HIMO: A New Benchmark for Full-Body Human Interacting with Multiple Objects

Xintao Lv[1*] ⓘ, Liang Xu[1,2*] ⓘ, Yichao Yan[1†] ⓘ, Xin Jin[2] ⓘ, Congsheng Xu[1] ⓘ,
Shuwen Wu[1] ⓘ, Yifan Liu[1] ⓘ, Lincheng Li[3] ⓘ, Mengxiao Bi[3] ⓘ,
Wenjun Zeng[2] ⓘ, and Xiaokang Yang[1] ⓘ

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong
University, Shanghai, China
[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
[3] NetEase Fuxi AI Lab
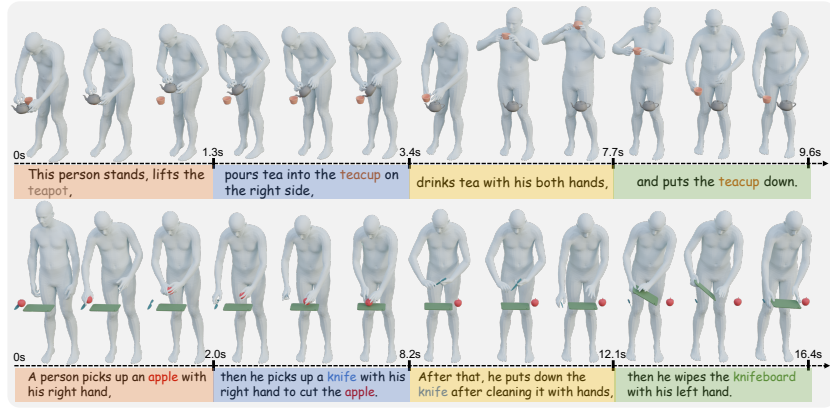https://lvxintao.github.io/himo

Fig. 1: Overview. HIMO is a large-scale dataset of full-body interacting with multiple objects, with fine-grained textual descriptions. The long interaction sequences and the corresponding texts are elaborately segmented temporally, facilitating the detailed timeline control of multi-step human-object interactions.

Abstract. Generating human-object interactions (HOIs) is critical with the tremendous advances of digital avatars. Existing datasets are typically limited to humans interacting with a single object while neglecting the ubiquitous manipulation of multiple objects. Thus, we propose **HIMO**, a large-scale MoCap dataset of full-body human interacting with multiple objects, containing **3.3K** 4D HOI sequences and **4.08M** 3D HOI frames. We also annotate **HIMO** with detailed textual descriptions and temporal segments, benchmarking two novel tasks of HOI synthesis conditioned on either the whole text prompt or the segmented text prompts as fine-grained timeline control. To address these novel tasks, we propose a dual-branch conditional diffusion model with a mutual interaction module for HOI synthesis. Besides, an auto-regressive generation pipeline is also designed to obtain smooth transitions between HOI segments. Experimental results demonstrate the generalization ability to

---

[*]Equal contribution

[†]Corresponding author

unseen object geometries and temporal compositions. Our data, codes, and models will be publicly available for research purposes.

Keywords: Human-Object Interaction · Multiple Objects · Text-driven HOI Synthesis · Temporal Segmentation

1   Introduction

Humans constantly interact with objects as daily routines. As a key component for human-centric vision tasks, the ability to synthesize human-object interactions (HOIs) is fundamental with numerous applications in video games, AR/VR, robotics, and embodied AI. However, most of the previous datasets and models [8, 17, 24, 26, 59] are limited to interacting with a single object, yet neglect the ubiquitous functionality combination of multiple objects. Intuitively, the multiple objects setting is more practical and allows for broader applications, such as manipulating multiple objects for robotics [44, 70].

The scarcity of such datasets mainly results in the underdevelopment of the synthesis of human interacting with multiple objects, as listed in Tab. 1. GRAB [59] builds the dataset of 4D full-body grasping of daily objects ("4D" refers to 3D geometry and temporal streams, "full-body" emphasizes the body movements and dexterous finger motion), followed by several 4D HOI datasets with RGB modality [8, 24], interacting with articulated [17] or sittable [26] objects. For text-driven HOI generation, [15, 46] manually label the textual descriptions for the BEHAVE [8] and CHAIRS [26] datasets, and Li *et al.* [31, 32] collect a dataset of human interacting with daily objects with textual annotations. Despite the significant development, all these datasets focus on single-object interactions. Thus, a dataset of full-body humans interacting with multiple objects incorporated with detailed textual descriptions is highly desired.

Capturing 4D HOIs is challenging due to the subtle finger motions, severe occlusions, and precise tracking of various objects. Compared with a single object, interacting with multiple objects is more complicated regarding the *spatial* movements between human-object and object-object, and the *temporal* schedule of several atomic HOI intervals. For example, the process of a person interacting with a teapot and teacup could be: "Lift the teapot" → "Pour tea to the teacup" → "Drink tea", as shown in Fig. 1. To facilitate the synthesis of **H**uman **I**nteracting with **M**ultiple **O**bjects, we build a large-scale dataset called **HIMO**. We adopt the optical MoCap system to obtain precise body movements and track the motion of objects attached by reflective markers. For the dexterous finger motions, we employ wearable inertial gloves to avoid occlusions. In total, **3.3K** 4D HOI sequences with **34** subjects performing the combinations of **53** daily objects are presented, resulting in **4.08M** 3D HOI frames. Various subjects, object combinations, and interaction patterns also ensure the diversity of **HIMO**.

