

---

# CoDA: Coordinated Diffusion Noise Optimization for Whole-Body Manipulation of Articulated Objects

## Supplementary Material

---

Huaijin Pi<sup>1</sup>   Zhi Cen<sup>2</sup>   Zhiyang Dou<sup>1</sup>   Taku Komura<sup>1</sup>

<sup>1</sup>The University of Hong Kong <sup>2</sup>Zhejiang University  
<https://phj128.github.io/page/CoDA>

<b>A</b>	<b>Introduction</b>	<b>2</b>
<b>B</b>	<b>Additional Related Work</b>	<b>2</b>
<b>C</b>	<b>Method Details</b>	<b>2</b>
<b>D</b>	<b>Implementation Details</b>	<b>3</b>
	D.1 Training details . . . . .	3
	D.2 Optimization details . . . . .	4
	D.3 User study details . . . . .	4
	D.4 More ablation study and details . . . . .	4
<b>E</b>	<b>More Analysis</b>	<b>6</b>
	E.1 Generalization to different object geometry . . . . .	6
	E.2 Object motion control . . . . .	6
	E.3 Simultaneous locomotion and manipulation . . . . .	6
	E.4 Text-based control . . . . .	8
	E.5 Diverse results . . . . .	9
	E.6 Deployment on simulated humanoids . . . . .	9
	E.7 Generating whole-body motion from hand-only dataset . . . . .	9
	E.8 Inference speed . . . . .	10
	E.9 Visualization of optimization process . . . . .	10
<b>F</b>	<b>More Discussions</b>	<b>10</b>
	F.1 Limitations . . . . .	10
	F.2 Broader impact . . . . .	11

## A Introduction

This supplementary document provides additional details and analysis of our proposed method. In Section B, we provide a more comprehensive review of related work. Section C includes more details about our method. In Section D, we describe implementation details and experimental settings, and provide further ablation studies to validate our design choices. Section E provides deeper analysis of the generated motions and showcases various capabilities enabled by our method. Finally, Section F discusses more limitations and broader impacts of this work.

We provide a supplementary video that demonstrates the capabilities of our method, comparisons with baseline approaches, and visualizations from ablation studies. In addition, we offer a supplementary webpage that presents high-resolution visualizations of all generated motion sequences.

## B Additional Related Work

**Motion generation.** Human motion generation is a long-standing research problem [21, 74, 57–59, 13, 60, 6, 65, 83]. Recent approaches adopt a wide range of neural architectures, including Mixture of Experts (MoE) [22, 74], recurrent neural network [42], transformer [62, 15, 50], and Mamba [12, 81]. To enhance motion diversity and realism, various generative paradigms have been explored, such as generative adversarial networks [11, 31], normalizing flow [27, 17], variational auto-encoder [26, 50, 16], VQ-VAE [53, 77, 23, 80, 36], diffusion models [19, 60, 78, 7, 70, 69, 1] and mask modeling [3, 14]. With the availability of the large-scale datasets [51, 40, 32, 13, 33], motion generation has been conditioned on diverse modalities such as text [77, 14, 5], music [61], and audio [4]. In parallel, physics-based methods [47, 48, 37, 54] have enabled simulated humanoids [41] to perform various motor skills [73, 84, 66] through reinforcement learning. Several recent works [49, 68, 71, 9, 72, 39] learn latent representations of human motion that support skill reuse.

**Diffusion noise optimization.** Diffusion models [19] have shown great success in various generative tasks [20]. To better control the generation process, classifier guidance [8] and classifier-free guidance [18] have been proposed. SDEdit [43] enables image editing by injecting noise into the reverse stochastic differential equation (SDE) process. DOODL [63] introduces an optimization-based approach that directly updates the input noise of diffusion models by leveraging the invertible ordinary differential equation (ODE) [64]. This framework has been extended to other domains such as music generation [45]. In the motion domain, DNO [24] applies this idea for body-only motion editing, while ProgMoGen [34] uses LLMs [46] to select constraints and performs noise-space optimization for open-set motion control. However, these works are limited to body motion. In contrast, our work tackles the more complex setting of whole-body manipulation of articulated objects. We perform coordinated optimization over three diffusion models specialized for the body, left hand, and right hand, enabling coherent and physically plausible motion across the whole body.

## C Method Details

We adopt a RoPE-based encoding scheme for object motion, inspired by 6-DoF CaPE [28], to effectively capture the temporal dynamics of objects.

Given the object pose at  $i$ -th frame, we could use  $P_i$  to denote its  $4 \times 4$  transformation matrix:

$$P_i = \begin{bmatrix} R_i & t_i \\ \mathbf{0} & 1 \end{bmatrix}, \quad (1)$$

where  $R_i$  is the rotation matrix and  $t_i$  is the translation vector. The relative position embedding function  $\pi(\mathbf{v}, \mathbf{P})$  should satisfy the following properties:

$$\langle \pi(\mathbf{v}_1, \mathbf{P}_1), \pi(\mathbf{v}_2, \mathbf{P}_2) \rangle = \langle \pi(\mathbf{v}_1, \mathbf{P}_2^{-1} \mathbf{P}_1), \pi(\mathbf{v}_2, \mathbf{I}) \rangle. \quad (2)$$

where  $\mathbf{v}$  is the position embedding. Under this constraint, the attention between two transformed features can be equivalently rewritten as:

