FORCE: Physics-aware Human-object Interaction

Xiaohan Zhang[1,2], Bharat Lal Bhatnagar[3], Sebastian Starke[3], Ilya Petrov[1,2],
Vladimir Guzov[1,2], Helisa Dhamo[4], Eduardo Pérez-Pellitero[4], Gerard Pons-Moll[1,2]

[1]Tübingen AI Center, University of Tübingen
[2]Max Planck Institute for Informatics, Saarland Informatics Campus
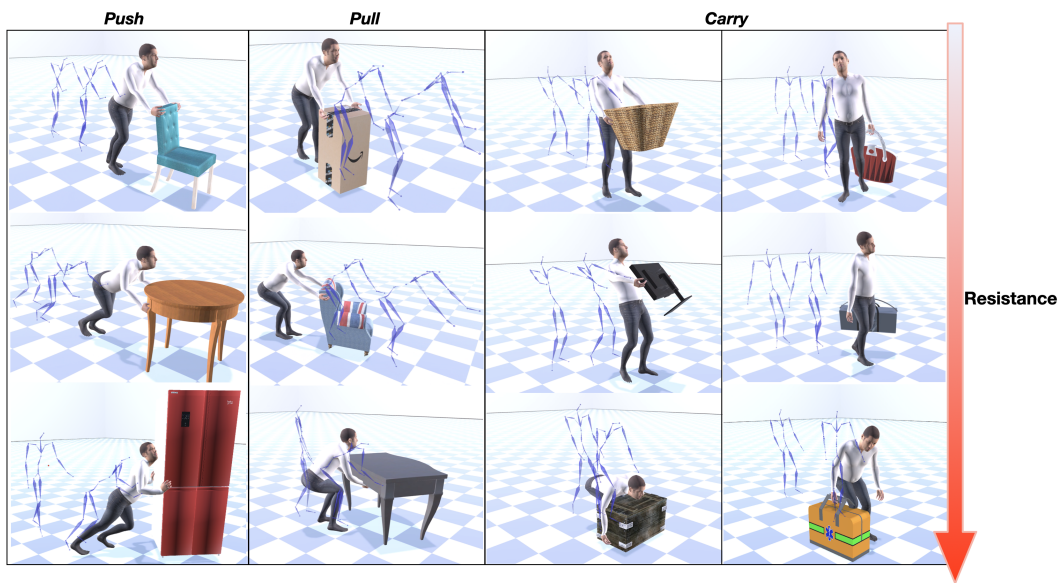[3]Meta Reality Labs Research
[4]Huawei Noah's Ark Lab

Figure 1: *FORCE synthesizes diverse, nuanced human-object interactions by modeling the primary physical attributes.*

Abstract

Interactions between human and objects are influenced not only by the object's pose and shape, but also by physical attributes such as object mass and surface friction. They introduce important motion nuances that are essential for diversity and realism. Despite advancements in recent human-object interaction methods, this aspect has been overlooked. Generating nuanced human motion presents two challenges. First, it is non-trivial to learn from multimodal human and object information derived from both the physical and non-physical attributes. Second, there exists no dataset capturing nuanced human interactions with objects of varying physical properties, hampering model development. This work addresses the gap by introducing the FORCE model, an approach for synthesizing diverse, nuanced human-object interactions by modeling physical attributes. Our key insight is that human motion is dictated by the interrelation between the force exerted by the human and the perceived resistance. Guided by a novel intuitive physics encoding, the model captures the interplay between human force and resistance. Experiments also demonstrate incorporating human force facilitates learning multi-class motion. Accompanying our model, we contribute a dataset, which features diverse, different-styled motion through interactions with varying resistances.

1. Introduction

Synthesizing realistic motion of human-object interactions (HOI) presents a significant challenge. The complexity lies in the intricate interplay between humans and objects. While previous work has primarily addressed fundamental aspects of interactions, such as object shape and location,

they overlook crucial physical attributes like weight, friction, and human force [30, 94, 97, 132]. Consequently, existing models struggle with distinguishing between carrying an empty suitcase and a full one, or determining if the interaction is feasible. These nuances are essential for diversity and realism. This work aims to fill this gap by leveraging the physical attributes to synthesize nuanced human motion in these scenarios.

As a parallel paradigm, physics-based methods using reinforcement learning coupled with imitation learning [24, 82, 83, 114] demonstrate remarkable generalization for different scenarios. However, they face challenges in providing direct fine-grained control, for example, the switching between left-handed carrying to using both hands. Additionally, these methods are computationally demanding, because distinctive policies need to be trained with tailored reward functions for navigation and interaction tasks.

Kinematic methods for human motion synthesis are known for their flexibility for fine-grained spatial or temporal motion control. They are scalable to train and easy to deploy. However, earlier approaches either disregard the surrounding environment [60, 69, 70, 80] or focus solely on static objects [30, 74, 97, 111, 131, 132]. Closest to our goal are methods [30, 36, 56, 94] that model object shapes, but overlook physical attributes of interaction. In reality, humans adapt their motion based on perceived resistance and applied force during an object interaction [25, 71]. Take the *pushing* action shown in Figure 1 (left-most column). When pushing around a heavier object, the human applies a greater force, naturally adjusting posture, shifting the center of mass [12, 101], and leaning forward to push against the friction. If the resistance exceeds the applied force, the object would not move and the human would have to give up the interaction because it is not feasible. Synthesizing such nuanced motion calls for a method that generalizes well to these physical attributes of the interaction.

Developing such a method presents multiple challenges. Firstly, it is non-trivial to reason about multi-modal human-object information, including different actions, object poses, and crucial physical attributes. The increased complexity makes it challenging to disambiguate similar human poses, leading to motion that lacks diversity and nuanced details. Secondly, when determining the feasibility of the interaction, resistance is not the sole factor. It also depends on how the human interacts. For instance, a human can more effectively carry a heavier object with two hands than with one. It is shown that naively conditioning on the resistance generates sub-optimal results (see Section 5). Additionally, there is no dataset capturing diverse daily interactions under varying physical conditions. The lack of such data hinders model development and evaluation. Challenges present even when collecting such data, when addressing issues such as object occlusion.

To address these challenges, we introduce FORCE, a method for synthesizing human-object interactions that focuses on the nuanced details in human motion. Our method is founded on a pivotal insight: Human motion is dictated by the interrelation between the force exerted by the human and the perceived resistance. Guided by a novel intuitive physics encoding derived from these crucial attributes, our model is able to synthesize a diverse spectrum of interactions. For example, under the category of "carrying", our model can plausibly generate motions such as carrying an object, carrying an object followed by a need to drop it, or attempting to carry but encountering failure. Additionally, we enable interactive control at run-time, the style of motion can be manipulated not only by varying the resistance of the object, but also by the desired action and contact mode (left, right, two-handed).

Complementing our FORCE model, we present a dataset featuring diverse motion nuances through interactions involving 3-6 levels of resistance. We adopt a customized hybrid tracker comprising of 4 Kinect RGB-D cameras [1] paired with 17 Inertial Measurement Units (IMUs) [79]. Our novel dataset contains 450 motion sequences (192k frames) of pervasive interactions of carrying, pushing and pulling objects. For each frame, we provide high-quality human and object poses. This dataset can serve as a benchmark for various human-object interaction tasks. The overall contributions of our work are:

1. We introduce FORCE, the first kinematic method to synthesize human-object interaction by modeling physical attributes such as resistance and the applied human force. It achieves state-of-the-art results quantitatively and qualitatively.
2. To enable the synthesis of diverse, nuanced human motion, we propose a novel intuitive physics encoding.
3. We present a dataset that accurately captures the daily interactions of pushing, pulling and carrying objects. It features diverse, different-styled interaction motions with varying resistances.
4. We will release our code, dataset, and models to stimulate further research.

