - \hat{\boldsymbol{x}}^{ro}_{\text{acc}}\|_2^2, \quad (5)$$

where $\boldsymbol{x}^{ro}$ is 7-dim orientation and translation of motion $\boldsymbol{x}$.

Training process. A random motion segment, selected at a random sequence position, is chosen as the generation target. Our model is trained to reconstruct this segment. Both keyframe conditioning signals $\boldsymbol{c}$ and mask conditioning signal $\boldsymbol{m}$ are set to $\emptyset$ for 10% of training data to make our model suited for unconditioned motion generation.

Mismatch Detection. In practice, we calculate the IoU of 2D human mask and the 2D projection of SMPL mesh for mismatch detection. For our high-difficulty datasets, the threshold is 0.5 with greater tolerance. For general datasets like AIST++ and kungfu, we take 0.7 as the threshold.

4. Motion Prediction Prior

Following Multi-Transmotion [1], we designed a motion prediction model based on transformer architecture, which was trained on high-quality motion datasets and then participated as a motion prior in our reinforcement learning training process. The inputs of the model are human key point positions, joint point velocities, rotation angles, rotation angular velocities, and root trajectories, and the outputs of the model are future human poses and root trajectories. Specifically, we use a modality-specific multilayer perceptron layer to implement the tokenization of each input, and the resulting token will be fed into the Transformer Encoder along with the initialized future motion token to predict the future token. Finally, the predicted token is fed into the trajectory decoder and pose decoder to get the final output.

5. Implementation Details

Training Our method consists of two modules, MCM and PTM, which are trained separately. For the Mask-conditioned Motion Correction Module, we select the ground truth data from the Human3.6M [2], AIST++ [4, 12], and Motion-X [5] Kungfu datasets as the training set. Additionally, we use SAM [3] to process these datasets and generate the corresponding mask data. The mask data serves as a condition to guide the motion correction process. For the Physics-based Motion Transfer Module, we use the AMASS [9], Human3.6M, AIST++, and Kungfu datasets for training. Our method is not limited to a fixing short-motion clip, on the contrary, our approach supports long-duration motion repair. The repair capability depends on the quality of the reference motion and the complexity of the action, rather than the length of the motion sequence.

Imitation Our simulation environment is NVIDIA's Isaac Gym [10], where the humanoid model is adapted to a format compatible with SMPL, as it natively supports various body shapes and is widely used in pose estimation research. As a result, our restoration outputs can be easily converted to the SMPL format and rendered into realistic meshes. Currently, we do not consider complex SMPL body shapes in our approach. Our discriminator is an MLP with two hidden layers with ReLU activation function. The policy of our PTM is constructed with 6-layer multilayer perceptions and ReLU functions.

Adaptation Time. We tested the time consumed by adaptation on both the difficult and general datasets. For the difficult data, our method needs to execute an average

of 2.5k steps for about 30 minutes to repair a motion with a length of about 15s. For the general dataset, our method needs to execute an average of 500 steps for about 8 minutes. All of the above experiments were performed on an NVIDIA 4090.

Metrics Details In this section, we discuss the evaluation benchmarks we proposed from two aspects: physical realism and 2D similarity. For physical realism, we use the following metrics:

- Self-Penetration (SP): SP means some surface vertices of the SMPL mesh (rendered from the human motion representation) appear inside the mesh, such as fingers passing through the torso. We calculate the proportion of vertices that penetrate the body mesh in each frame and average this ratio over time as the self-penetration metric.
- Ground-Penetration (GP): GP refers to vertices of the SMPL mesh that appear below the ground. We compute the distance between the ground and the lowest body mesh vertex that is below the ground.
- Float: Float happens when all body mesh vertices are above the ground. Following [13], we calculate the distance between the ground and the lowest body mesh vertex above the ground. Both GP and Float use a 5mm threshold, with values below this threshold being ignored.
- Foot-Skate (FS): We identify the foot joints in two adjacent frames that are in contact with the ground, and then calculate their average horizontal displacement across frames.