$$(\phi(\mathbf{P}_2^{-1} \mathbf{P}_1) \mathbf{v}_1)^\top (\phi(\mathbf{I}) \mathbf{v}_2) = (\mathbf{v}_1^\top \phi(\mathbf{P}_1^\top \mathbf{P}_2^{-\top})) \mathbf{v}_2 \quad (3)$$

$$= (\mathbf{v}_1^\top \phi(\mathbf{P}_1^\top)) (\phi(\mathbf{P}_2^{-\top}) \mathbf{v}_2) = (\phi(\mathbf{P}_1) \mathbf{v}_1)^\top (\phi(\mathbf{P}_2^{-\top}) \mathbf{v}_2) \quad (4)$$

$$= \langle \pi(\mathbf{v}_1, \phi(\mathbf{P}_1)), \pi(\mathbf{v}_2, \phi(\mathbf{P}_2^{-\top})) \rangle. \quad (5)$$



Therefore, the relative embedding function can be implemented as  $\pi(\mathbf{v}, \mathbf{P}) = \phi(\mathbf{P})\mathbf{v}$ , where the transformation  $\phi(\mathbf{P})$  is defined as:

$$\phi(\mathbf{P}) = \begin{bmatrix} \Psi & 0 & \cdots & 0 \\ 0 & \Psi & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \Psi \end{bmatrix}, \Psi = \begin{cases} \mathbf{P} & \text{if key} \\ \mathbf{P}^{-\top} & \text{if query} \end{cases} \quad (6)$$

Here, the embedding dimension is assumed to be divisible by 4, and  $\Psi$  is either  $\mathbf{P}$  or  $\mathbf{P}^{-\top}$  depending on whether the input is a key or query.

This formulation allows each frame’s object pose to attend to others within a temporal window using relative transformations, enabling more expressive modeling of object trajectories compared to simple velocity inputs. Following [55], we restrict the attention window to 120 neighboring frames during training and inference.

## D Implementation Details

### D.1 Training details

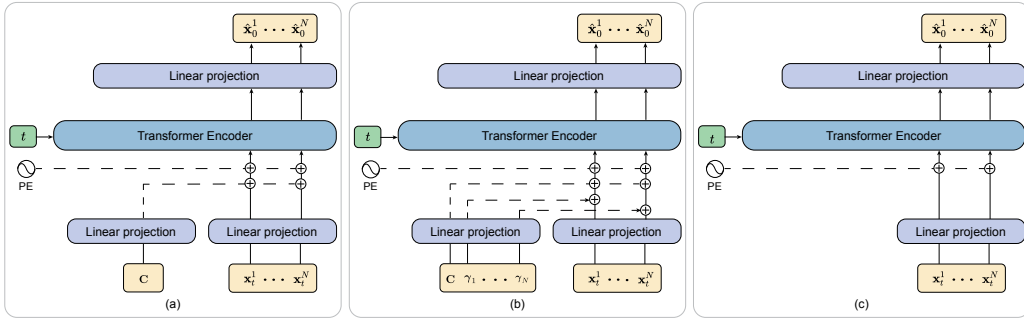


Figure 1: **Model architecture.** (a) The object motion diffusion model. (b) The end-effector trajectories diffusion model. (c) The body motion and hand motion diffusion models.

**Model architectures.** We adopt a transformer-based diffusion architecture [62, 19], similar to MDM [60], for all models in our framework, as illustrated in Figure 1. The object motion diffusion model is shown in Figure 1 (a). The condition  $\mathbf{C}$  includes the initial object pose, text CLIP [52] feature, and object BPS [79] at idle pose.  $t$  denotes the diffusion timestep. They will be processed by a linear projection layer to the embedding space and added to the noisy motion embedding. The end-effector trajectories diffusion model is shown in Figure 1 (b). The difference here is the frame-wise condition  $\gamma$ , which includes object geometry embedding and object global trajectories, would be added frame-wise to the noisy end-effector embedding. We also use a RoPE-based representation in this transformer to capture relative object trajectories of each frame. The body motion and hand motion diffusion models are shown in Figure 1 (c), which is an unconditional diffusion model, covers the motion manifolds. All diffusion models are 8 transformer encoder layers, with 4 attention heads and 512 hidden units, and the feed-forward layer has 1024 hidden units.

**Network training.** For object motion and end-effector trajectories diffusion models, we train them for 500 epochs with batchsize 32 and AdamW [35] optimizer, with a learning rate  $1 \times 10^{-4}$ . For body motion and hand motion diffusion models, we train them for 2000 epochs with batchsize 128 and AdamW optimizer, with a learning rate  $1 \times 10^{-4}$ . All diffusion models use 1000 sampling steps during training [19], with a variance schedule increasing from  $\beta_1 = 0.0001$  to  $\beta_T = 0.02$  using the cosine schedule [44]. At inference time, we accelerate sampling for object motion and fingertip distance generation using DDIM [56] with  $T = 100$ . All our experiments are conducted on a single NVIDIA A100 GPU.

## D.2 Optimization details

**End-effector trajectories generation.** Given the generated end-effector BPS, we use the Adam [25] optimizer with a cosine-decayed learning rate 0.05 for 800 steps to calculate the end-effector trajectories.