2. Related Work

Human Motion Synthesis. Human motion synthesis has been a long-standing challenge in computer vision, evolving from early non-contextual approaches [2, 4, 13, 34, 48, 52, 84, 91, 135], to more sophisticated methods. Recent advancements include text-to-motion [5, 14, 20, 35, 39, 49, 62, 67, 68, 90, 100, 128, 129], motion-to-text [43, 134], motion synthesis with spatial [23, 46, 102, 113] or temporal control [86], audio to motion [16, 72]. Other directions explore human-to-human interaction [27, 65] and learning 3D human motions from images and videos [117, 120, 121].

In the context of human-object interactions, research

has progressed from predicting static affordances in 3D scenes [31, 32, 57, 61, 88, 118, 130, 133, 137] to dynamic human motions. However, most existing work focuses on human interactions with static scenes [18, 31, 38, 51, 54, 74, 104, 105, 107], while some studies have attempted to enhance motion generation quality by concentrating on static object interactions, such as sitting and lying on furniture [30, 50, 78, 87, 122, 131, 132, 138].

A parallel paradigm of research explores reinforcement learning (RL) for object manipulation [8, 9, 17], carrying objects [33, 73, 103, 114], and performing locomotion [63, 89]. RL has also been applied to interactions with static objects [78, 112] sports movements [106] and text to motion [19]. Another avenue of exploration centers around full-body grasping synthesis [3, 58, 97, 99, 111] and simultaneous synthesis of human and object motions [26, 59, 64, 92]. Specialized efforts are dedicated to synthesizing dexterous hand-object interactions using object motion [85, 98, 126, 140].

While these approaches have made significant strides, they primarily deal with static object interactions or specific types of movements. Our work addresses more challenging scenarios involving humans interacting and moving objects.

Interactions with moving objects. Recent datasets have stimulated development in human-object interaction synthesis, capturing single-person [7, 40, 44, 45, 47, 56, 136] and multi-person [124, 127] with dynamic objects. Among the current methods for synthesizing dynamic human-object interactions, the Neural State Machine [94] can model both static and dynamic interactions. Diffusion models [22, 81, 110, 115, 116, 119] are applied to synthesize short-term human-object interactions with a fixed set of objects [7], often assuming stable contact between humans and objects. More recent work [55, 56] conditions on object trajectories or waypoints to predict full-body object manipulation.

However, these previous datasets and methods overlook how human motion is dictated by the physical attributes of the interaction, such as mass or friction. Our work addresses this gap by capturing a dataset that focuses on interactions with objects of varying resistance, including pushing, pulling, and carrying. We emphasize the nuanced details and motion diversity resulting from these properties.

Intuitive physics guided learning. The study of intuitive physic [6, 71, 108] has experienced a resurgence with recent attempts to incorporate human-level intuitive physics capabilities into deep learning frameworks. Some work focuses on inferring intrinsic physical properties of objects such as mass or spring [25, 139] while others predict underlying object dynamics such as the trajectory [75, 76], forces [21, 28, 42, 53, 109], or physical stability [28, 42, 53, 109].

Applications of intuitive physics have expanded to 3D computer vision, including recovering object trajectories by modeling gravity [21], and guiding 3D human pose estima-
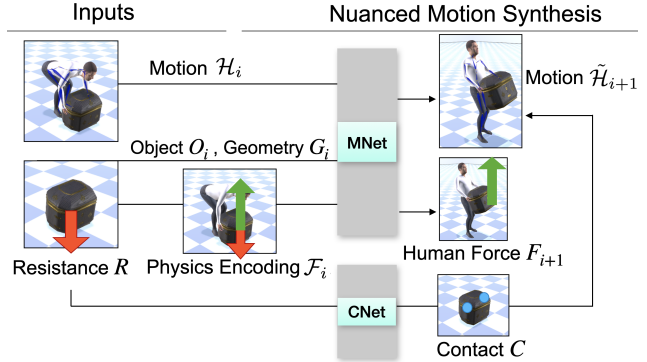

Figure 2. Given the input human and object poses and the object geometry, our method which consists of *MNet* and *CNet*, synthesizes auto-repressively a diverse spectrum of nuanced interactions.

tion with center of mass and center of pressure [101].

Our work draws key insights from intuitive physics studies, modeling the intricate interplay between dynamic human forces and object resistance, by explicitly incorporating these physical attributes to generate more realistic and diverse human-object interactions.

3. Method

Our goal is to synthesize diverse, nuanced human-object interactions by modeling the physical attributes of resistance and the applied human force. The style of motion can be manipulated not only by varying the resistance, but also by the desired action and contact mode (e.g., left, right, two-handed).

3.1. Key Insight

Learning simultaneously from multi-modal human-object information, including the object pose, different modes of action and contact, can result in motions lacking in diversity and nuanced details. However, the intricate interplay between the force exerted by the human and the perceived resistance can help disambiguate between similar poses and facilities learning from multi-class motions. There are two advantages to utilizing these two physical attributes. a) Human force is a 3D representation that informs the model to which direction the force is applied, which dictates the human motion. b) Given the applied force, the resistance further refines the human motion. For instance, the human leans the body to compensate for the increased resistance.

Founded on this key insight, our method (Figure 2) utilizes a physics-aware motion network, *MNet* (Section 3.2) to auto-regressively synthesize diverse, nuanced interaction motion that is plausible to the physical conditions. More specifically, our method takes as input the initial human pose $\mathcal{H}_0$, the object pose $\mathbf{O}_0$ and its geometry $\mathbf{G}_0$, it auto-regressively predicts the future body motion $\mathcal{H}_{i>0}$. Notably, its physics awareness is achieved by our novel physics en-

coding. To ensure the motion is physically plausible, we use another neural network, *CNet* (Section 3.3) to predict the hand contact positions on the object surface.

3.2. MNet: Physics-aware Motion Prediction

At current frame i, given the poses of the human $\mathcal{H}_i$ and the object $\mathbf{O}_i$, the object geometry $\mathbf{G}_i$, *MNet* leverages the interrelation between the applied human force and the perceived resistance to auto-regressively synthesize the future human motion $\tilde{\mathcal{H}}_{i+1}$. At the core of *MNet*, is the intuitive physics encoding $\mathcal{F}_i$, which encapsulates the crucial physics attributes of the interaction.

