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

Huaijin Pi[1] Zhi Cen[2] Zhiyang Dou[1] Taku Komura[1]

[1]The University of Hong Kong [2]Zhejiang University
https://phj128.github.io/page/CoDA

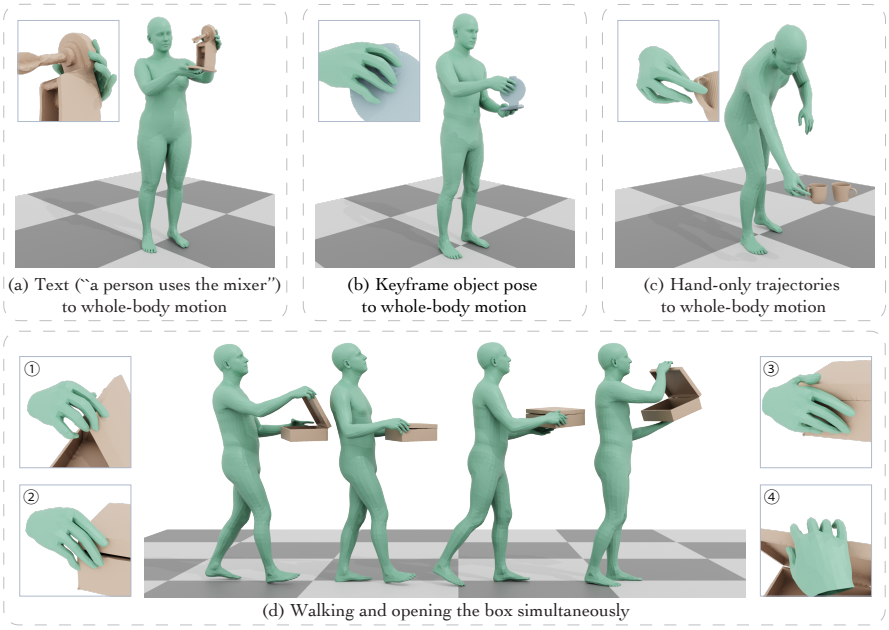


Figure 1: Our approach enables: (a) generating whole-body manipulation of articulated objects from text input (e.g., “a person uses the mixer”); (b) manipulating the object to a target pose and articulation (the blue object is the target pose); (c) synthesizing whole-body motion guided by trajectories from hand-only data; (d) generating motions involving simultaneous walking and object manipulation (e.g., opening a box while walking).

Abstract

Synthesizing whole-body manipulation of articulated objects, including body motion, hand motion, and object motion, is a critical yet challenging task with broad applications in virtual humans and robotics. The core challenges are twofold. First, achieving realistic whole-body motion requires tight coordination between the hands and the rest of the body, as their movements are interdependent during manipulation. Second, articulated object manipulation typically involves high degrees of freedom and demands higher precision, often requiring the fingers to be placed at specific regions to actuate movable parts. To address these challenges, we propose a novel coordinated diffusion noise optimization framework. Specifically, we perform noise-space optimization over three specialized diffusion models for the body, left hand, and right hand, each trained on its own motion dataset to improve generalization. Coordination naturally emerges through gradient flow

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

along the human kinematic chain, allowing the global body posture to adapt in response to hand motion objectives with high fidelity. To further enhance precision in hand-object interaction, we adopt a unified representation based on basis point sets (BPS), where end-effector positions are encoded as distances to the same BPS used for object geometry. This unified representation captures fine-grained spatial relationships between the hand and articulated object parts, and the resulting trajectories serve as targets to guide the optimization of diffusion noise, producing highly accurate interaction motion. We conduct extensive experiments demonstrating that our method outperforms existing approaches in motion quality and physical plausibility, and enables various capabilities such as object pose control, simultaneous walking and manipulation, and whole-body generation from hand-only data. The code will be released for reproducibility.

1 Introduction

Human-object interaction (HOI) motion generation [47, 54] has broad applications in virtual reality, character animation [5, 68, 48], and robotics. These interactions range from simple activities like sitting on a chair [54, 47] to more complex tasks involving articulated object manipulation [6, 21], such as opening a box or a microwave. This paper focuses on the challenging setting of whole-body manipulation of articulated objects. Given an initial pose of the human and the object, along with a textual instruction, our goal is to synthesize realistic, physically plausible interaction sequences that involve coordinated body, hand, and articulated object motion.

Most prior works on HOI generation [16, 47, 74, 4, 28, 29, 8] suffer from two key limitations. First, they typically focus on either body-only motion [47, 28, 29] or hand-only manipulation [73, 81, 6, 21, 8]. Although hand-only methods can produce plausible contact behaviors in short-range scenarios, they fail to capture important whole-body dynamics such as bending down, reaching forward, or walking while manipulating objects. Such whole-body behaviors are essential for generating realistic human-object interactions, especially when manipulation is not restricted to a fixed space. Second, most existing works target rigid objects [70, 29, 8], while articulated objects introduce more complex motion patterns and require continuous in-hand adjustments.

Whole-body manipulation of articulated objects is highly challenging. First, it demands coordinated motion between the body and hands to reflect natural physical behaviors. Body movement affects how the hands approach and manipulate objects, and conversely, hand-object interactions can influence global posture. Second, precise control of finger positions is essential to maintain accurate, physically plausible contact throughout the sequence. This is especially important for articulated objects, where the manipulation often requires placing the fingers at specific regions to actuate the articulation while avoid colliding with other parts.

To address these challenges, we propose a novel framework called CoDA (**Co**ordinated **D**iffusion noise optimization for whole-body manipulation of **A**rticulated objects), which jointly synthesizes the motions of the human body, hands, and articulated objects. Our core idea is to optimize the input noise vectors of three specialized diffusion models, which independently model the body, left hand, and right hand, to generate coordinated whole-body motion. This decoupled design allows each component to be trained on its own data source, such as using large-scale datasets like AMASS [38] for body motion, manipulation datasets like ARCTIC [13] and GRAB [56] for hand motion, thereby improving generalization across diverse motions. Coordination naturally emerges during optimization, as gradients from hand motion objectives flow through the human kinematic chain, allowing the global posture to adapt in response to fine-grained hand motion. This optimization further enables precise control over hand-object contact, while the diffusion noise space [25] provides strong motion priors to preserve naturalness in the generated sequences.

