

POMP: Physics-consistent Human Motion Prior through Phase Manifolds

Supplementary Material

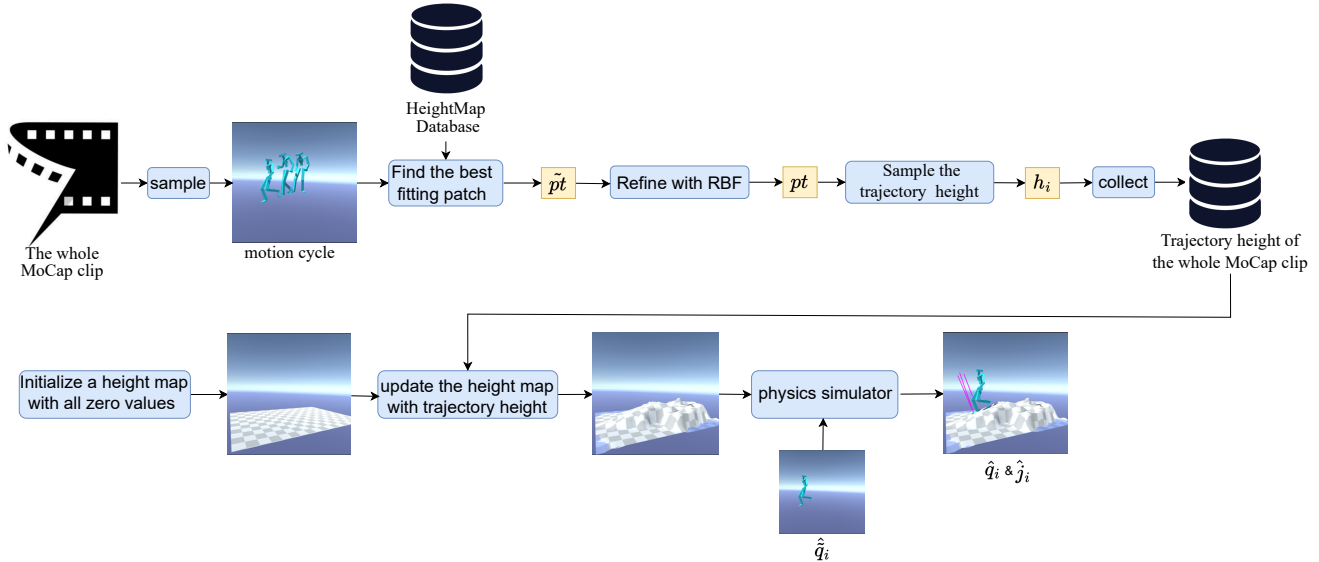


Figure 1. Detailed pipeline for terrain fitting and supplement of dynamic information(e.g. simulated poses and contact impulses)

In the main paper, we present POMP, a novel kinematics-based framework that synthesizes physically consistent motions by leveraging phase manifolds to align motion priors with physics constraints. This supplementary material provides additional details of our proposed approach: (1). Data Pre-processing; (2). Details of Character Modeling; (3). Network Architecture of Kinematic Module; (4). More results; (5). Discussion . For a complete demonstration of our model’s capabilities, we direct readers to the supplementary video.

1. Data Pre-processing

To accurately simulate realistic human-environment interactions, the data preparation involves two key steps, as illustrated in Fig. 1. First, we generate terrain geometry compatible with existing motion capture data. Second, we gather dynamic information, including simulated poses and full-body contact impulses, within a Unity virtual scene.

To reconstruct the topography of the interactive scene, we begin by employing the terrain fitting technique described in the PFNN [2]. The fitting procedure consists of two phases. Initially, the motion capture sequence is divided into individual motion cycles, with each cycle defined as the interval between consecutive right foot liftoffs. We then identify the optimal terrain segment by minimizing an error function that takes into account factors such as

foot-ground contact, airborne foot movement and jumping. Subsequently, a Radial Basis Function (RBF) mesh manipulation method is applied to refine the terrain, ensuring precise foot placement during contact periods. Once the terrain is fitted for each motion cycle, the height is sampled along the trajectory at three points—left, right, and center—each spaced 25 cm apart. These sampled heights are then combined to create the final height profile for the entire MoCap sequence. Finally, within the Unity scene, we initialize a height map with zero values and use the height profile to gradually “scan” and update the height map, resulting in the generation of the final terrain data.