Intuitive Physics Encoding $\mathcal{F}_i$. Denoted as $\mathcal{F}_i = \{\mathbf{F}_i, \mathbf{R}, \mathbf{c}_i\}$, the encoding comprises the 3D human force $\mathbf{F}_i \in \mathbb{R}^{\mathcal{T} \times 3}$, the magnitude of resistance $\mathbf{R} \in \mathbb{R}$, and binary hand contact labels $C \in \{0,1\}^{\mathcal{T} \times 2}$. The control signal is designed according to previous work [37, 94, 96, 125, 132], the signals are from the $\tau = 13$ uniformly sampled frames within the temporal window $[max(0, i-30), i+30]$. This sampling ensures the model's contextual awareness, allowing it to capture the temporal nuances of the interaction. Next, we define the rest of the auto-regressive inputs:

Human motion $\mathcal{H}_i$: At frame i, the human motion comprises of the pose and the root trajectory: $\mathcal{H}_i = \{\mathbf{J}_i, \mathbf{T}_i\}$. The human pose $\mathbf{J}_i = (\boldsymbol{j}_i^p, \boldsymbol{j}_i^v, \boldsymbol{j}_i^r)$ contains root-relative joint positions $\boldsymbol{j}_i^p \in \mathbb{R}^{J \times 3}$, velocities $\boldsymbol{j}_i^v \in \mathbb{R}^{J \times 3}$, and rotations $\boldsymbol{j}_i^r \in \mathbb{R}^{J \times 6}$ (forward and upward vectors of the rotation matrix) for the $J = 22$ joints in SMPL skeleton [66]. The root trajectory $\mathbf{T}_i = (\boldsymbol{t}_i^p, \boldsymbol{t}_i^d, \boldsymbol{t}_i^a)$ encodes the root position $\boldsymbol{t}_i^p \in \mathbb{R}^{\mathcal{T} \times 2}$ and direction $\boldsymbol{t}_i^d \in \mathbb{R}^{\mathcal{T} \times 2}$ projected onto the ground. $\boldsymbol{t}_i^a \in [0,1]^{\mathcal{T} \times 5}$ describes the current action (*idle, walk, carry, push, pull*). We use continuous values between 0 and 1 to facilitate the transition between actions.

Object $\mathbf{O}_i$: The target object is denoted as $\mathbf{O}_i = (\boldsymbol{o}_i^p, \boldsymbol{o}_i^d, \boldsymbol{o}_i^a)$, where $\boldsymbol{o}_i^p \in \mathbb{R}^{\mathcal{T} \times 3}$, $\boldsymbol{o}_i^d \in \mathbb{R}^{\mathcal{T} \times 6}$ are its positions and orientations relative to the root. $\boldsymbol{o}_i^a \in \{0,1\}^{\mathcal{T} \times 5}$ are the binary variables describing the *desired future action*. For example, for carrying, the target action is "carry" when approaching the object.

Geometry $\mathbf{G}_i$: The object geometry $\mathbf{G}_i$ is encoded by voxelizing object shape (in an $8 \times 8 \times 8$ dimensional grid). Each voxel stores its occupancy ($\mathbb{R}$) and the relative vector between the human root joint and the voxel ($\mathbb{R}^3$). This enables reasoning about the shape of the object. We vectorize this grid to obtain our geometry encoding $\mathbf{G}_i \in \mathbb{R}^{2048}$.

Phase $\boldsymbol{\Phi}_i$: We introduce a learned variable, phase $\boldsymbol{\Phi}_i \in \mathbb{R}^4$, encoding the human joint trajectories following [93]. It encapsulates the spatial-temporal context of the motion.

Training. The *MNet* adopts a mixture-of-experts architecture [30, 37, 94, 95, 131, 132]. It is trained with direct su-

pervision by minimizing the MSE loss on the outputs:

$$\{\tilde{\mathcal{H}}_{i+1}, \mathbf{F}_{i+1}, \mathbf{O}_{i+1}, \boldsymbol{\Phi}_{i+1}\} = f^{\text{MNet}}(\mathcal{H}_i, \mathcal{F}_i, \mathbf{O}_i, \mathbf{G}_i, \boldsymbol{\Phi}_i), \quad (1)$$

where $\tilde{\mathcal{H}}_{i+1} = \{\mathbf{J}_{i+1}, \mathbf{T}_{i+1}, \tilde{\boldsymbol{j}}_{i+1}^p, \tilde{\mathbf{T}}_{i+1}\}$ denotes the future human motion. Here, $\mathbf{J}_{i+1}$, $\mathbf{T}_{i+1}$ are the future body joints and the root trajectory. $\mathbf{F}_{i+1}$ denotes the future 3D human forces. $\tilde{\boldsymbol{j}}_{i+1}^p$ and $\tilde{\mathbf{T}}_{i+1}$ are joint positions and root trajectory relative to the object. Supervising with these object-centric signals ensures the human reaches the target. $\mathbf{O}_{i+1}$ and $\boldsymbol{\Phi}_{i+1}$ are the object pose, and phase at frame $i+1$. More details of the network architecture and training can be found in the supplementary.

Derivation of Human Force. For training, the human force is not directly observable. However, it can be estimated from the observable object acceleration. a) When the object moves, we leverage Newton's laws to compute the human force. We assume the direction of force acts opposite to the direction of the object acceleration. For example, for *pushing* and *pulling*, the force is derived as $m \cdot g \cdot \eta \cdot \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} + m \cdot \boldsymbol{a}$, where m, g, η, $\boldsymbol{a}$ denote the mass, gravity, frictional coefficient and object acceleration. b) When the object is in contact while remaining static, we derive the force by linearly interpolating between the zero, and the derived force from the nearest frame at which the object moves. c) When the resistance is too great for a successful interaction, we leverage the maximal observed human force for each action class as the control signal. Note, to eliminate the effect of noise of object acceleration, we apply a Butterworth filter of kernel size 4.

3.3. Resistance-conditioned Contact Prediction

Contacts are imperative to ensure the physical plausibility of the human-object interaction, and the way the human contacts with an object is influenced by its resistance. For example, when pushing a chair when the resistance is low, the human tends to push through the top of the chair since it is more efficient. However, when the resistance of the chair increases, the human tends to hold onto the lower part of the chair to stabilize to push more efficiently. Inspired by this and the previous work on predicting hand contact conditioned on the object geometry [132], we use a variational auto-encoder *CNet* to predict the hand contact positions, $\mathcal{C} \in \mathbb{R}^{2 \times 3}$, on the object surface. Formally, the contacts are predicted via $\hat{\mathcal{C}} = f^{\text{CNet}}(\mathbf{R}, \mathbf{G}_i, \boldsymbol{j}_i^p, \boldsymbol{o}_i^a)$. Here, $\mathbf{R}$, $\boldsymbol{j}_i^p$, $\boldsymbol{o}_i^a$ are the resistive force, human joint positions and the target actions as defined previously. The object geometry $\mathbf{G}_i$ in this case is encoded relative to the center of the object instead of the human's root. During training, the network is trained to minimize the following loss,