To enable precise manipulation while accounting for object geometry and articulation, we adopt a basis point set (BPS) representation [49, 81] to encode both the object surface and end-effector trajectories in a unified form. Specifically, we represent the positions of the end-effectors, namely the wrists and fingertips, by their distances to the same BPS used for encoding the object geometry. The unified representation captures the relative spatial relationship between the hand and the object geometry as well as its articulation during complex manipulation tasks. The generated trajectories, based on this representation, provide a continuous target signal for optimizing whole-body motion.

We evaluate our approach on both the ARCTIC [13] dataset of articulated object manipulation and the GRAB [56] dataset of rigid object interactions. Our method achieves state-of-the-art performance on both benchmarks, outperforming existing approaches in motion quality and physical plausibility. Beyond benchmark evaluation, our framework enables several compelling capabilities, as illustrated in Figure 1. It supports object pose control at specific times, and coordinated whole-body behaviors involving simultaneous locomotion and manipulation, which are absent from the ARCTIC dataset. In addition, our framework allows us to leverage hand-only datasets [2] to generate whole-body motion, enabling broader data usage and generalization. To the best of our knowledge, this is the first work to jointly generate body, hand, and articulated object motions for whole-body manipulation tasks.

2 Related work

Human-object interaction. Human-object interaction (HOI) generation [54, 28, 9] has received increasing attention due to its potential to enable virtual humans to perform various actions in 3D environments. Early works focus on generating static interactions such as sitting or lying on furniture [54, 16, 82, 84, 80], using either auto-regressive pipelines or whole-sequence generation [61, 62, 40, 1, 86]. Recent methods explore diffusion-based models [22, 47, 27, 4, 74, 23] and apply guidance techniques [11, 19] to improve human-scene contact quality. Beyond static objects, several works consider dynamic objects [70, 71] or generate human motion conditioned on given object trajectories [28, 10]. For example, OMOMO [28] proposes a two-stage framework that first generates wrist trajectories and then completes body motion accordingly. Other approaches [45, 70, 29, 12, 53, 72] jointly generate body and object motion, and incorporate contact-aware guidance into the diffusion process to improve the quality. Another line of research [17, 42, 69, 59, 43, 63] enables physically simulated characters to perform scene-level interactions by learning control policies through environment interaction. These methods mainly focus on navigation and interactions with large-scale objects such as furniture or obstacles. While generating plausible body motion, they ignore finger motion, which is crucial for fine-grained manipulation.

Hand-object interaction. ManipNet [77] synthesizes object manipulation given wrist and object trajectories, using multiple representations to model the hand-object relationship. GRIP [58] design a temporal hand-object spatial feature for stable grasping. Some works [87, 33] address the task of denoising noisy hand motion to recover clean interaction sequences. While these methods explore various representations for modeling hand-object spatial relationships, they rely on access to predefined wrist and object trajectories. [85, 81, 79] explore settings where only the object trajectory is provided. CAMS [85] introduces a canonicalized representation to enable precise contact generation. [81, 79] generate manipulation by predicting contact maps as intermediate representations. Other works generate hand and object motion jointly, without relying on predefined trajectories. DiffH2O [8] applies grasp guidance to diffusion models for more coherent hand-object interactions.. Text2HOI [6] employs cascaded diffusion to iteratively refine the results. HOIGPT [21] leverages separate codebooks for hands and objects, and jointly predicts motion and text. Physics-based approaches [7, 78] generate grasping motions through reinforcement learning in simulated environments. Despite their differences, all these methods ignore the body context, resulting in floating hand motions.

Whole-body interaction. Although there are several whole-body manipulation datasets [56, 13, 24, 23, 76, 37, 34], only a few works consider body and hand interaction simultaneously. [57, 66] assume the object is static and only synthesize approaching and grasp motion. IMoS [14] demonstrates full-body manipulation with given finger motion; it generates body motion auto-regressively and optimizes object trajectories by assuming a static hand-object contact frame. TOHO [30] synthesizes whole-body interactions using implicit representations [18], relying on the same contact assumption to recover object motion. DiffGrasp [83] generates whole-body motion conditioned on given object trajectories using diffusion models, and introduces hand-object guidance to improve interaction quality. Wu et.al. [67] employs LLM [41] to analyze the scene and plan motions for grasping and relocating rigid objects. Other works [64, 65, 75] employ physics-based tracking to mimic manipulation behaviors. [3, 36, 31] explores humanoid grasping, but the generated motions remain unnatural and do not involve complex manipulation. Most of the above methods focus exclusively on rigid object interaction and do not address articulated objects. Compared to rigid object interaction, articulated object manipulation is more complex, as it often requires placing the fingers at specific regions to actuate the articulation.