**Whole-body motion generation.** During inference, we perform noise optimization using DDIM [56] with  $T=10$  for 800 steps and a cosine-decayed learning rate 0.05, following the DNO [24] strategy. The optimization loss with different weights  $\lambda_{ee}$ ,  $\lambda_{pen}$ , and  $\lambda_{reg}$  is as follows:

$$\mathcal{L} = \lambda_{ee}\mathcal{L}_{ee} + \lambda_{pen}\mathcal{L}_{pen} + \lambda_{reg}\mathcal{L}_{reg}, \quad (7)$$

where  $\mathcal{L}_{ee}$ ,  $\mathcal{L}_{pen}$ , and  $\mathcal{L}_{reg}$  are the end-effector tracking, penetration, and regularization losses. For first 300 steps, we set fingertip part as zero and only use the wrist targets, as the body motion might be very different from the generated end-effector trajectories, with  $\lambda_{ee} = 1$ ,  $\lambda_{pen} = 0$ ,  $\lambda_{reg} = 0$ . For 300 to 500 steps, we optimize with all end-effectors, and set  $\lambda_{ee} = 1$ ,  $\lambda_{pen} = 0$ ,  $\lambda_{reg} = 0$ . For 500 to 800 steps, we enable the penetration loss and regularization loss, and set  $\lambda_{ee} = 1$ ,  $\lambda_{pen} = 5.0$ ,  $\lambda_{reg} = 1.0$ .

## D.3 User study details

Figure 2 shows the user study webpage. We conduct a user study using 22 sequences, covering 11 articulated objects from the ARCTIC dataset, with two motion sequences per object. A total of 16 participants completed the study anonymously. For each sequence, participants were asked to rank the generated motions from different methods based on motion quality and physical plausibility.

## D.4 More ablation study and details

**Ablation study implementation details.** We present more details about the implementation of each variant in the ablation study. (a) A single model to jointly predict object motion and end-effector trajectories. In our proposed pipeline, we utilize different models for object motion and end-effector trajectories. Instead, we train a variant where a single model predicts both object motion and end-effector trajectories. The difference is that in our pipeline, the input of the end-effector trajectory model includes the frame-wise object geometry embedding, as the object geometry of articulated objects is dynamic and changes over time. (b) Predicting the relative coordinate of end-effectors to the object center without end-effector BPS. In this variant, we directly predict the coordinate of end-effectors in the object coordinate system. This variant validates our design of using end-effector BPS to represent the end-effector trajectories. (c) Using object velocity and rotational velocity as the trajectory input without RoPE-based representation. This variant is to validate the effectiveness of the RoPE-based representation of object trajectories, which could capture the relative object trajectories in a local temporal window of each frame. (d) Removing the optimization process and using a conditional diffusion model with fingertip trajectories as input. This variant is to validate the effectiveness of the optimization process, which could generate more realistic whole-body motion. (e) Using a single diffusion model for the entire body without the decoupled body-hand representation. This variant is to validate the effectiveness of the decoupled representation. (f) Excluding the AMASS [40] dataset during training the body motion model. This variant validates the effectiveness of including more body data (the AMASS [40] dataset) for training the body motion model. (g) Replacing the end-effector representation with a distance field [75]. This variant validates the effect of different distance-based representations. Both encode trajectories as distances, but the distance field uses a fixed voxel grid instead of a basis point set. (h) Conditioning the hand motion diffusion model on object trajectories. This variant validates the impact of training a conditional model instead of a pure prior, examining whether explicit conditioning on object motion affects coordination quality.

**More ablation study.** We present more ablation study results in Table 1. (i) w/ VAE. We replace the diffusion model with a VAE and optimize the VAE latent space instead. This variant is to validate the effectiveness of the diffusion noise space. (j) w/ Guidance. We replace noise optimization with a guidance strategy in the diffusion process. This variant is to validate the effectiveness of the optimization strategy.

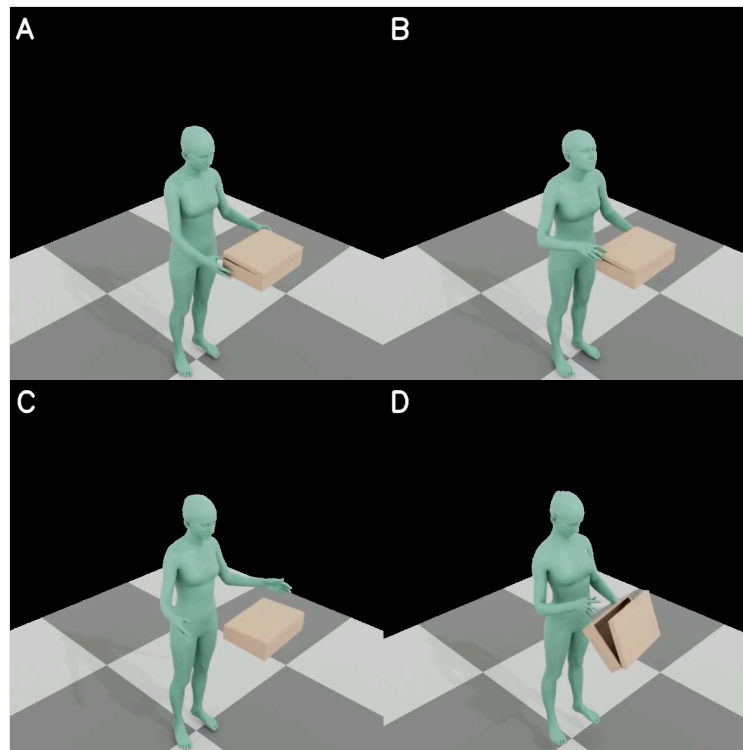
## User Study Instructions

For each question, you will see one video containing four motion clips (A, B, C, D). Please rank them based on:

- **Motion Quality:** smoothness, realism, style consistency
- **Physical Plausibility:** balance, force realism, no impossible motions

You will assign a **Best**, **Second Best**, and **Third Best** for each aspect.