$$L_{\text{CNet}} = \|\hat{\mathcal{C}} - \mathcal{C}\|_2^2 + \lambda KL(q(\boldsymbol{z}|\vec{c}, \mathbf{R}, \mathbf{G}, \boldsymbol{j}_i^p, \boldsymbol{o}_i^a)\|p(\boldsymbol{z})), \quad (2)$$
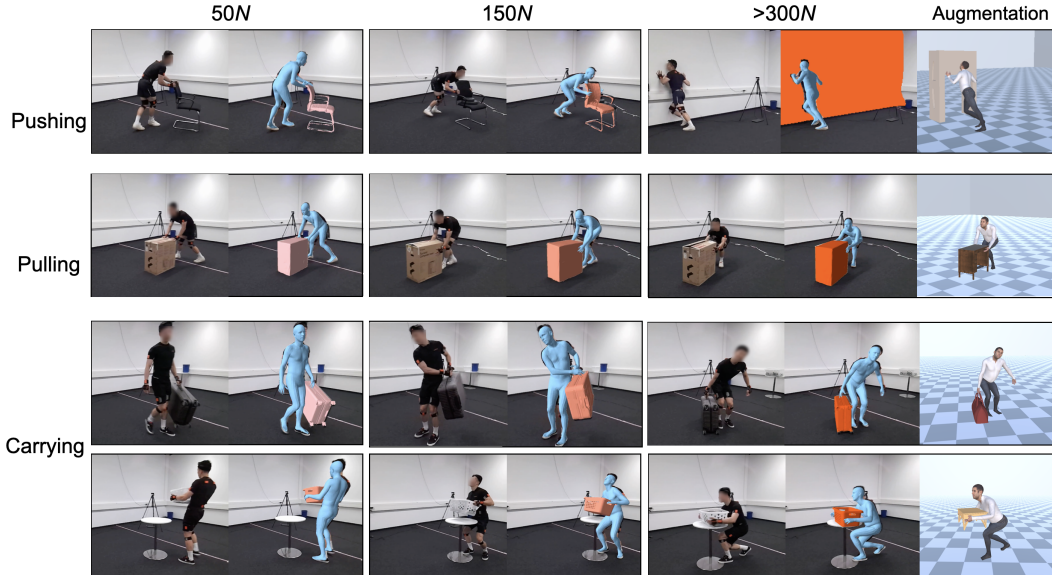
Figure 3. The FORCE dataset accurately captures diverse, nuanced interaction motion under varying levels of resistance (N for Newtons). The last column shows results from the object shape augmentation (see Section 4), where nuanced motion details are preserved.

where KL denotes the Kullback-Leibler divergence. For training, *CNet* is trained with ground truth contact positions derived from Section 4. For "one-handed" contact, the contact position for the other hand is zeroed. During inference, the contacts are sampled once the human is within the vicinity of the object, and is remained constant for the rest of the motion sequence. We apply a real-time inverse kinematics [30, 94, 131, 132] to ensure the contact constraints are satisfied. We optimize the pose of the arm joints to minimize the distance between the linearly interpolated position between the current hand joint and contact. The weight of the interpolation is inversely proportional to the distance to ensure the smoothness and realism of the hand motion.

4. Dataset

We introduce a dataset that comprises 450 motion sequences of human pushing, pulling and carrying. Its main characteristic is the diverse motion nuances through interactions with varying resistances. It includes 8 objects, captured with 3-6 levels of resistance (from 5kg to 35kg) with removable weights. The testing set consists of 30 sequences covering a range of interactions and scenarios.

Human and object tracking. We leverage a customized tracker, integrating 17 human-mounted IMU [79] sensors with the cameras. This enhancement improves the accuracy of the captured data. After initially fitting the SMPL[66] to the captured point clouds, we synchronize the tracking results with motion sequences recorded by IMU sensors. Subsequently, we optimize the body poses from the IMU tracking to align with the Kinect-fitted result.
For objects, we initiate the tracking by pre-scanning eight

objects (carry: bin, suitcase, stool, basket, container, backpack; push/pull: chair, box) to obtain their template meshes. Subsequently, we fit them to annotated keypoints on the captured images using camera projection. The fitting is refined by running ICP to fit to the segmented object point cloud [15]. We consider our objects to be rigid.
Further technical details about our tracking methodology are available in the supplementary materials.

Diverse interactions with varying resistance. The capture covers diverse motion variation. Firstly, we vary the resistance of the object in each interaction instance. The same motion is performed for each object of the same action category. Secondly, the dataset captures different contact modes including one-handed and two-handed interactions. We systematically place them in varying positions relative to the human. To ensure the authenticity of the motion, only high-level guidance is provided instead of specific instructions. Figure 3 showcases a selection of interactions.
Note, that our data collection is performed on a uniform surface, and we determine the frictional coefficient by measuring force meter readings during the uniform pulling of a 10kg object. For consistency, we approximate gravity (g) and the frictional coefficient (η) as 10N/kg and 1, respectively.

Object shape augmentation. To generalize to unseen shapes, each motion sequence is augmented with 10 differently shaped objects from one of the table, chair, bag, and shelf categories of ShapeNet [10] (see the last column of Figure 3). First, we detect the ground truth hand contact positions on the object surface based on a distance threshold of 0.05cm. Next, we transfer contacts by aligning the
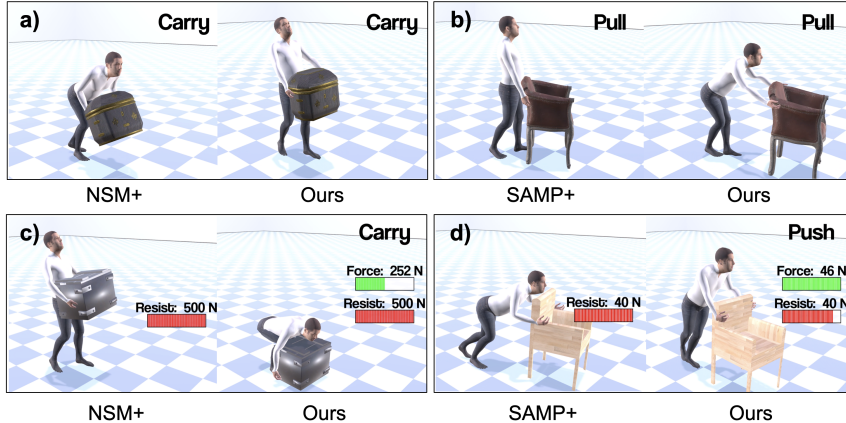
Figure 4. Comparison with baselines. Our *intuitive physics encoding* $\mathcal{F}$ disambiguates different multi-class motions and enables the synthesis of the desired interaction (carrying in **a**, pulling in **b**). **c)** and **d)** show the coupled encoding of human force and resistance allows the synthesis of visually plausible interactions. In **c)**, when the resistance is greater than the human force, FORCE synthesizes a more plausible "infeasible" interaction. In **d)**, with low resistance, the human pose maintains an upright position.

new object with the source via centering and scaling. The ground truth contacts are projected onto the nearest surface of the new object. At every frame, following object sampling and rescaling, the human poses are recomputed to satisfy the contacts using a full-body inverse kinematics [94]. We assume the resistance remains unchanged after augmentation to preserve the context of human motion.

5. Experiments

For evaluation, we first introduce the baselines (Section 5.1). We then compare with the baselines quantitatively (Section 5.2) and qualitatively (Section 5.3). Last, we perform ablations on the designs of our method (Section 5.4) and show results on generalization (Section 5.5).

5.1. Baselines

To our knowledge, the most related works to ours are NSM [94] and SAMP [30] as they both enable real-time generation human-object interaction with a user-controller. We also compare our method with InterDiff [115], a state-of-the-art method that leverages the diffusion model.

We train the baseline methods on our captured dataset using their publicly released codes. These methods neglect the physics attributes of the interaction. To measure the effectiveness of our intuitive physics encoding, we adapt NSM and SAMP by additionally conditioning on the resistive force and the binary hand contact. We refer to these modified baselines as NSM+ and SAMP+. Moreover, we evaluate two ablative variations of FORCE. One model is without the physics encoding $\mathcal{F}$, and the other one is without the contact prediction.