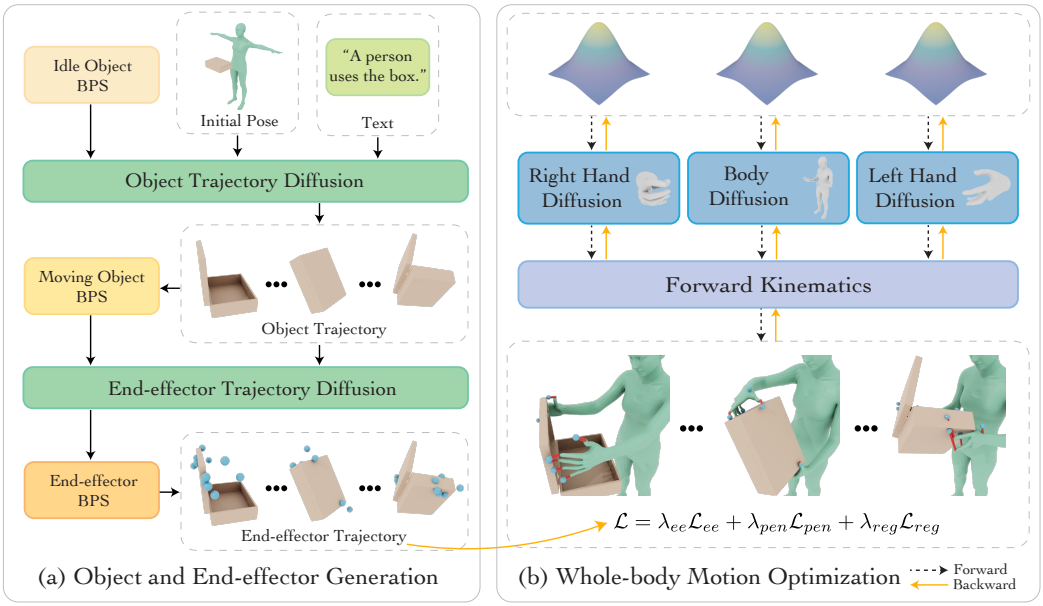


Figure 2: **Pipeline overview.** (a) Given the initial human pose, object pose, and text, we first generate the articulated object trajectory and the corresponding end-effector trajectories via two conditional diffusion models. (b) We then optimize the latent noise inputs of three decoupled diffusion models by propagating gradients through the kinematic chain, guided by end-effector tracking, penetration, and regularization losses. Finally, we forward the optimized noise through the diffusion models to synthesize coherent whole-body motion aligned with the generated object motion.

3 Preliminary

In this section, we define the input and output in this paper. Given the initial pose of a human and an articulated object, along with a textual instruction, our goal is to generate a full sequence including the whole-body human motion (body and fingers) and the articulated object motion over time.

Object representations. The objects from the ARCTIC [13] dataset are two-part articulated objects with 7 degrees of freedom. We use $\mathbf{S}_o = \{\mathbf{T}_o, \mathbf{R}_o, \mathbf{a}\}$ to indicate the object pose, where the object state $\mathbf{S}_o \in \mathbb{R}^7$ consists of object translation $\mathbf{T}_o \in \mathbb{R}^3$, object rotation $\mathbf{R}_o \in \mathbb{R}^3$, and the angle of the rotational joint $\mathbf{a} \in \mathbb{R}^1$ between the two parts of the object.

Motion representations. We use SMPL-X [44], which is a parametric human body model to represent the whole body, including the face and fingers. SMPL-X is a differentiable function that takes input shape, pose, and expression parameters and outputs a 3D mesh with $10,475$ vertices and $20,908$ triangles. The vertices are posed with linear blend skinning with a rigged skeleton which is learned from the data. As we focus on the body motion with two hands, we remove the face related parameters. $\mathbf{\Theta} = \{\boldsymbol{\theta}, \mathbf{t}\}$ is the pose parameters to drive the SMPL-X model, where $\boldsymbol{\theta} \in \mathbb{R}^{52\times3}$ represents joint angles and $\mathbf{t} \in \mathbb{R}^3$ is the root translation.

Text descriptions. In the ARCTIC [13] and the GRAB [56] dataset, each sequence is annotated with an action label. Following previous work [14, 6], we construct the text description using the template "A person <action> the <object>.". For example, "A person uses the box.".

4 Method

The overview of our pipeline is shown in Figure 2. We first generate the motion of the articulated object (Section 4.1), then predict the end-effector trajectories (Section 4.2), and finally synthesize the whole-body motion by optimizing the noise of decoupled diffusion models (Section 4.3).

4.1 Object motion generation

Given the initial object pose and the textual instruction, we train a diffusion model [60] to generate the object future trajectory. The input includes the CLIP [50] feature of the text, the initial object pose, and the object geometry embedding. We represent the object geometry using the normalized part-based BPS descriptor [81], which will be formally defined in Section 4.2 and Figure 3. The output is a sequence of object states over time.

4.2 End-effector trajectory generation

Given the generated object trajectory, we extract its geometry representation and combine it with the trajectory itself and the textual instruction as input to a diffusion model that predicts end-effector trajectories. Instead of directly predicting 3D joint coordinates [28], we design a distance-based representation that encodes end-effector positions in the same space as the object geometry.

Unified BPS-based representation for object and end-effectors. We first present the object geometry representation. Following previous work [81], we adopt the normalized part BPS [49] to represent the object geometry. Specifically, the object mesh is first normalized to the unit scale by dividing all vertex coordinates by the maximum distance from the object origin to any vertex. Then a pre-defined fixed set of basis points $\mathbf{P} \in \mathbb{R}^{K\times 3}$, shared across all objects, are uniformly sampled within the unit sphere centered at the object origin. The BPS representation is computed as the distances from each basis point to the nearest vertex on each of the two rigid object parts, resulting in an object geometry vector $\mathbf{O} \in \mathbb{R}^{K\times 2}$.

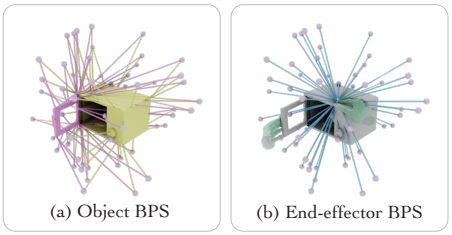
We then introduce *end-effector BPS*, a distance-based representation tailored for encoding the positions of end-effectors in the object coordinate system. The end-effectors include both wrists and fingertips, comprising a total of 12 joints (2 wrists and 10 fingertips). As shown in figure 3, at each frame, for each of the 12 end-effectors, we compute a K-dimensional vector of Euclidean distances to the basis points. We use the same pre-defined set of basis points $\mathbf{P} \in \mathbb{R}^{K\times 3}$ in object geometry representation [81]. This results in a $(12 \times K)$-dimensional end-effector BPS vector per frame. The diffusion model outputs a sequence of end-effector BPS over time, along with binary contact labels for each fingertip, indicating whether it is close to the object surface.