For 2D similarity, we use the following metrics:

- 2D Keypoint OKS: We first compute the 3D absolute positions from the motion representation, then project these 3D coordinates onto the 2D image plane. The 2D Keypoint OKS is then calculated as the similarity between the 3D projections and the 2D annotations. To ensure compatibility with various 2D/3D keypoint representations, the OKS calculation uses 12 keypoints: left/right hip, knee, ankle, shoulder, elbow, and wrist. The 2D annotations is obtained from multiple advanced keypoints detection methods and we corrected them manually.
- Mask-Pose Similarity (MPS): To provide a more granular description of the motion and take into account the human shape, we introduce MPS to calculate the similarity between the 2D projection of the human mesh and the segmentation mask of the human body in the image. MPS is calculated by determining the ratio of 3D mesh vertices within the human segmentation mask. A larger ratio indicates higher 2D similarity and more accurate human shape estimation.

6. User Study

To obtain subjective evaluations of different methods, we recruited 44 participants to watch 30 pairs of videos, with one video produced by our method and the other by GVHMR, PhysPT, or PHC+. For each pair of videos, we asked participants to answer three questions: *which video was of better overall quality? Which video has better physical realism? Which video is more restored to the original video?* Thus we ended up with 3960 comparisons. For

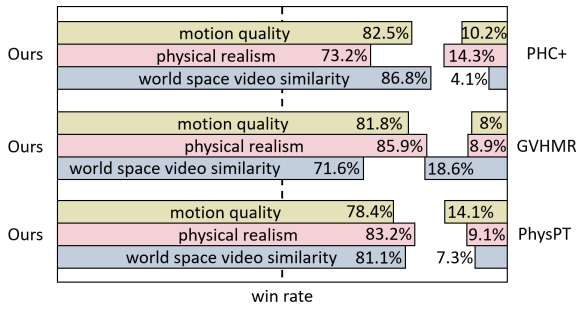

Figure 7. **The physical restoration user study of high-difficulty and in-the-wild motions.**

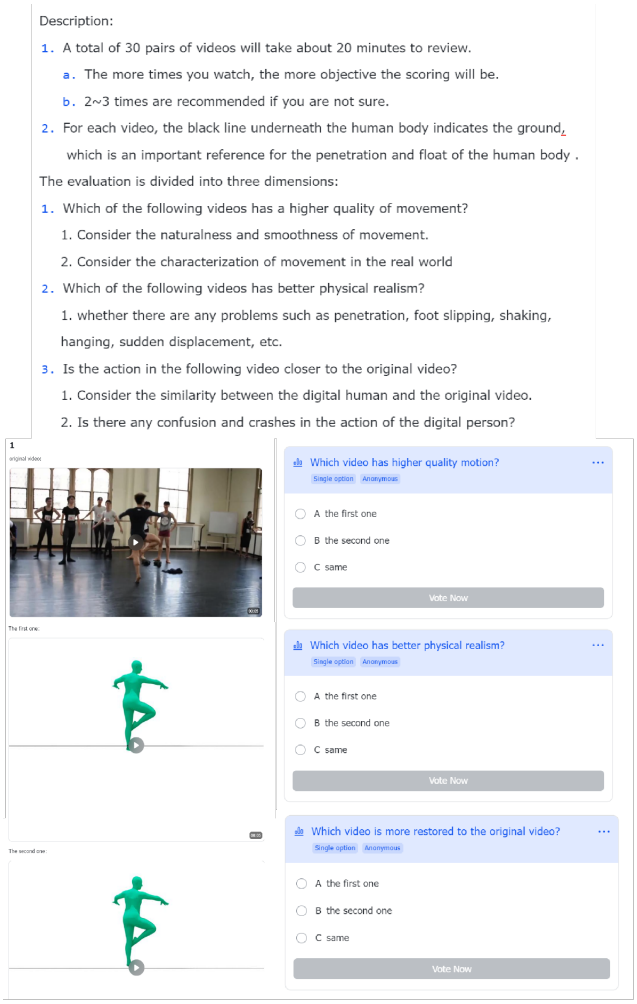

Figure 8. **Screenshot of user study.**

each comparison, we asked participants to choose between three options, *the first being better, the second being better, or both being about the same*. The evaluation results are shown in Figure 7, our method outperforms all other methods in more than 70%, and only in less than 20% of the cases are we defeated by other methods.

References

[1] Yang Gao, Po-Chien Luan, and Alexandre Alahi. Multi-transmotion: Pre-trained model for human motion prediction. *arXiv preprint arXiv:2411.02673*, 2024. 7

[2] 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. 7

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 7

[4] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 7

[5] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 2023. 7

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[7] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 1, 3

[8] 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. 3

[9] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 7

[10] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 7

[11] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024. 1

[12] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, 2019. 7

[13] 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*, pages 16010–16021, 2023. 8

[14] Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2305–2317, 2024. 3