5.2. Quantitative Evaluation

We quantitatively evaluate our model performance with two setups. **a)** offline motion synthesis. This is performed on

Table 1. Quantitative comparisons for offline motion synthesis. Results evaluated on the testing sequences.

Method	MPJPE(cm)$^\downarrow$	Non-collision(%)$^\uparrow$	Foot-slide(cm)$^\downarrow$
NSM [94]	7.71	64.3	4.59
SAMP [30]	8.67	57.7	5.79
InterDiff [115]	8.62	75.3	6.82
NSM+	7.08	69.6	4.71
SAMP+	8.34	58.3	6.02
Ours no Phys. Enc. $\mathcal{F}$	6.75	75.4	4.60
Ours no Contact Pred.	-	77.2	4.48
Ours no **R**	7.10	76.4	4.76
Ours	**6.02**	**84.0**	**4.46**

Table 2. Quantitative comparisons for online full interaction synthesis. Note, diversity on the dataset is *0.906*. The closeness to this figure indicates a better result.

Method	Diversity$^\rightarrow$	Success Rate(%)$^\uparrow$	Exec. Time(s)$^\downarrow$
NSM [94]	0.707	75.1	5.94
SAMP [30]	0.719	66.2	7.72
NSM+	0.785	79.6	6.01
SAMP+	0.776	68.7	7.56
Ours no Phys. Enc. $\mathcal{F}$	0.847	86.7	5.81
Ours no Contact Pred.	0.886	94.2	5.50
Ours	**0.891**	**97.5**	**5.42**

the testing sequences given the ground truth control signals. We evaluate the accuracy (MPJPE) and motion quality (foot-slide and collision) of motion synthesis on testing sequences following our training/testing split explained in Section 4. The model synthesizes poses auto-regressively given the trajectories. The sequences are each distinctive, covering diverse pushing, pulling and carrying motions at different resistances. **b)** online full synthesis. We synthesize full approaching and interacting in a random setup. For fairness, 10 objects are randomly sampled. They are placed randomly relative to the human (within 6 meters) at a random orientation. The synthesis is performed 120 times for each method, repeating each action type (pushing, pulling, one-handed, and two-handed carrying) 30 times. We mea-
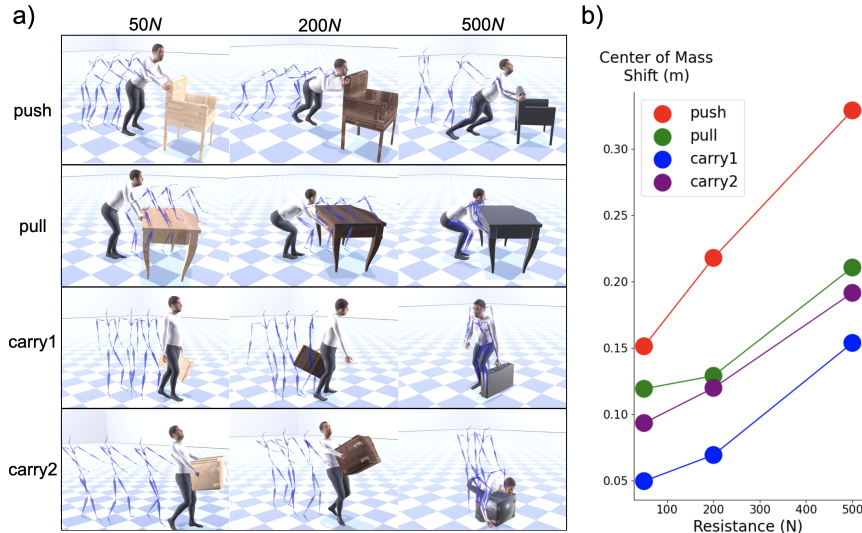
Figure 5. Nuanced interactions with the same object shape, but different resistance. a) The human struggles with the interaction as the resistance increases from 50 N to 200 N. At 500 N, the human fails to interact with the object. b) Plot depicts the horizontal shift of the center of mass of the human, estimated following [12, 101]. Results averaged over 1000 frames of the synthesized interaction.

sure the diversity, the success rate and the execution time. Note, InterDiff [115] does not allow online synthesis, we compare against it for offline synthesis on the testing set.

MPJPE(cm) [41, 80]. The accuracy of human motion synthesis. The result on "Ours no Contact" is omitted because the ablative baseline utilizes the same MNet as "Ours".
Non-Collision Score(%) [32, 130, 133]. The percentage of frames that the body does not penetrate with the object.
Foot-slide(cm) [30, 33, 37, 94, 131, 132]. The average traveled distance by the pivotal foot per step is calculated.
Diversity [29, 30, 123, 133]. We compute the Average Pairwise Distance (APD) on the root-relative joint positions.
Success rate(%) [11, 30, 33]. The percentage of both the desired object interaction and the hand contact is synthesized within 10 seconds.
Execution time [32, 77]. Time till a successful interaction.

From Table 1, it can be seen that FORCE attains the lowest MPJPE of 6.02 cm, surpassing both baseline models and ablative variations of FORCE . This emphasizes the crucial role of our intuitive physics encoding in facilitating the learning of multi-class motions. Additionally, our method minimizes object interpenetration, as reflected in the highest non-collision scores. It is notable that, compared with the ablative baseline without contact prediction, our method has a higher non-collision score. This indicates that with contact prediction, our model synthesizes more physically plausible motion. Furthermore, our model demonstrates improved motion quality, evident in its reduced foot-slide. From Table 2, it can be seen under the online synthesis setup, FORCE has the highest success rate of interaction of 97.5% and the shortest average execution time of 5.42 sec-

onds, illustrating the better controllability of our approach. Our model synthesizes motion of diversity (0.891) which is the closest match to the training data distribution (0.906). This underscores the significance of the physics encoding in enabling the synthesis of diverse, nuanced interactions.

5.3. Qualitative Evaluation

Figure 4 illustrates our qualitative comparison with the baselines NSM+ [94] and SAMP+ [30]. In Figure 4 a), the poses when lifting up an object and pushing an object are similar, introducing ambiguity between these two classes of motions. However, the physics encoding informs the model about the crucial directional information of the human force, dictating the model to synthesize the desired carrying motion. Similarly in Figure 4 b), our model synthesizes the desired pulling motion. The physics encoding also provides awareness of the feasibility given the resistance, because of the coupled encoding of human force and resistances. In Figure 4 c), when the resistance is greater than the exerted human force, FORCE is aware that the interaction is no longer feasible. In Figure 4 d), our synthesis is more visually plausible given the low resistance, with the body pose remaining upright.