Given the generated end-effector BPS sequence, we recover the end-effector trajectories by solving a simple optimization problem. For each end-effector at each frame, we minimize the following loss to infer its 3D position:

$$\mathbf{p}_e^* = \arg\min_{\mathbf{p}_e} \mathcal{L}(\mathbf{p}_e), \tag{1}$$

$$\mathcal{L}(\mathbf{p}_e) = \sum_j \| \|\mathbf{p}_e - \mathbf{P}_j\|_2 - d_j \|_2, \tag{2}$$

where $\mathbf{p}_e$ is the optimized 3D position and d_j is the predicted distance to the j-th basis point $\mathbf{P}_j$. By sharing the basis point set with the object BPS representation, our method provides a consistent spatial reference frame that facilitates geometric alignment between end-effectors and object parts.



(a) Object BPS (b) End-effector BPS

Figure 3: **The illustration of the end-effector BPS.** (a) is the object BPS [81]. (b) is the proposed end-effector BPS representation. Gray points denote the basis points; pink/yellow are two object parts; blue indicates a fingertip. Only one end-effector and 64 basis points are visualized for simplicity.

RoPE-based object motion encoding. To better encode the object trajectory, we adapt the idea of CaPE [26], which encodes relative camera pose information via RoPE [55]. In our case, each object pose is also represented as a 4×4 transformation matrix. Inspired by CaPE [26], we use the object pose to transform the query and key features in each attention layer. This enables the model to encode the relative object motion within a local temporal window, providing temporally-aware conditioning for the generation. We refer readers to the supplementary material for more details.

4.3 Whole-body motion generation

The goal of this stage is to generate coherent whole-body motion that aligns with the predicted end-effector trajectories and articulated object motion. Rather than directly predicting whole-body

poses conditioned on end-effectors [28], we adopt an optimization-based approach inspired by DNO [25]. Specifically, we optimize the noise input to the diffusion models (Figure 2 (b)), and then forward the optimized noise through the diffusion models to generate the final motion. To further improve motion quality, we decouple the body into three components: body, left hand, and right hand, and train separate diffusion models for each. This decoupled design enables us to train each module using individual data, such as training the hand models using the ARCTIC [13] and GRAB [56], and the body-only model without hands on the AMASS [38]. Such specialization improves generalization by allowing novel combinations of finger and body motion to be synthesized. Moreover, this formulation facilitates gradient flow through the kinematic chain during optimization, which improves coordination between the body and hands.

Decoupled motion diffusion model. We adopt a decoupled human representation for whole-body motion, dividing the human pose into three components: body, left hand, and right hand. Formally, for each frame i, the whole-body pose $\mathbf{\Theta}_i$ is represented as:

$$\mathbf{x} = \{\mathbf{x}_b, \mathbf{x}_{lh}, \mathbf{x}_{rh}\}, \tag{3}$$

$$\mathbf{x}_b = \{\dot{r}^x, \dot{r}^z, r^y, \dot{r}^a, \boldsymbol{\theta}_b\}, \tag{4}$$

$$\mathbf{x}_{lh} = \{\boldsymbol{\theta}_{lh}\}, \tag{5}$$

$$\mathbf{x}_{rh} = \{\boldsymbol{\theta}_{rh}\}, \tag{6}$$

where x_b denotes the body component, including root velocities $\dot{r}^x, \dot{r}^z \in \mathbb{R}$ (projected on the XZ-plane), root height $r^y \in \mathbb{R}$, angular velocity $\dot{r}^a \in \mathbb{R}$, and body joint rotations $\boldsymbol{\theta}_b \in \mathbb{R}^{6\times J_b}$, while $\boldsymbol{\theta}_{lh} \in \mathbb{R}^{6\times J_{lh}}$ and $\boldsymbol{\theta}_{rh} \in \mathbb{R}^{6\times J_{rh}}$ represent the left and right hand joint rotations, respectively. All joint rotations are encoded using the 6D representation [88], with $J_b = 22$, $J_{lh} = J_{rh} = 15$ joints for the body and each hand. We train three separate diffusion models, $\mathbf{M}_b$, $\mathbf{M}_{lh}$, and $\mathbf{M}_{rh}$, to model the motion manifolds of the body and hands individually.

Optimization over diffusion noise. Given the trained diffusion models for body, left hand, and right hand, we optimize the noise vectors $\mathbf{z} = \{\mathbf{z}_b, \mathbf{z}_{lh}, \mathbf{z}_{rh}\}$ to generate whole-body motion as shown in Figure 2 (b). Let $f(\mathbf{z})$ denote the process that maps the input noise to global joint positions through diffusion models and forward kinematics:

$$f(\mathbf{z}) = \mathcal{FK}(\mathbf{M}_b(\mathbf{z}_b), \mathbf{M}_{lh}(\mathbf{z}_{lh}), \mathbf{M}_{rh}(\mathbf{z}_{rh})), \tag{7}$$

where $\mathcal{FK}(\cdot)$ converts root translation and local joint rotations into global joint positions. We formulate motion generation as minimizing a loss $\mathcal{L}$ over the diffusion noise:

$$\mathbf{z}^* = \arg\min_{\mathbf{z}} \mathcal{L}(f(\mathbf{z})). \tag{8}$$

The overall loss function consists of three components with different weights λ_{ee}, λ_{pen}, and λ_{reg}:

$$\mathcal{L} = \lambda_{ee}\mathcal{L}_{ee} + \lambda_{pen}\mathcal{L}_{pen} + \lambda_{reg}\mathcal{L}_{reg}, \tag{9}$$

where $\mathcal{L}_{ee}$, $\mathcal{L}_{pen}$, and $\mathcal{L}_{reg}$ are the end-effector tracking, penetration, and regularization losses.