Once the terrain is created, MoCap data is imported into the scene, and a physics simulator is employed to automatically collect ground truth contact impulses, denoted as \hat{j}_i , as well as to generate the corresponding ground truth simulated poses, represented as \hat{q}_i .

2. Details of Character Modeling

Fig. 2 illustrates our articulated rigid-body system, comprising 15 box or capsule limb colliders. This structure envelops the character mesh, facilitating comprehensive collision detection and full-body impulse collection during character-environment interactions. The system employs Unity’s configurable joints for inter-collider connections, which serve dual purposes: passive response to external

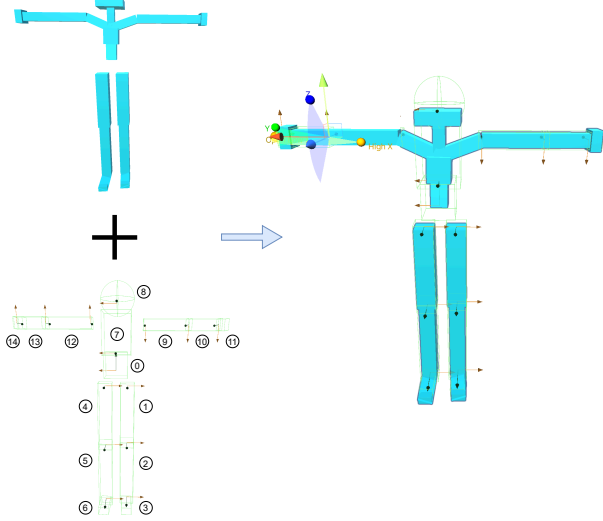


Figure 2. **Character modeling.** The articulated rigid-body system is composed of 15 limb colliders, each interconnected through configurable joints. These joints impose constraints on the degrees of freedom (DoF) of each limb, while simultaneously applying driving forces to realize movement.

forces and active force application through simulated joint motors. Additionally, we can constrain the joints’ degrees of freedom (DoF) to enhance the realism of the articulated body’s movements, making them more human-like.

3. Network Architecture of Kinematic Module

The kinematic module consists of an ortho-MoE-based encoder and a diffusion-based decoder. As illustrated in Fig. 3, we delve into the detailed architectures of ortho-MoE-based encoder and the noise prediction network in the diffusion reverse step.

Ortho-MoE-based encoder. The encoder consists of two components: a 3-layer Ortho-MoE network E_m and a gating network E_g . E_m implements a Mixture of Experts (MoE) architecture with eight parallel MLP branches, incorporating layer normalization (LN) and Exponential Linear Unit (ELU) activation layers. On the other hand, E_g calculates the blending weights for these experts using multi-layer perceptron (MLP) layers with ELU and softmax activation layers. Additionally, to achieve disentanglement of the principle phase components, we further enforce orthogonality between every pair of principle components in the final MoE layer by imposing the constraint $\langle w_i, w_j \rangle = 0$ for $i \neq j$, where $\langle \cdot, \cdot \rangle$ signifies the dot product.

Noise prediction network in the diffusion reverse step.

According to Eq. (4) in the main paper, we need to build a noise prediction network for the diffusion reverse step. We

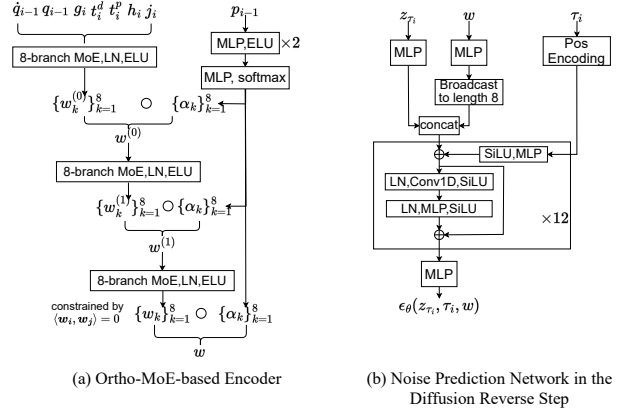


Figure 3. **Network architecture of kinematic module.** Our articulated rigid-body system is composed of 15 limb colliders, which are connected by configurable joints. These configurable joints apply drive forces and constrain the DoF of each joint.