5.4. Ablations

Physics encoding $\mathcal{F}$. As shown in Figure 5, FORCE synthesizes nuanced human-object interactions with objects of the same shape but varying resistances. Empowered by the intuitive physics encoding $\mathcal{F}$, our method synthesizes diverse interactions. It generalizes not only to different resistances of the object but also to different actions and contacts. The graph on the right provides a zoomed-in analy-

sis of the motion, where we evaluate the horizontal shift of the center of mass of the human (for each 1000-frame sequence). The center of mass of the human is estimated following [12, 101]. As shown in the plot, there is a positive correlation between the perceived resistance of the human during the interaction and the shift of center of mass.
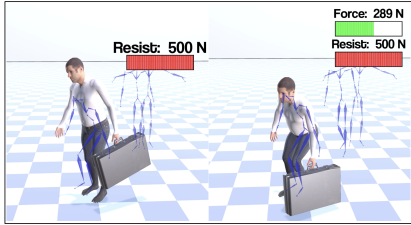

Figure 6. Without Physics Encoding (left) vs. FORCE (right). With physics encoding, synthesizes visually plausible motion, when the resistance is greater than the applied human force.
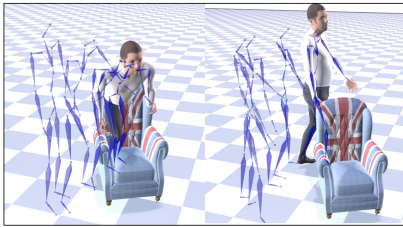
Figure 7. Without geometry representation **G** (left) vs. FORCE (right). The feature enables the reasoning about collision.

Figure 6 also emphasizes the importance of intuitive physics encoding. On the left, without the physics encoding, the model overlooks the resistance surpassing the applied force of the human. With the physics encoding on the right, our method is aware that the resistance is not feasible, as the perceived resistance reaches $500\ N$ and is greater than the applied human force.

Resistance $\mathcal{R}$. As shown in table 1, without conditioning on the resistance, the motion quality degrades.

Geometry representation G. With the representation introduced in Section 3.2, our method reasons about the collision as the voxels provide coarse geometry details. As a result, it exhibits less interpenetration (Figure 7).
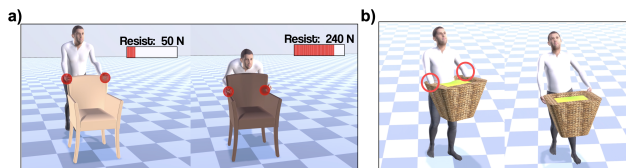

Figure 8. a) FORCE predicts different hand contacts at different resistances for the same pushing motion. b) No contact prediction (left) vs. FORCE (right). FORCE exhibits less penetration.

Contact prediction. As shown Figure 8 a), different contact positions are predicted by conditioning on different re-

sistances using our method. It can be concluded that the contact depends on the resistance. Here, as the resistance increases, the human tends to push from the lower part of the chair to push more efficiently. Figure 8 b) shows, with the contact prediction, FORCE synthesizes motion that satisfies better the hand contact constraints on the surface, and exhibits less interpenetration (see Table 1).

5.5. Generalization to shapes and locations.

As illustrated in Figure 9, FORCE generalizes to different shapes. Thanks to the object shape augmentation introduced in Section 4, our model is able to synthesize lifting a chair, an object category that is not seen during training of carrying sequences. It is also highlighted in Figure 9 that FORCE interacts with objects at different locations, for example, picking up from different heights.

Figure 9. (Top) FORCE generalizes to unseen shapes, demonstrated via pushing, carrying, and pulling. (Bottom) FORCE generalizes to different locations, performing one-handed and two-handed carry from the table and from the floor.

6. Conclusion

This work, FORCE, tackles the problem of synthesizing human object interactions with nuanced details, by modeling intuitive physics. It solves two main challenges. First, it proposes the first kinematic-based method that synthesizes human-object interaction conditioned on physical attributes such as object resistance and human force. By leveraging the novel physics encoding, our method generates a diverse spectrum of interaction motions under varying resistance. Its performance in diversity and realism surpasses previous methods. Our accompanying dataset features over 450 motion sequences. Together, they provide a benchmark for training and evaluating HOI methods.

For future work, this work paved the way for research in nuanced human-object interaction in more complex scenarios. First, while our model demonstrates its efficacy in generating diverse human-object interactions, there is an opportunity for expansion by considering a broader range of subjects, because each individual possesses unique strengths and approaches to interactions. Second, interaction can be extended to complex scenarios of dynamic resistance. For example, when carrying a tank of water, the resistance fluctuates and influences human motion.

References

[1] Microsoft azure kinect. https://en.wikipedia.org/wiki/Azure_Kinect/ 2

[2] Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Trans. Graph. **42**(4) (2023) 2

[3] Araújo, J.P., Li, J., Vetrivel, K., Agarwal, R., Wu, J., Gopinath, D., Clegg, A., Liu, C.K.: CIRCLE: capture in rich contextual environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR (2023) 3

[4] Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 2

[5] Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings (2024) 2

[6] Battaglia, P.W., Pascanu, R., Lai, M., Rezende, D.J., Kavukcuoglu, K.: Interaction networks for learning about objects, relations and physics. In: Advances in Neural Information Processing Systems 2016 3

[7] Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022) 3

[8] Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis 3

[9] Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: International Conference on 3D Vision (3DV 2024) (2024) 3

[10] Chang, A.X., Funkhouser, T.A., Guibas, L.J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3d model repository. CoRR **abs/1512.03012** (2015) 5

[11] Chao, Y., Yang, J., Chen, W., Deng, J.: Learning to sit: Synthesizing human-chair interactions via hierarchical control. CoRR **abs/1908.07423** (2019) 7

[12] Chebel, E., Tunc, B.: Evaluation of center of mass estimation for obese using statically equivalent serial chain. Scientific Report **12** (2022) 2, 7, 8

[13] Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 2

[14] Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2

[15] Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022) 5

[16] Chhatre, K., Daněček, R., Athanasiou, N., Becherini, G., Peters, C., Black, M.J., Bolkart, T.: AMUSE: Emotional speech-driven 3D body animation via disentangled latent diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2

[17] Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 3

[18] Cong, P., Wang, Z., Dou, Z., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment (2024) 3

[19] Cui, J., Liu, T., Liu, N., Yang, Y., Zhu, Y., Huang, S.: Anyskill: Learning open-vocabulary physical skill for interactive agents. In: Conference on Computer Vision and Pattern Recognition(CVPR), year=2024 3

[20] Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2023) 2

[21] Dabral, R., Shimada, S., Jain, A., Theobalt, C., Golyanik, V.: Gravity-aware monocular 3d human-object reconstruction. In: International Conference on Computer Vision (ICCV) (2021) 3

[22] Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation (2024) 3

[23] Diomataris, M., Athanasiou, N., Taheri, O., Wang, X., Hilliges, O., Black, M.J.: WANDR: Intention-guided human motion generation. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 2

[24] Dou, Z., Chen, X., Fan, Q., Komura, T., Wang, W.: C·ase: Learning conditional adversarial skill embeddings for physics-based characters. CoRR **abs/2309.11351** (2023) 2

[25] Duan, J., Dasgupta, A., Fischer, J., Tan, C.: A survey on machine learning approaches for modelling intuitive physics. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelli-

gence, IJCAI 2022, Vienna, Austria, 23-29 July 2022 2, 3

[26] Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Imos: Intent-driven full-body motion synthesis for human-object interactions. In: Eurographics (2023) 3

[27] Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In: European Conference on Computer Vision (ECCV) (2024) 2

[28] Groth, O., Fuchs, F.B., Posner, I., Vedaldi, A.: Shapestacks: Learning vision-based physical intuition for generalised object stacking. In: European Conference on Computer Vision - ECCV 2018 3