### Question 1 - A person uses the box.



#### ◆ Motion Quality

The best:  
 ▼  
Second best:  
 ▼  
Third best:  
 ▼

#### ◆ Physical Plausibility

The best:  
 ▼  
Second best:  
 ▼  
Third best:  
 ▼

Figure 2: **User study interface.** The webpage of our user study, where participants are asked to rank the generated motions based on motion quality and physical plausibility.

Table 1: **More ablation study on the ARCTIC [10] dataset.** The right arrow  $\rightarrow$  means the closer to real motion the better. The best and second-best results are highlighted green and yellow, respectively.

Methods	FID↓	R-Precision↑	Diversity→	Foot skating↓	IV↓	ID↓	CR↑
Real	—	0.531	8.664	0.002	4.68	11.47	0.085
(i) w/ VAE	5.082	0.430	6.717	0.032	6.28	12.79	0.090
(j) w/ Guidance	11.737	0.242	4.506	0.775	16.88	16.36	0.009
Ours	2.283	0.477	7.208	0.002	5.25	12.87	0.086

Table 2: **Ablation study on the optimization steps.**

Steps	FID↓	R-Precision↑	Diversity→	Foot skating↓	IV↓	ID↓	CR↑
Real	—	0.531	8.664	0.002	4.68	11.47	0.085
400	3.750	0.453	7.025	0.002	5.46	12.38	0.086
1200	2.264	0.445	7.305	0.001	5.42	13.22	0.087
Ours (800)	2.283	0.477	7.208	0.002	5.25	12.87	0.086

We further investigate the effect of the optimization steps on the performance. As shown in Table 2, with more optimization steps, the performance is improved. But 800 steps obtains similar performance to 1200 steps. So we set 800 steps as the default optimization steps.

We also analyze the effect of the BPS number on the performance. Previous work [30, 29] use 1024 points for the BPS. As shown in Table 3, increasing BPS number obtains similar performance in our setting.

## E More Analysis

### E.1 Generalization to different object geometry

To validate the generalization to different object geometries, we train the object motion and end-effector trajectory models on the hand-only dataset [82]. Following previous work [82, 76, 79], we use 7 training and 3 testing objects. Thanks to our multi-stage design, the optimization only relies on the object motion and end-effector trajectories. Therefore, our method could generate whole-body motion, as shown in Figure 3.

### E.2 Object motion control

To enable controllable object motion, we apply diffusion noise optimization to the object motion generation model by specifying keyframe object poses as targets. As shown in Figure 4, our method successfully generates manipulation sequences where the object is guided to reach the desired poses, producing plausible whole-body interactions.

### E.3 Simultaneous locomotion and manipulation

We demonstrate simultaneous locomotion and manipulation in Figure 5. Starting from the generated object motion, we manually add horizontal translation to simulate object movement in different directions. This translation is then set as the root position target in the optimization process. Our

Table 3: **Ablation study on the BPS number.**

BPS number	FID↓	R-Precision↑	Diversity→	Foot skating↓	IV↓	ID↓	CR↑
Real	—	0.531	8.664	0.002	4.68	11.47	0.085
1024	2.483	0.430	7.471	0.003	4.98	12.22	0.091
Ours (256)	2.283	0.477	7.208	0.002	5.25	12.87	0.086

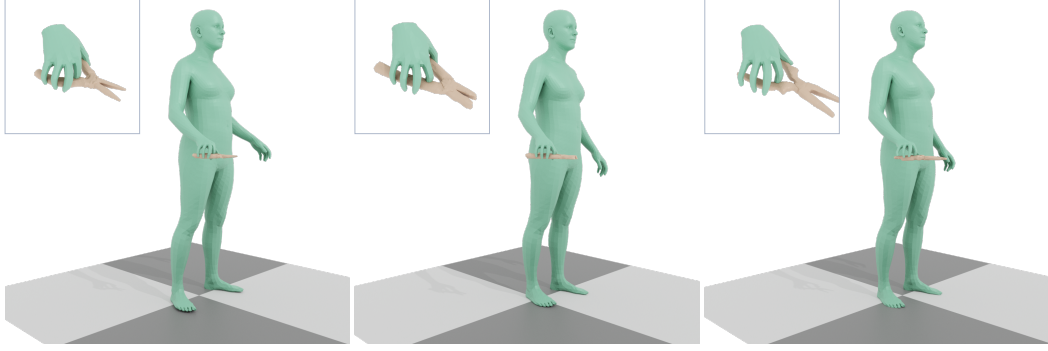


Figure 3: **Generalization to different object geometry.** We train the object motion and end-effector trajectory models on hand-only data [82] with different object geometries. These models are integrated into our framework to provide optimization targets, enabling realistic whole-body motion synthesis for unseen object geometry.