We encourage the generated global fingertip positions $\hat{\mathbf{p}}_f$ to follow the predicted trajectories $\mathbf{p}_f$ from the previous stage. We also constrain the relative fingertip positions to the wrist joints:

$$\mathcal{L}_{ee} = \|\hat{\mathbf{p}}_f - \mathbf{p}_f\|_1 + \|\hat{\mathbf{p}}_f^r - \mathbf{p}_f^r\|_1, \tag{10}$$

where $\mathbf{p}_f^r$ and $\hat{\mathbf{p}}_f^r$ denote the relative fingertip positions with respect to the wrist.

To reduce hand-object interpenetration, we penalize fingertip joints that fall inside the object mesh:

$$\mathcal{L}_{pen} = \sum_j \left\|\min\left(\mathrm{SDF}(\mathbf{J}^j) - 0.01, 0.00\right)\right\|_1 \tag{11}$$

where $\mathrm{SDF}(\mathbf{J}^j)$ is the signed distance at the j-th hand joint, assuming 1cm finger thickness.

We add a regularization term to discourage foot floating and foot sliding:

$$\mathcal{L}_{reg} = \|\min\left(\mathbf{J}^y\right) - 0.02\|_1 + \mathbf{1}_{\text{left}} \cdot \left\|\mathbf{J}_i^l - \mathbf{J}_{i-1}^l\right\|_1 + \mathbf{1}_{\text{right}} \cdot \left\|\mathbf{J}_i^r - \mathbf{J}_{i-1}^r\right\|_1, \tag{12}$$

where $\mathbf{J}^y$ denotes the height of all joints in the body, and $\mathbf{J}_i^l$ and $\mathbf{J}_i^r$ denote the 3D positions of the left and right foot joints at frame i, respectively. The binary indicators $\mathbf{1}$ denote whether the left or right foot is in contact with the ground, based on a height threshold of 0.02 meters.

We adopt DDIM [52] sampling to efficiently generate motion sequences during optimization following DNO [25]. The loss is computed on the final output, and gradients are propagated back through the DDIM solver to update the noise. After optimization, we pass the optimized noise into the decoupled diffusion models to generate the final whole-body motion. Combined with the previously generated object trajectory, this yields a complete human-object manipulation sequence. This noise-space optimization avoids high-dimensional pose regression, reduces artifacts, and produces natural whole-body motions aligned with the object manipulation process.

5 Experiments

5.1 Implementation details

We adopt a transformer-based diffusion architecture similar to MDM [60] for all models in our framework. During inference, we perform noise optimization using DDIM [52] with T=10 for 800 steps and a cosine-decayed learning rate, following the DNO [25] strategy. All experiments are conducted on a single NVIDIA A100 GPU. More training details are in the supplementary material.

5.2 Dataset and evaluation metrics

Dataset. We evaluate on ARCTIC [13] for articulated object manipulation and on GRAB [56] for rigid object interaction. ARCTIC contains around 2 hours of motion data featuring 10 subjects interacting with 11 articulated objects, including complex motions such as bimanual grasps and in-hand manipulation. Following the protocol in [81], we randomly sample 4 sequences per object category to construct the test set. The GRAB dataset covers about 4 hours of interaction from 10 subjects with 51 rigid objects, focusing primarily on grasping and simple lifting actions. Similar to [14], we use data from the last subject as the test set. For training object motion and end-effector trajectories generation, ARCTIC is used for articulated objects, and GRAB is used for rigid objects. The body motion model is trained on ARCTIC, GRAB, and AMASS [38], while the two hand motion models are trained on ARCTIC and GRAB.

Evaluation metrics. Similar to [8, 6], we evaluate the motion quality using the following metrics: (1) **Frechet Inception Distance (FID)** measures the feature-level distance between generated and real motions, using a motion feature extractor trained on the dataset following [15]. (2) **R-Precision** quantifies the alignment between generated motion and the corresponding textual prompt, measured using Top-3 accuracy. (3) **Diversity** reflects the variation among generated motion samples. (4) **Foot skating** indicates motion realism by detecting undesired foot sliding, following the computation in [32, 47]. We additionally report physical realism metrics following [8]: (5) **Interpenetration volume (IV)** computes the number of hand vertices that penetrate the object mesh. (6) **Interpenetration depth (ID)** measures the maximum penetration depth of hand vertices into the object. (7) **Contact ratio (CR)** is defined as the average proportion of hand vertices within 5 mm of the object surface. We also conduct a user study involving 16 participants to evaluate the generated motion sequences.

5.3 Comparison with baselines

Baselines. As there is no existing method that jointly generates body, hand, and articulated object motion, we adapt several representative methods to our task: IMoS [14], MDM [60], OMOMO [28], Text2HOI [6], and CHOIS [29]. IMoS is a CVAE-based [51] auto-regressive model, while MDM is a full-sequence diffusion-based [20] model. Text2HOI is originally designed for hand-object interaction with multiple diffusion models for iterative refinement. CHOIS is a diffusion-based model that incorporates contact guidance during inference. We extend them to jointly generate whole-body motion and object motion. OMOMO first generates wrist motion and then synthesizes body motion. We extend it to a three-stage model: first generating object motion, then predicting fingertip and wrist trajectories, and finally producing whole-body motion. OMOMO+DNO further extends OMOMO by using its diffusion model as a latent prior and applying DNO [25] to refine the generated results.