start by using MLP layers to project z_{τ_i} and w into separate high-dimensional feature spaces. Subsequently, we expand the feature of w to a length of 8, enabling the feature concatenation of z_{τ_i} and w . This ensures that each expert component of z_{τ_i} can effectively incorporate phase motion features during the denoising process in the phase domain. To allow the following residual blocks to utilize information about the iteration order, we transform τ_n into positional embeddings. Each residual block includes a 1D convolutional layer and an MLP layer, which extract serial and spatial features, respectively. To prevent the vanishing gradient problem, we also incorporate skip connections [1].

4. More Results

In this section, we present additional visualization results showcasing POMP’s responses to dynamic changes across various motion types.

Bump reactions. Fig. 4 illustrates POMP’s varied responses to different impact levels (slight/moderate/heavy). The synergy between the kinematics and dynamics modules enables POMP to react both actively and passively to forces of varying magnitudes and directions. When an impact is substantial enough to displace the center of mass, the kinematics module actively generates a new target pose, producing a compensatory driving force to restore balance. Conversely, minor impacts do not alter the kinematics module’s target pose, allowing affected limbs to gradually resume their original motion under the influence of the initial driving force. Additional details are available in the supplementary video.

Target tracking. POMP performs target tracking through a two-step process. Initially, it defines the target position for

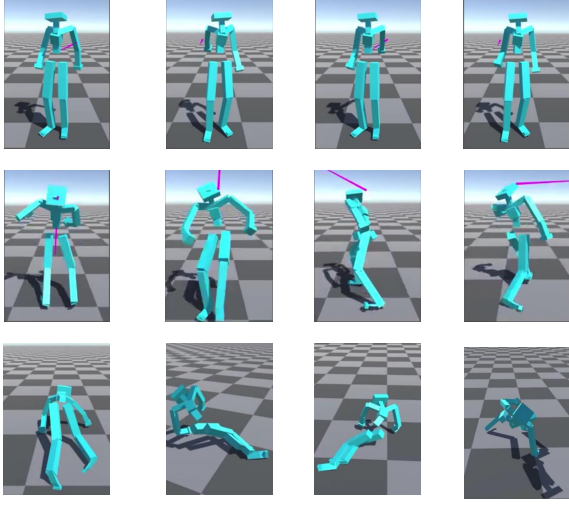


Figure 4. **Bump reaction.** POMP can perform both active and passive responses to impacts of varying magnitudes and directions.

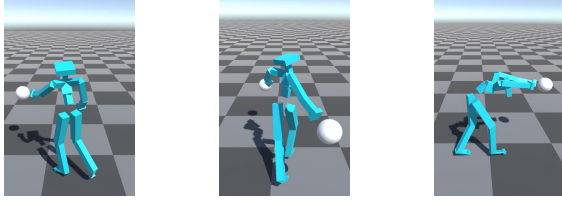


Figure 5. **Target tracking.** POMP can handle various target tracking scenarios, including tracking a single target with one end effector, two targets with two effectors, or a single target with both effectors.

Models	train success \uparrow	test success \uparrow
POMP (human-terrain)	100%	94.3%
POMP-wo-PEM	100%	71.6%

Table 1. Success rate on the human-terrain interaction dataset.

an end effector (e.g., left or right hand). Then, it employs inverse dynamics to compute the external force f required to move the joint to the desired location. This force f is subsequently applied to the target joint, enabling forward dynamics. POMP supports various target tracking scenarios as illustrated in Fig. 5. These include tracking a single target with one end effector, two targets with two end effectors, or a single target with two effectors. For a more comprehensive demonstration, please refer to the supplementary video.

Human-terrain Interactions. The results presented in Figs. 6 to 8 demonstrate POMP’s capability to generate physically plausible locomotion patterns across diverse

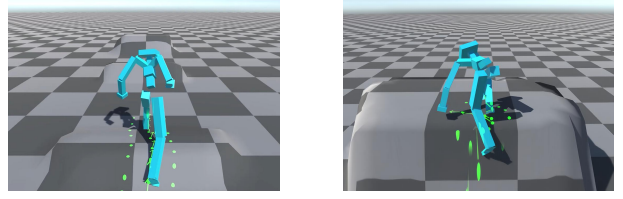


Figure 6. **Obstacle crossing.** POMP can manage obstacle crossing of all types, from low barriers to high walls.