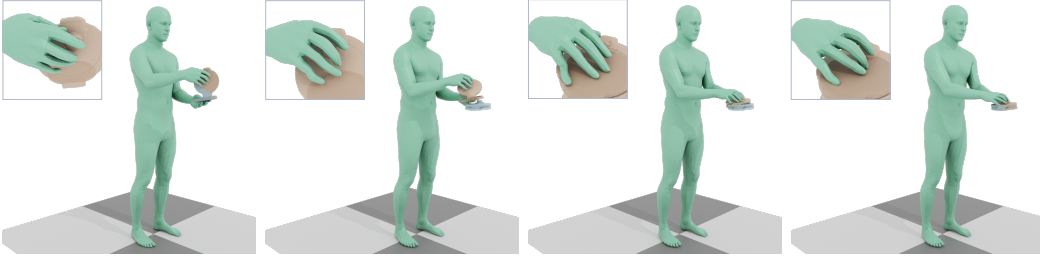


Figure 4: **Object motion control.** Our method could generate coherent whole-body motion with the object motion keyframe. The blue object indicates the object motion keyframe.

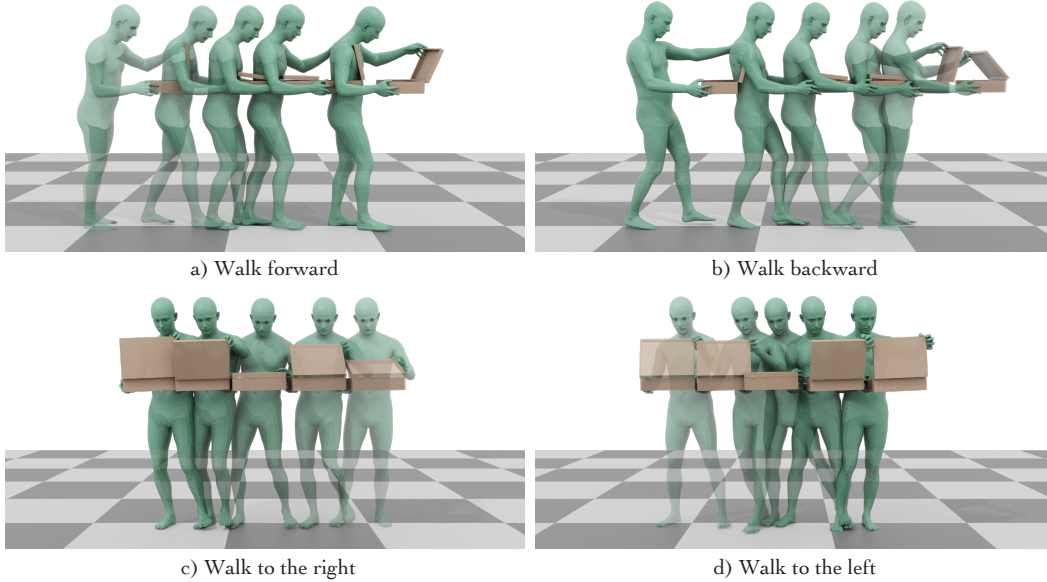


Figure 5: **Simultaneous locomotion and manipulation.** Our method enables the human to manipulate objects while simultaneously a) walking forward, b) walking backward, c) walking to the right, and d) walking to the left. The transparency of the meshes indicates time progression, where more transparent frames correspond to earlier frames.

method successfully produces whole-body motions that combine manipulation with walking forward, backward, left, and right. Notably, such combinations are not present in the ARCTIC [10] dataset, which only features manipulation while standing still.

#### E.4 Text-based control

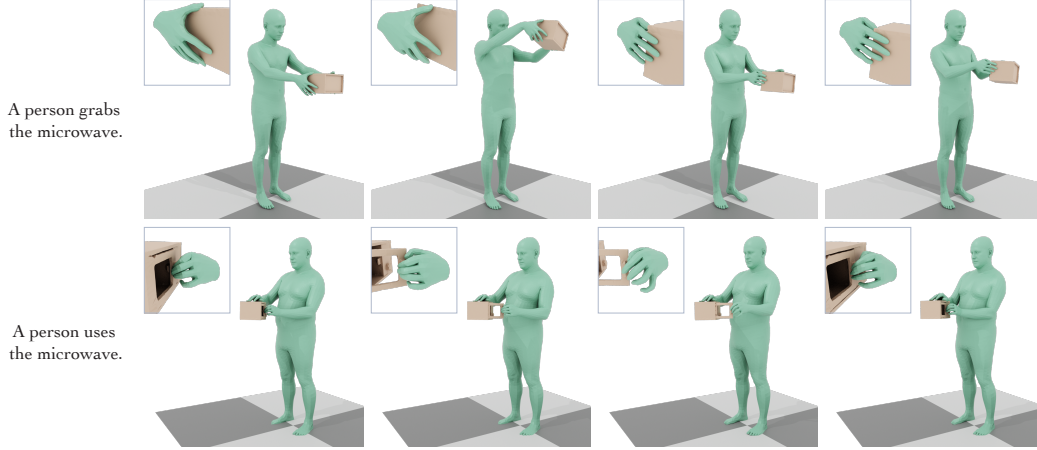


Figure 6: **Text-based control on the ARCTIC dataset.** Given different textual instructions (e.g., “grab the microwave” vs. “use the microwave”), our method generates distinct interaction behaviors that reflect the intended mode of manipulation.