Quantitative results. We report quantitative results on ARCTIC and GRAB in Table 1 and Table 3, and user study results in Table 2. Our method achieves the best performance on nearly all metrics across both datasets. While it ranks slightly lower in diversity, it significantly outperforms all

Table 1: **Comparison on the ARCTIC [13] dataset.** The right arrow → means the closer to real motion the better. IV, ID, and CR denote interpenetration volume, interpenetration depth, and contact ratio. 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
IMoS [14]	6.686	0.305	6.144	1.469	14.28	13.24	0.010
MDM [60]	3.972	0.209	8.167	0.027	16.90	15.85	0.033
Text2HOI [6]	6.654	0.234	5.923	0.028	12.72	17.14	0.010
OMOMO [28]	3.710	0.406	6.110	0.028	13.77	15.16	0.061
OMOMO + DNO [25]	2.873	0.391	7.004	0.022	8.95	13.30	0.075
CHOIS [29]	3.758	0.367	7.423	0.023	17.19	15.84	0.030
Ours	2.283	0.477	7.208	0.002	5.25	12.87	0.086

Table 2: **User study on the ARCTIC [13] dataset.**

Metrics	Ours	CHOIS [29]	OMOMO [28]	Text2HOI [6]
Best Motion Realism Rate ↑	88.7%	1.1%	9.9%	0.3%
Best Physical Plausibility Rate ↑	87.3%	1.4%	10.2%	1.1%

baselines in the user study, indicating superior perceptual quality and physical plausibility. To assess the physical feasibility of our method, we conduct a mimic-based evaluation following [46, 64], where a humanoid policy is trained to reproduce the generated motions in the IsaacGym [39] simulator. We use eight generated sequences (each 10 seconds long) involving boxes and microwaves, and measure the tracking duration—the time (in seconds) during which both object and joint position errors remain below a 10 cm threshold. Results in Table 4 show that our motions lead to longer tracking durations compared to OMOMO, demonstrating improved physical plausibility and better compatibility with downstream humanoid execution. In addition, to evaluate the generalization capability of our method, we conduct an additional experiment on the ARCTIC dataset by holding out the box object for testing and training on the remaining objects. As shown in Table 5, our method achieves a substantially lower FID compared to the baseline, demonstrating better object-level generalization.

Qualitative results. As demonstrated in Figure 4, our method achieves significantly better hand-object contact compared to baselines. We provide more results in the supplementary material.

5.4 Ablation study

We ablate key components of our framework to understand their impact on overall performance: (a) A single model to jointly predict object motion and end-effector trajectories. (b) Predicting relative coordinate of end-effectors to the object center without end-effector BPS. (c) Using object velocity and rotational velocity as the trajectory input without RoPE-based representation. (d) Removing the optimization process and using a conditional diffusion model with fingertip trajectories as input. (e) Using a single diffusion model for the entire body without the decoupled body-hand representation. (f) Excluding the AMASS [38] dataset during training the body motion model. (g) replacing the end-effector representation with a distance field [77], where the trajectory is encoded as a fixed grid of distances in the object's local coordinate frame, while keeping the object geometry encoded using

Table 3: **Comparison on the GRAB [56] dataset.**

Methods	FID↓	R-Precision↑	Diversity→	Foot skating↓	IV↓	ID↓	CR↑
Real	–	0.727	15.045	0.010	5.84	13.41	0.049
IMoS [14]	52.290	0.180	8.374	0.152	11.57	20.35	0.000
MDM [60]	26.734	0.289	8.627	0.109	12.96	16.03	0.001
Text2HOI [6]	30.101	0.320	10.302	0.086	12.52	14.55	0.000
OMOMO [28]	25.017	0.391	9.294	0.094	11.03	14.03	0.004
CHOIS [29]	25.835	0.320	9.887	0.055	9.31	14.37	0.002
Ours	21.544	0.438	9.387	0.046	4.93	10.23	0.040

Table 4: **Physical feasibility evaluation on the ARCTIC [13] dataset.**

Metrics	OMOMO [28]	Ours
Tracking Duration ↑	4.75	8.75

Table 5: **Comparison on the ARCTIC [13] dataset with held-out box object.**

Methods	FID↓	R-Precision↑	Diversity→	Foot skating↓	IV↓	ID↓	CR↑
OMOMO [28]	44.009	0.547	6.234	0.028	26.92	8.23	0.116
Ours	16.091	0.547	4.964	0.002	26.24	8.56	0.128

BPS. (h) conditioning the hand motion diffusion model on object trajectories. As shown in Table 6, each component contributes to the performance improvement.

5.5 More discussions

Generalization to different object geometry. To further validate generalization to unseen object geometries of the same category, we train the object motion and end-effector trajectory models on the hand-only dataset [85], using 7 training and 3 testing objects. Despite the dataset containing only hand motion, our method successfully generates whole-body motion, as shown in Figure 5.

Various capabilities. Our approach enables various capabilities. First, it allows control over keyframe object poses by setting them as optimization targets for object trajectory generation. Second, it can synthesize whole-body motions that involve simultaneous locomotion and manipulation, even though such combinations are not present in the training dataset [13]. Third, it enables generating whole-body motion guided by hand-only datasets [2], using wrist and fingertip trajectories as optimization targets. We provide more qualitative results in the supplementary material.

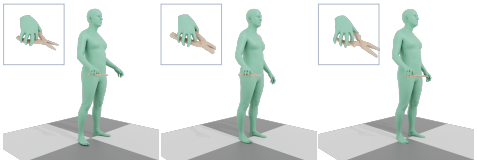


Figure 5: **Generalization to different object geometry.** We train the object motion and end-effector trajectory models on hand-only data [85] with diverse object geometries. These models are integrated into our framework to provide optimization targets, enabling realistic whole-body motion synthesis for unseen object geometry.

Deployment on simulated humanoids As shown in Figure 6, our generated whole-body motion can serve as a reference for controlling humanoids in physics-based simulators. We apply physical motion tracking methods [46, 35, 64] to track the synthesized motions. The humanoid is able to physically interact with objects and perform coordinated manipulation behaviors in the simulated environment.