[29] Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 7

[30] Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.: Stochastic scene-aware motion prediction. In: Proceedings of the International Conference on Computer Vision 2021 (2021) 2, 3, 4, 5, 6, 7

[31] Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: International Conference on Computer Vision (Oct 2019) 3

[32] Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021) (2021) 3, 7

[33] Hassan, M., Guo, Y., Wang, T., Black, M.J., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. CoRR **abs/2302.00883** (2023) 3, 7

[34] Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: probabilistic and controllable motion synthesis using normalising flows. ACM Trans. Graph. **39**(6) (2020) 2

[35] Hoang, N.M., Gong, K., Guo, C., Mi, M.B.: Motionmix: Weakly-supervised diffusion for controllable motion generation. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth Conference on Artificial Intelligence, AAAI 2024 2

[36] Holden, D., Kanoun, O., Perepichka, M., Popa, T.: Learned motion matching. ACM Trans. Graph. **39**(4) (2020) 2

[37] Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Trans. Graph. **36**(4) (2017) 4, 7

[38] Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3

[39] Huang, Y., Wan, W., Yang, Y., Callison-Burch, C., Yatskar, M., Liu, L.: Como: Controllable motion generation through language guided pose code editing (2024) 2

[40] Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction. In: German Conference on Pattern Recognition (GCPR) (2022) 3

[41] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (jul 2014) 7

[42] Janner, M., Levine, S., Freeman, W.T., Tenenbaum, J.B., Finn, C., Wu, J.: Reasoning about physical interactions with object-oriented prediction and planning. In: International Conference on Learning Representations, ICLR 2019 3

[43] Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems (2024) 2

[44] Jiang, N., Liu, T., Cao, Z., Cui, J., Chen, Y., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: ICCV (2023) 3

[45] Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3

[46] Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 2

[47] Kim, J., Kim, J., Na, J., Joo, H.: Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions (2024) 3

[48] Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. arXiv preprint arXiv:2209.00349 (2022) 2

[49] Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, 2

[50] Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: Nifty: Neu-

ral object interaction fields for guided human motion synthesis (2023) 3

[51] Lee, J., Joo, H.: Locomotion-action-manipulation: Synthesizing human-scene interactions in complex 3d environments. arXiv preprint arXiv:2301.02667 (2023) 3

[52] Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. In: Williams, B., Chen, Y., Neville, J. (eds.) Thirty-Seventh Conference on Artificial Intelligence, AAAI 2023 2

[53] Lerer, A., Gross, S., Fergus, R.: Learning physical intuition of block towers by example. In: Balcan, M., Weinberger, K.Q. (eds.) International Conference on Machine Learning, ICML 2016 3

[54] Li, G., Zhao, K., Zhang, S., Lyu, X., Dusmanu, M., Zhang, Y., Pollefeys, M., Tang, S.: EgoGen: An Egocentric Synthetic Data Generator. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3

[55] Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis 3

[56] Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics **42**(6) (Dec 2023) 2, 3

[57] Li, L., Dai, A.: GenZI: Zero-shot 3D human-scene interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3

[58] Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024 3

[59] Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations (2023) 3

[60] Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++ (2021) 2

[61] Li, X., Liu, S., Kim, K., Wang, X., Yang, M.H., Kautz, J.: Putting humans in a scene: Learning affordance in 3d indoor environments. In: IEEE Conference on Computer Vision and Pattern Recognition (2019) 3

[62] Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision (2024) 2

[63] Ling, H.Y., Zinno, F., Cheng, G., van de Panne, M.: Character controllers using motion vaes. ACM Trans. Graph. **39**(4) (2020) 3

[64] Liu, X., Yi, L.: Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. In: The Twelfth International Conference on Learning Representations (2024) 3

[65] Liu, Y., Chen, C., Yi, L.: Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. CoRR **abs/2312.08983** (2023) 2

[66] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34** (Oct 2015) 4, 5

[67] Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Humantomato: Text-aligned whole-body motion generation. arxiv:2310.12978 (2023) 2

[68] Ma, S., Cao, Q., Zhang, J., Tao, D.: Contact-aware human motion generation from textual descriptions. arXiv preprint arXiv:2403.15709 (2024) 2

[69] Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 2

[70] Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017 2

[71] McCloskey, M., Washburn, A., Felch, L.: Intuitive physics: The straight-down belief and its origin. Journal of experimental psychology. Learning, memory, and cognition **9**, 636–49 (10 1983) 2, 3

[72] Mehta, S., Deichler, A., O'Regan, J., Moëll, B., Beskow, J., Henter, G.E., Alexanderson, S.: Fake it to make it: Using synthetic data to remedy the data shortage in joint multimodal speech-and-gesture synthesis (2024) 2

[73] Merel, J., Tunyasuvunakool, S., Ahuja, A., Tassa, Y., Hasenclever, L., Pham, V., Erez, T., Wayne, G., Heess, N.: Catch & carry: reusable neural controllers for vision-guided whole-body tasks. ACM Trans. Graph. (2020) 3

[74] Mir, A., Puig, X., Kanazawa, A., Pons-Moll, G.: Generating continual human motion in diverse 3d scenes. In: International Conference on 3D Vision (3DV) (March 2024) 2, 3

[75] Mottaghi, R., Bagherinezhad, H., Rastegari, M., Farhadi, A.: Newtonian image understanding: Unfolding the dynamics of objects in static images. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016 3

[76] Mottaghi, R., Rastegari, M., Gupta, A., Farhadi, A.: "what happens if..." learning to predict the effect of forces in images. In: European Conference on Computer Vision - ECCV 2016 3

[77] Pan, L., Wang, J., Huang, B., Zhang, J., Wang, H., Tang, X., Wang, Y.: Synthesizing physically plausible human motions in 3d scenes (2023) 7

[78] Pan, L., Wang, J., Huang, B., Zhang, J., Wang, H., Tang, X., Wang, Y.: Synthesizing physically plausible human motions in 3d scenes. In: International Conference on 3D Vision (3DV) (2024) 3

[79] Paulich, M., Schepers, M., Rudigkeit, N., Bellusci, G.: Xsens mtw awinda: Miniature wireless inertial-magnetic motion tracker for highly accurate 3d kinematic applications (2018) 2, 5

[80] Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. In: British Machine Vision Conference (BMVC) (2018) 2, 7

[81] Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. arXiv preprint arXiv:2312.06553 (2023) 3

[82] Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. ACM Trans. Graph. (2022) 2

[83] Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: AMP: adversarial motion priors for stylized physics-based character control. ACM Trans. Graph. (2021) 2

[84] Pérez, G.V., Henter, G.E., Beskow, J., Holzapfel, A., Oudeyer, P., Alexanderson, S.: Transflower: probabilistic autoregressive dance generation with multimodal attention. ACM Trans. Graph. **40**(6) (2021) 2

[85] Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 3

[86] Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Peng, X.B., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: CVPR Workshop on Human Motion Generation (2024) 2

[87] Pi, H., Peng, S., Yang, M., Zhou, X., Bao, H.: Hierarchical generation of human-object interactions with diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2023) 3

[88] Qi, S., Zhu, Y., Huang, S., Jiang, C., Zhu, S.C.: Human-centric indoor scene synthesis using stochastic grammar. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3

[89] Rempe, D., Luo, Z., Peng, X.B., Yuan, Y., Kitani, K., Kreis, K., Fidler, S., Litany, O.: Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3