**On the ARCTIC dataset.** Due to dataset constraints, most interactions fall into two primary modes: (1) grasp mode, where the articulated object is treated as a rigid object and manipulated without actuating its joints, and (2) use mode, which involves interacting with the object along its articulation axes, such as opening or closing. As shown in Figure 6, our method is capable of generating distinct interaction behaviors aligned with different textual instructions like “grab the microwave” and “use the microwave,” effectively capturing the semantic difference between these two modes. Although Text2HOI [2] provides rule-based annotations for left- and right-hand actions, they often result in very short sequences, making it difficult to generate temporally coherent long-horizon interactions. To support longer and more natural sequences, we instead use pre-defined text templates to assign instructions at the sequence level.

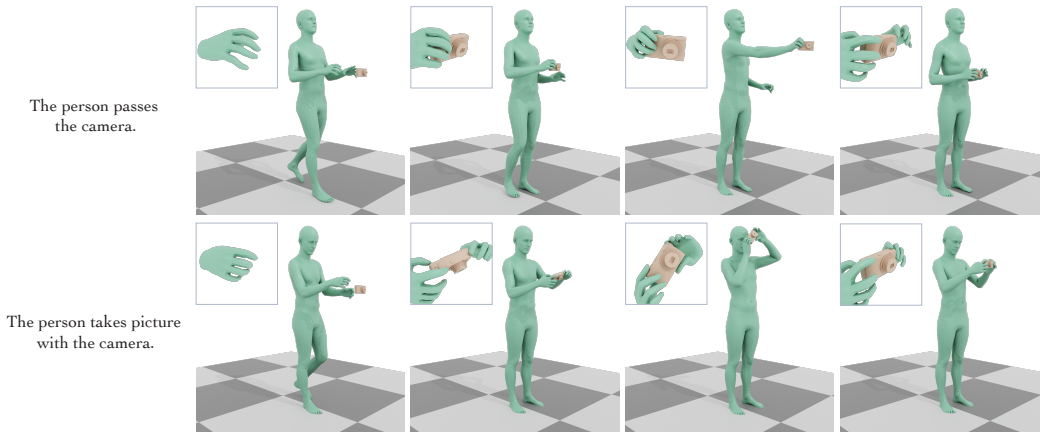


Figure 7: **Text-based control on the GRAB dataset.** Given different textual instructions (e.g., “pass the camera” vs. “take picture with the camera”), our method generates distinct interaction behaviors that reflect the intended mode of manipulation.



**On the GRAB dataset.** Figure 7 shows the results on the GRAB dataset. Our method generates different interaction behaviors in response to different instructions, such as “pass the camera” and “take a picture with the camera,” demonstrating the model’s ability to interpret and act upon textual input.

### E.5 Diverse results

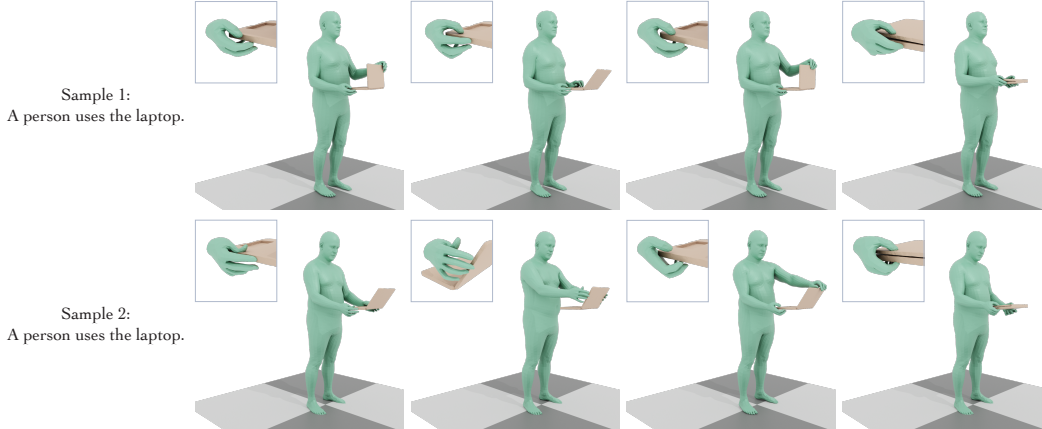


Figure 8: **Diverse results.** Our model generates different motions from identical text prompts, illustrating diversity in interaction synthesis.

Our model is capable of generating different human-object interactions conditioned on the same textual instruction. As shown in Figure 8, even with identical prompts, the resulting motions vary in execution, demonstrating the model’s ability to capture multiple interaction patterns.

### E.6 Deployment on simulated humanoids

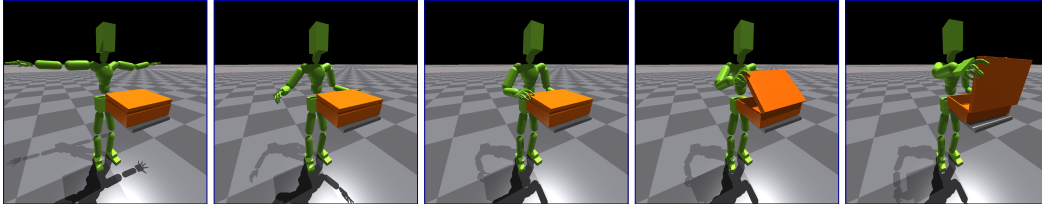


Figure 9: **Deployment on simulated humanoids.** We apply existing motion tracking techniques to deploy the generated motion to a simulated humanoid. The articulated object is physically manipulated by the humanoid within the physics simulator [41].