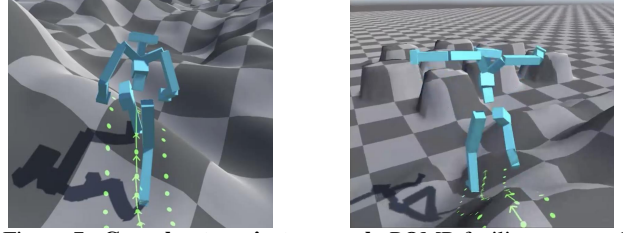


Figure 7. **Complex terrain traversal.** POMP facilitates smooth and continuous motion transitions over diverse terrains.

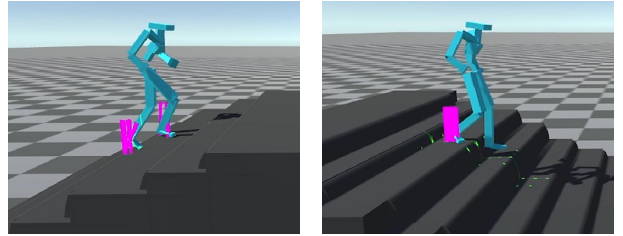


Figure 8. **Stair ascent and descent.** POMP effectively addresses the issue of model penetration, a common challenge encountered during stair ascent and descent.

challenging terrains, including obstacle crossing, complex terrain traversal and stair ascent and descent. Additionally, as demonstrated in Fig. 9, the kinematic module effectively learns diverse motion priors, enabling POMP to generalize across a wide range of motion patterns over a large-scale dataset. We further conduct the ablation study on the phase encoding module (PEM), as detailed in Tab. 1. The quantitative analysis reveals that the PEM significantly enhances POMP’s generalization performance, particularly in complex human-scene interaction scenarios.

Comparative performance. In the supplementary video, we further provide qualitative comparisons between POMP and other methods: the kinematic-based PFNN, and the physics-based PhysicsVAE, DROP, and MaskedMimic [3–5]. Extensive evaluations demonstrate the effectiveness of POMP across various contents, terrains and physical interactions.

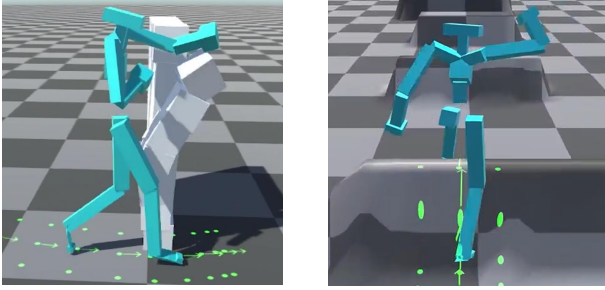


Figure 9. **Diverse motion priors.** Due to the diverse motion priors learned by the kinematic module, POMP is capable of actively generating a wide range of realistic motions across various interactive scenarios

5. Discussion

Limitation. Despite the significant progress achieved by our current work, POMP, several limitations remain, offering directions for future improvement. First, the model lacks a diverse range of dynamic features. While it successfully captures full-body contact impulses, other critical dynamic elements, such as joint torques and reaction forces, are not yet collected. Incorporating these additional features could provide deeper insights into the underlying mechanisms of human motion production, thereby enabling kinematic-based models to generate more realistic movements during complex interactions. Second, POMP does not currently integrate task-based motion controllers. The present approach primarily aims to bridge the gap between kinematic motion priors and physical constraints, demonstrating its potential to generalize across various motion patterns. Future research will focus on developing a multimodal motion controller that can adapt character movements based on task requirements, current states, and terrain topology. These limitations highlight areas for further investigation, which could enhance both the realism and versatility of POMP.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [2](#)
- [2] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. [1](#)
- [3] Yifeng Jiang, Jungdam Won, Yuting Ye, and C Karen Liu. Drop: Dynamics responses from human motion prior and projective dynamics. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. [3](#)
- [4] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *arXiv preprint arXiv:2409.14393*, 2024.

- [5] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Trans. Graph.*, 41(4), 2022. [3](#)