Inference speed We report the inference time of each module in our pipeline, measured on a single NVIDIA A100 GPU for generating a 300-frame motion sequence. The object motion model

Table 6: **Ablation study on the ARCTIC [13] dataset.**

Methods	FID↓	R-Precision↑	Diversity→	Foot skating↓	IV↓	ID↓	CR↑
Real	–	0.531	8.664	0.002	4.68	11.47	0.085
(a) w/o separate models	3.790	0.438	6.939	0.002	8.21	13.16	0.103
(b) w/o end-effector BPS	4.069	0.453	6.888	0.002	8.09	13.54	0.093
(c) w/o RoPE motion	2.714	0.469	7.021	0.002	6.12	12.66	0.093
(d) w/o optimization	4.883	0.414	6.406	0.030	16.39	16.13	0.095
(e) w/o decoupled	2.699	0.438	7.142	0.008	12.45	16.29	0.082
(f) w/o AMASS	3.305	0.453	6.859	0.003	5.46	13.04	0.089
(g) w/ SDF	2.350	0.422	7.064	0.003	6.69	13.97	0.086
(h) w/ extra condition	2.998	0.453	7.212	0.002	8.95	13.30	0.075
Ours	2.283	0.477	7.208	0.002	5.25	12.87	0.086

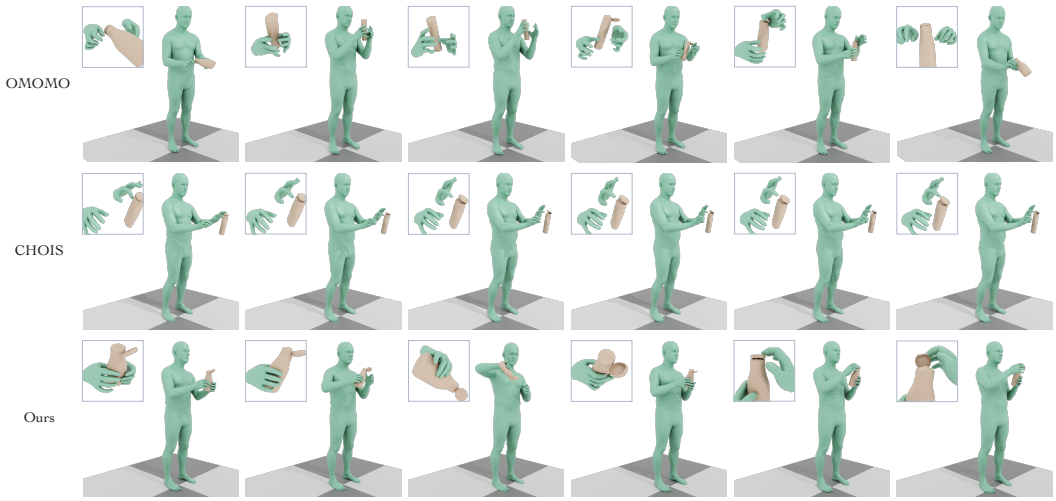


Figure 4: **Qualitative comparison.** Given the text "A person uses the ketchup.", our method generates the whole-body motion with better hand-object contact compared to baselines.

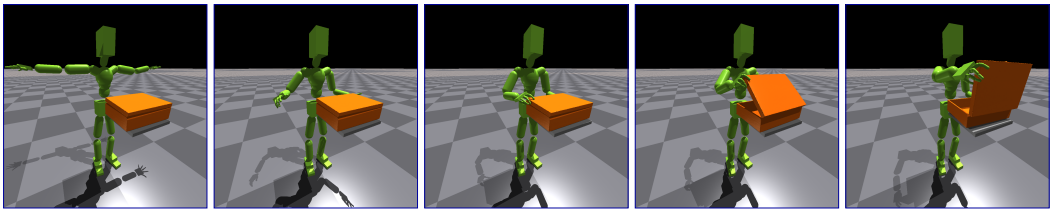


Figure 6: **Deployment on simulated humanoids.** We apply existing motion tracking techniques to deploy the generated motion to a simulated humanoid. The articulated object is physically manipulated by the humanoid within the physics simulator [39].

requires approximately 0.52 seconds, the end-effector model takes about 3.66 seconds, and the whole-body motion optimization, which involves iterative diffusion sampling and gradient-based updates, takes around 16.9 minutes. Most of the computation time is spent on the whole-body optimization stage. Although slower than feed-forward approaches such as CHOIS [29], this optimization process produces motions with substantially higher quality and physical plausibility.

Limitations. First, the optimization process is slower than other generative methods [60], limiting real-time applications. Second, due to the limited object diversity in existing datasets [13], the model struggles to generalize to novel object categories. Third, our framework only focuses on single-object manipulation; extending it to handle multiple interacting objects or multi-step sequential interactions remains an open direction. Finally, enabling both the body and fingers to reason about and avoid obstacles in complex scenes, such as surrounding geometry or other objects, is still a difficult problem.

6 Conclusion

In this paper, we present a coordinated diffusion noise optimization framework for synthesizing whole-body manipulation of articulated objects. By optimizing over the noise space of separately trained diffusion models for the body, left hand, and right hand, our method enables natural coordination between the body and hands. We introduce a unified distance-based representation built on basis point sets to generate end-effector trajectories, facilitating precise hand-object interactions. Extensive experiments demonstrate that our approach achieves state-of-the-art performance in motion quality and physical plausibility. It also supports various capabilities such as object pose control, simultaneous manipulation and locomotion, and whole-body motion generation from hand-only data.

7 Acknowledgments

This work is partly supported by the Innovation and Technology Commission of the HKSAR Government under the ITSP-Platform grant (Ref: ITS/335/23FP) and the InnoHK initiative (TransGP project). Part of the research was conducted in the JC STEM Lab of Robotics for Soft Materials, funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *CVPR*, 2023. 3

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025. 3, 9

[3] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 3

[4] Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Zhu Shuai, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *CVPR*, 2024. 2, 3

[5] 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

[6] 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. 2, 3, 4, 7, 8

[7] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 3

[8] Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2, 3, 7

[9] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: Language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024. 3

[10] Peishan Cong, Ziyi Wang, Yuexin Ma, and Xiangyu Yue. Semgeomo: Dynamic contextual human motion generation with semantic and geometric guidance. *arXiv preprint arXiv:2503.01291*, 2025. 3

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 3

[12] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 3

[13] 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. 2, 3, 4, 6, 7, 8, 9, 10

[14] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, 2023. 3, 4, 7, 8

[15] 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. 7

[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, 3

[17] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023. 3

[18] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022. 3

[19] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv e-prints*, art. arXiv:2207.12598, July 2022. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 7