As shown in Figure 9, our generated whole-body motion can serve as a reference for controlling humanoids in physics-based simulators. We apply physical motion tracking methods [47, 37, 67] to track the synthesized motions. The humanoid is able to physically interact with objects and perform coordinated manipulation behaviors in the simulated environment.

### E.7 Generating whole-body motion from hand-only dataset

We demonstrate that our framework can generate whole-body motion from hand-only datasets. Specifically, we use the object trajectories provided by the dataset, and extract the end-effector trajectories (fingertips and wrists) as optimization targets. Our method only requires 3D positions of the fingertips and wrists, without the need for full joint rotations or detailed hand mappings, making it easier to apply. As shown in Figure 10, our method successfully produces realistic whole-body motions aligned with the provided hand-object interactions.

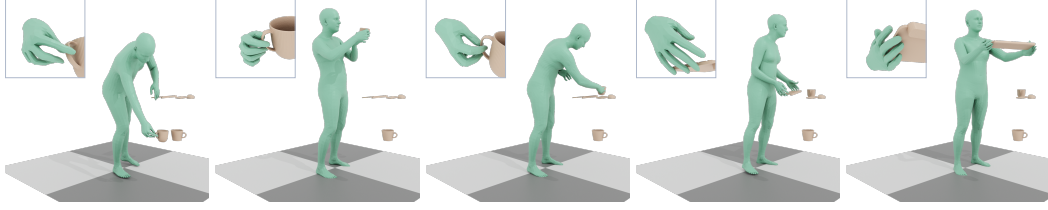


Figure 10: **Generating whole-body motion from hand-only dataset.** We use the fingertip and object trajectories from the dataset and assign them as the optimization targets. After the optimization, we could get the whole-body motion.

Table 4: **Inference time.**

Module	Object motion	End-effector	Whole-body motion
Time	0.52 secs	3.66 secs	16.93 mins

## E.8 Inference speed

Table 4 presents the inference time of each module in our pipeline. All results are measured on a single NVIDIA A100 GPU, generating a motion sequence of 300 frames. The majority of the time cost comes from the whole-body motion optimization process, which involves iterative diffusion sampling and gradient-based updates. Although slower than feed-forward models [30], this process enables high-quality, physically plausible whole-body interactions.

## E.9 Visualization of optimization process

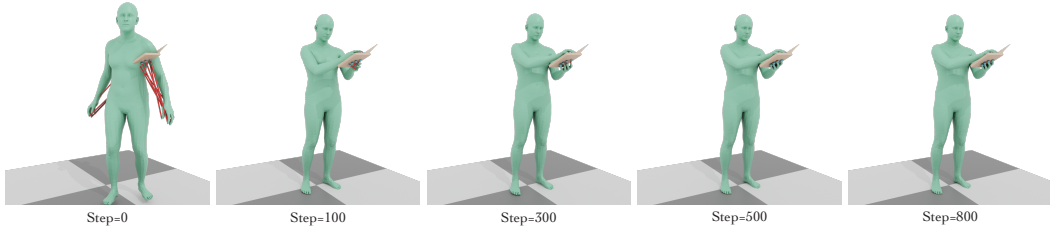


Figure 11: **Visualization of optimization process.** We visualize the optimization process in the optimization process. The blue points indicate the optimization targets while the red lines indicate the error.

We visualize the optimization process in Figure 11. In the early steps, the body motion has rapid adjustments to align with the wrist trajectories. As optimization progresses, finer adjustments appear in the finger motions. By the final steps, the motion satisfies the end-effector targets while minimizing hand-object penetration, resulting in realistic and physically plausible interaction.

## F More Discussions

### F.1 Limitations

The optimization process is relatively slow compared to feed-forward generative models such as MDM [60], which may hinder real-time deployment or interactive applications. Reducing inference time while maintaining quality remains an important direction for future work.

Due to the limited diversity of object categories in current datasets like ARCTIC [10], our model struggles to generalize to novel objects with significantly different geometries, topologies, or manipulation affordances. Our method only considers the rotational articulated objects, but other manipulation



tasks, such as pushing or pulling, are not considered. Scaling to broader object types would require more diverse and high-quality motion data.

Our current framework focuses on single-object manipulation. Extending it to multi-object scenarios, such as opening a drawer and retrieving an item, or performing sequential multi-step tasks, poses both modeling and optimization challenges and remains an open problem.

Our method does not explicitly model obstacle avoidance. While the resulting body and hand motions are physically plausible, the character may occasionally intersect with surrounding geometry in cluttered or constrained scenes. Enabling both the fingers and the full body to reason jointly about nearby obstacles and environment geometry is an important direction for improving interaction realism.

Our text conditioning is currently limited by the simplicity of available annotations. While Text2HOI [2] provides rule-based captions for hand-object interactions, they are typically segmented into short atomic motions, whereas we aim to model longer and more coherent manipulation sequences. Developing richer textual annotations and grounding them to temporally extended actions is a promising avenue for future work.