[90] Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023) 2

[91] Shi, Y., Wang, J., Jiang, X., Dai, B.: Controllable motion diffusion model. CoRR **abs/2306.00416**, https://doi.org/10.48550/arXiv.2306.00416 2

[92] Shimada, S., Mueller, F., Bednarik, J., Doosti, B., Bickel, B., Tang, D., Golyanik, V., Taylor, J., Theobalt, C., Beeler, T.: Macs: Mass conditioned 3d hand and object motion synthesis. In: International Conference on 3D Vision (3DV) (2024) 3

[93] Starke, S., Mason, I., Komura, T.: Deepphase: periodic autoencoders for learning motion phase manifolds. ACM Trans. Graph. (2022) 4

[94] Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Trans. Graph. **38**(6) (2019) 2, 3, 4, 5, 6, 7

[95] Starke, S., zhao, Y., Komura, T., Zaman, K.A.: Local motion phases for learning multi-contact character movements. ACM Trans. Graph. 4

[96] Starke, S., zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. ACM Trans. Graph. 4

[97] Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: GOAL: Generating 4D whole-body motion for hand-object grasping. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 2, 3

[98] Taheri, O., Zhou, Y., Tzionas, D., Zhou, Y., Ceylan, D., Pirk, S., Black, M.J.: Grip: Generating interaction poses conditioned on object and body motion. In: International Conference on 3D Vision (3DV 2024) (2024) 3

[99] Tendulkar, P., Surís, D., Vondrick, C.: Flex: Full-body grasping without full-body grasps. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 3

[100] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023) 2

[101] Tripathi, S., Müller, L., Huang, C.H.P., Omid, T., Black, M.J., Tzionas, D.: 3D human pose estimation via intuitive physics. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 3, 7, 8

[102] Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. CoRR **abs/2311.17135** (2023) 2

[103] Wan, W., Yang, L., Liu, L., Zhang, Z., Jia, R., Choi, Y.K., Pan, J., Theobalt, C., Komura, T., Wang, W.: Learn to predict how humans manipulate large-sized objects from interactive motions. IEEE Robotics and Automation Letters (2022) 3

[104] Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2021) 3

[105] Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., Dai, B.: Towards diverse and natural scene-aware 3d human motion synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE 3

[106] Wang, Y., Lin, J., Zeng, A., Luo, Z., Zhang, J., Zhang, L.: Physhoi: Physics-based imitation of dynamic human-object interaction. arXiv preprint arXiv:2312.04393 (2023) 3

[107] Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. In: Advances in Neural Information Processing Systems (NeurIPS) (2022) 3

[108] Watters, N., Zoran, D., Weber, T., Battaglia, P.W., Pascanu, R., Tacchetti, A.: Visual interaction networks: Learning a physics simulator from video. In: Advances in Neural Information Processing Systems 2017 3

[109] Wu, J., Lu, E., Kohli, P., Freeman, B., Tenenbaum, J.: Learning to see physics via visual de-animation. In: Advances in Neural Information Processing Systems 2017 3

[110] Wu, Q., Shi, Y., Huang, X., Yu, J., Xu, L., Wang, J.: THOR: text to human-object interaction diffusion via relation intervention. CoRR **abs/2403.11208** (2024) 3

[111] Wu, Y., Wang, J., Zhang, Y., Zhang, S., Hilliges, O., Yu, F., Tang, S.: Saga: Stochastic whole-body grasping with contact. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022) 2, 3

[112] Xiao, Z., Wang, T., Wang, J., Cao, J., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. Arxiv (2023) 3

[113] Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. In: The Twelfth International Conference on Learning Representations (2024) 2

[114] Xie, Z., Tseng, J., Starke, S., van de Panne, M., Liu, C.K.: Hierarchical planning and control for box locomanipulation. CoRR **abs/2306.09532** 2, 3

[115] Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023) 3, 6, 7

[116] Xu, S., Wang, Z., Wang, Y.X., Gui, L.Y.: Interdreamer: Zero-shot text to 3d dynamic human-object interaction. arXiv preprint arXiv:2403.19652 (2024) 3

[117] Xu, X., Joo, H., Mori, G., Savva, M.: D3d-hoi: Dynamic 3d human-object interactions from videos. arXiv preprint arXiv:2108.08420 (2021) 2

[118] Xuan, H., Li, X., Zhang, J., Hongwen Zhang, Y.L., Li, K.: Narrator: Towards natural control of human-scene interaction generation via relationship reasoning. In: ICCV (2023) 3

[119] Yang, J., Niu, X., Jiang, N., Zhang, R., Siyuan, H.: F-hoi: Toward fine-grained semantic-aligned 3d human-object interactions. European Conference on Computer Vision (2024) 3

[120] Yang, Y., Zhai, W., Luo, H., Cao, Y., Zha, Z.J.: Lemon: Learning 3d human-object interaction relation from 2d images (2024) 2

[121] Ye, Y., Li, X., Gupta, A., Mello, S.D., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: CVPR (2023) 2

[122] Yi, H., Thies, J., Black, M.J., Peng, X.B., Rempe, D.: Generating human interaction motions in scenes with text control. arXiv:2404.10685 (2024) 3

[123] Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 7

[124] Zhang, C., Liu, Y., Xing, R., Tang, B., Yi, L.: Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. arXiv preprint arXiv:2406.19353 (2024) 3

[125] Zhang, H., Starke, S., Komura, T., Saito, J.: Mode-adaptive neural networks for quadruped motion control. ACM Trans. Graph. **37**(4) (2018) 4

[126] Zhang, H., Ye, Y., Shiratori, T., Komura, T.: Manipnet: neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph. (2021) 3

[127] Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., Yu, J., Xu, L., Wang, J.: Hoi-m3: Capture multiple humans and objects interaction within contextual environment. arXiv preprint arXiv:2404.00299 (2024) 3

[128] Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 2

[129] Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: 2023 IEEE/CVF International Conference on Computer Vision, ICCV 2

[130] Zhang, S., Zhang, Y., Ma, Q., Black, M.J., Tang, S.: PLACE: Proximity learning of articulation and contact in 3D environments. In: International Conference on 3D Vision (3DV) (2020) 3, 7

[131] Zhang, W., Dabral, R., Leimkühler, T., Golyanik, V., Habermann, M., Theobalt, C.: Roam: Robust and object-aware motion generation using neural pose descriptors. arXiv preprint arXiv:2308.12969 (2023) 2, 3, 4, 5, 7

[132] Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions. In: European Conference on Computer Vision (ECCV) (October 2022) 2, 3, 4, 5, 7

[133] Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 3, 7

[134] Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators (2023) 2

[135] Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis. In: SIGGRAPH, Technical Papers (2024) 2

[136] Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., Wang, J., Xu, L.: I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 3

[137] Zhao, K., Wang, S., Zhang, Y., Beeler, T., Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: European conference on computer vision (ECCV) (2022) 3

[138] Zhao, K., Zhang, Y., Wang, S., Beeler, T., , Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: International conference on computer vision (ICCV) (2023) 3

[139] Zheng, Q., Wu, W., Pan, H., Mitra, N.J., Cohen-Or, D., Huang, H.: Inferring object properties from human interaction and transferring them to new motions (2021) 3

[140] Zhou, K., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In: European Conference on Computer Vision (ECCV) (2022) 3