[21] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, Kris Kitani, Matt Feiszli, and Hao Tang. Hoigpt: Learning long sequence hand-object interaction with language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, 2025. 2, 3

[22] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *CVPR*, 2023. 3

[23] Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction, 2024. URL https://arxiv.org/abs/2410.03187. 3

[24] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 3

[25] 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, 6, 7, 8

[26] 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. 5

[27] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024. 3

[28] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 2, 3, 5, 6, 7, 8, 9

[29] 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. 2, 3, 7, 8, 10

[30] Quanzhou Li, Jingbo Wang, Chen Change Loy, and Bo Dai. Task-oriented human-object interactions generation with implicit neural representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3035–3044, 2024. 3

[31] Yitang Li, Mingxian Lin, Zhuo Lin, Yipeng Deng, Yue Cao, and Li Yi. Learning physics-based full-body human reaching and grasping from brief walking references. *arXiv preprint arXiv:2503.07481*, 2025. 3

[32] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 2020. 7

[33] Xueyi Liu and Li Yi. Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[34] Jintao Lu, He Zhang, Yuting Ye, Takaaki Shiratori, Sebastian Starke, and Taku Komura. Choice: Coordinated human-object interaction in cluttered environments for pick-and-place actions. *arXiv preprint arXiv:2412.06702*, 2024. 3

[35] 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. 9

[36] 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. 3

[37] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. Himo: A new benchmark for full-body human interacting with multiple objects, 2024. URL https://arxiv.org/abs/2407.12371. 3

[38] 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, 6, 7, 8

[39] 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. 8, 10

[40] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, pages 903–913. IEEE, 2024. 3

[41] OpenAI. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt, 2022. 3

[42] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, pages 1498–1507. IEEE, 2024. 3

[43] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *CVPR*, 2025. 3

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 4

[45] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023. 3

[46] 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. 8, 9

[47] Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, and Hujun Bao. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *ICCV*, 2023. 2, 3, 7

[48] Huaijin Pi, Ruoxi Guo, Zehong Shen, Qing Shuai, Zechen Hu, Zhumei Wang, Yajiao Dong, Ruizhen Hu, Taku Komura, Sida Peng, et al. Motion-2-to-3: Leveraging 2d motion data to boost 3d motion generation. *arXiv preprint arXiv:2412.13111*, 2024. 2

[49] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4332–4341, 2019. 2, 5

[50] 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. 5

[51] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015. 7

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP. 7

[53] Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 811–820, June 2024. 3

[54] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 2019. 2, 3

[55] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5

[56] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 2, 3, 4, 6, 7, 8

[57] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, 2022. 3

[58] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. GRIP: Generating interaction poses using latent consistency and spatial cues. In *International Conference on 3D Vision (3DV)*, 2024. URL https://grip.is.tue.mpg.de. 3

[59] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024. 3

[60] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023. 5, 7, 8, 10

[61] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *CVPR*, 2021. 3

[62] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *CVPR*, 2022. 3

[63] Wenjia Wang, Liang Pan, Zhiyang Dou, Zhouyingcheng Liao, Yuke Lou, Lei Yang, Jingbo Wang, and Taku Komura. Sims: Simulating human-scene interactions with real world script planning. *arXiv preprint arXiv:2411.19921*, 2024. 3

[64] Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*, 2023. 3, 8, 9

[65] Yinhuai Wang, Qihan Zhao, Runyi Yu, Ailing Zeng, Jing Lin, Zhengyi Luo, Hok Wai Tsui, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning reusable basketball skills from demonstrations. *arXiv preprint arXiv:2408.15270*, 2024. 3

[66] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *ECCV*, 2022. 3

[67] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. *arXiv preprint arXiv:2406.17840*, 2024. 3

[68] 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

[69] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vCnDyQkjg. 3

[70] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023. 2, 3

[71] Sirui Xu, Yu-Xiong Wang, Liangyan Gui, et al. Interdreamer: Zero-shot text to 3d dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37:52858–52890, 2024. 3

[72] Mengqing Xue, Yifei Liu, Ling Guo, Shaoli Huang, and Changxing Ding. Guiding human-object interactions with rich geometry and relations. *arXiv preprint arXiv:2503.20172*, 2025. 3

[73] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[74] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2025. 2, 3

[75] Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. *arXiv preprint arXiv:2505.02094*, 2025. 3

[76] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 445–456, June 2024. 3

[77] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Trans. Graph.*, 2021. 3, 8

[78] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *2024 International Conference on 3D Vision (3DV)*, pages 235–246. IEEE, 2024. 3

[79] 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. 3

[80] Wanyue Zhang, Rishabh Dabral, Thomas Leimkühler, Vladislav Golyanik, Marc Habermann, and Christian Theobalt. Roam: Robust and object-aware motion generation using neural pose descriptors. In *2024 International Conference on 3D Vision (3DV)*, pages 1392–1402. IEEE, 2024. 3

[81] 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. 2, 3, 5, 7

[82] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *ECCV*, 2022. 3

[83] Yonghao Zhang, Qiang He, Yanguang Wan, Yinda Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. Diffgrasp: Whole-body grasping synthesis guided by object motion using a diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10320–10328, 2025. 3

[84] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *ICCV*, 2023. 3

[85] 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. 3, 9

[86] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *ECCV*, 2022. 3

[87] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3

[88] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 6

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: To jointly generate body, hand, and articulated object motion, we introduce a coordinated diffusion noise optimization framework equipped with a unified BPS-based representation. We conduct experiments on public datasets and show various results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation is discussed in the Section 5.5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not include theoretical results in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe our method in detail and we will release the code and models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our code will be released. While the code is not included in the submission, the implementation details are provided in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Experimental details are provided in Section 5.1. More details are provided in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report our results by averaging over 10 runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are provided in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are discussed in the supplementary material.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release models that have a high risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We do not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The user study instructions are provided in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We conducted a user study in which participants were asked to compare and rate generated motion sequences. The study posed minimal risk and no personal information was collected.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not use LLMs in our method.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.