Our method primarily focuses on manipulation tasks, with the body posture adapting naturally to support the behaviors. However, some interactions require direct contact between the object and other body parts, such as pressing a box against the torso or holding an item between the arm and the body. Modeling such whole-body contact behaviors remains largely unexplored and could further expand the expressiveness of interaction generation.

## F.2 Broader impact

Our method can be used to create a realistic manipulation sequence, which could be rendered as a video. It also has the potential to be transferred to humanoid robots [38]. Therefore, our method has a potential positive social impact to help build the development of character animation and humanoid robotics.

## References

- [1] Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-react: Online reaction policy for two-character interaction generation. In *ICLR*, 2025. 2
- [2] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. 8, 11
- [3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 2
- [4] Changan Chen, Juze Zhang, Shrinidhi Kowshika Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *arXiv*, 2024. 2
- [5] Ling-Hao Chen, Shunlin Lu, Wenxun Dai, Zhiyang Dou, Xuan Ju, Jingbo Wang, Taku Komura, and Lei Zhang. Pay attention and move better: Harnessing attention for interactive motion generation and training-free editing. *arXiv preprint arXiv:2410.18977*, 2024. 2
- [6] Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 2
- [7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your Commands via Motion Diffusion in Latent Space. *arXiv e-prints*, art. arXiv:2212.04048, December 2022. doi: 10.48550/arXiv.2212.04048. 2
- [8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

- [9] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C-ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2
- [10] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 8, 10
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 2020. 2
- [12] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2
- [13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 2
- [14] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2
- [15] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Trans. Graph.*, 2020. 2
- [16] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 2
- [17] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Trans. Graph.*, 2020. 2
- [18] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv e-prints*, art. arXiv:2207.12598, July 2022. 2
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3
- [20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. *arXiv e-prints*, art. arXiv:2204.03458, April 2022. 2
- [21] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Trans. Graph.*, 2017. 2
- [22] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79. 2
- [23] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2
- [24] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1345, 2024. 2, 4
- [25] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, 2014. 4
- [26] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv e-prints*, 2013. 2
- [27] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *NeurIPS*, 2018. 2

- [28] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 2
- [29] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 6
- [30] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pages 54–72. Springer, 2024. 6, 10
- [31] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: Neural motion synthesis from a single sequence. *ACM Trans. Graph.*, 2022. 2
- [32] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021. 2
- [33] 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*, 36, 2024. 2
- [34] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2024. 2
- [35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>. 3
- [36] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Juntao Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. *arXiv preprint arXiv:2412.14559*, 2024. 2
- [37] 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. 2, 9
- [38] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris M. Kitani, and Weipeng Xu. Omnigrasp: Simulated humanoid grasping on diverse objects. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=Glt37xoU7e>. 11
- [39] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=OrOd8Px002>. 2
- [40] 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. 2, 4
- [41] 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. 2, 9
- [42] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017. 2
- [43] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL [https://openreview.net/forum?id=aBsCjcPu\\_tE](https://openreview.net/forum?id=aBsCjcPu_tE). 2

- [44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 3
- [45] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*, 2024. 2
- [46] OpenAI. Openai: Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022. 2
- [47] 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, 9
- [48] 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)*, 40(4):1–20, 2021. 2
- [49] 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)*, 41(4):1–17, 2022. 2
- [50] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021. 2
- [51] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 2016. 2
- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3
- [53] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 2
- [54] 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. 2
- [55] 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. 3
- [56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=StlgiaRCHLP>. 3, 4
- [57] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 2019. 2
- [58] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Trans. Graph.*, 2020.
- [59] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.*, 2022. 2
- [60] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 3, 10
- [61] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, 2023. 2
- [62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3

- [63] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023. 2
- [64] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 2
- [65] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. In *ECCV 2024*, pages 37–54. Springer Nature Switzerland, 2024. 2
- [66] Jiashun Wang, Jessica Hodgins, and Jungdam Won. Strategy and skill learning for physics-based table tennis animation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [67] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physshoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 9
- [68] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Trans. Graph.*, 41(4), July 2022. ISSN 0730-0301. doi: 10.1145/3528223.3530067. URL <https://doi.org/10.1145/3528223.3530067>. 2
- [69] Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025. 2
- [70] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [71] 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)*, 41(6):1–16, 2022. 2
- [72] Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Transactions on Graphics (TOG)*, 43(4):1–21, 2024. 2
- [73] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Trans. Graph.*, 2023. 2
- [74] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Trans. Graph.*, 2018. 2
- [75] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 2021. 4
- [76] Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *arXiv preprint arXiv:2409.09300*, 2024. 6
- [77] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023. 2
- [78] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint*, 2022. 2

- [79] Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 6
- [80] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 2
- [81] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2025. 2
- [82] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023. 6, 7
- [83] Wenyang Zhou, Zhiyang Dou#, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *ECCV 2024*, pages 18–38. Springer Nature Switzerland, 2024. 2
- [84] Qingxu Zhu, He Zhang, Mengting Lan, and Lei Han. Neural categorical priors for physics-based character control. *ACM Trans. Graph.*, 42(6), dec 2023. ISSN 0730-0301. doi: 10.1145/3618397. URL <https://doi.org/10.1145/3618397>. 2