To facilitate the study of text-driven HOI synthesis [12, 15, 20, 62, 73], we annotate the HOI sequences with fine-grained textual descriptions. Different from existing datasets, we additionally segment the long interaction sequences temporally and align the corresponding texts semantically, which allows for fine-grained

Table 1: Dataset comparisons. We compare the HIMO dataset with existing human-object interaction datasets. "Multi-object", "Segment" and "#" mean human interacting with multiple objects, temporal segmentation, and the number of.

Dataset	Modality				Scale			
	Hand	Multi-object	Text	Segment	#Subject	#Object	#Seq	#Frame
GRAB [59]	✓	✗	✗	✗	10	51	1,334	-
BEHAVE [8]	✗	✗	✗	✗	8	20	321	15K
InterCap [24]	✗	✗	✗	✗	10	10	223	67K
ARCTIC [17]	✓	✗	✗	✗	9	11	339	2.1M
CHAIRS [26]	✓	✗	✗	✗	46	81	1,390	-
Li *et al.* [32]	✓	✗	✓	✗	17	15	-	-
HIMO(Ours)	✓	✓	✓	✓	34	53	3,376	4.08M

timeline control, and a flexible schedule of multiple atomic HOI sequences. Re-visiting the example of pouring water, after training based on our datasets, *novel* HOI compositions such as "Drink water" → "Lift the teapot" → "Add more tea to the teacup" could be synthesized benefited from our fine-grained temporal segments. Although temporal segmentation is widely adopted in video action domain [16,35,38,57], we are the first to introduce the annotations to 4D HOIs.

Our HIMO dataset enables two novel generative tasks: 1) Text-driven HOI synthesis with multiple objects (denoted as `HIMO-Gen`); 2) `HIMO-Gen` conditioned on segmented texts as timeline control (denoted as `HIMO-SegGen`). Concretely, instead of single text as the condition for `HIMO-Gen`, `HIMO-SegGen` requires a series of multiple consecutive texts as guidance. For `HIMO-Gen`, we generate the synchronized human motion and object motions conditioned on the language descriptions and the initial states of the human and the objects. Inspired by previous works on modeling human-human interaction generation [36,56,65,67] and human interacting with single object [31,32,46,64,68], it is straightforward to design a dual-branch conditional diffusion model to generate the human and object motion, respectively. However, this naive solution suffers from the following challenges. First, the generated human motion and object motions could be spatio-temporally misaligned. Second, implausible contacts between human-object and object-object are commonly encountered. To guarantee the coordination between human motion and object motions, we further design a mutual interaction module implemented as a stack of multiple mutual-attention layers. We also add an object-pairwise loss based on the relative distance between the geometry of objects to generate plausible object movements. Experiments show that our method can synthesize realistic and coordinated HOIs.

Furthermore, for the task of `HIMO-SegGen`, each segmented HOI clip paired with the corresponding text can be viewed as a sample to train the `HIMO-Gen` model, while the key of `HIMO-SegGen` lies in the smoothness between the generated clips. We introduce a simple yet effective scheme to auto-regressively generate one HOI clip at a time, conditioning the next clip generation with the last few frames of the previous clip. Experiment results show that conditioning on 10 frames achieves the best performance. Besides, we also show the gener-

alization ability of our model to unseen objects and novel HOI compositions. Overall, our contributions can be summarized as follows:

- We collect a large-scale dataset of full-body human interacting with multiple daily objects called **HIMO**, with precise and diverse 4D-HOI sequences.
- We annotate the detailed textual descriptions for each long HOI sequence, together with the temporal segmentation labels, facilitating two novel downstream tasks of `HIMO-Gen` and `HIMO-SegGen`.
- We propose a dual-branch conditional diffusion model with a mutual interaction module to fuse the human and object features. We also propose a pipeline to auto-regressively generate the composite HOI sequences.

2   Related Work

2.1   Human-Object Interaction Datasets

Many human-object interactions together with the hand-object interaction datasets have been proposed to facilitate the development of modeling HOIs. Several works [10, 21, 22, 25, 39, 40] focusing on hand-object interaction have been proposed. Though effective, modeling hand-object interactions falls significantly short of comprehending the nature of human-object interaction, since the interaction process involves more than just hand participation. Full-body interactions [8, 17, 24, 26, 42, 59, 69, 74, 76] may enhance our understanding of HOIs. GRAB [59], InterCap [24] and ARCTIC [17] broaden the range of 4D-HOIs to full-body interactions. However, all of these full-body datasets only involve one single object, while the combination of several different objects is common in our daily lives. The KIT Whole-body dataset [42] contains motion in which subjects manipulate several objects simultaneously. However, the human body is represented as a humanoid model instead of a more realistic SMPL-X model. We address the deficiency by introducing a full-body 4D-HOI dataset that involves full-body interaction with multiple objects. Comparisons with the existing HOI datasets are listed in Tab. 1.

2.2   Text-driven Motion Generation

Text-driven human motion generation aims at generating human motions based on textual descriptions. Benefiting from large-scale motion-text dataset like KIT-ML [49], BABEL [52] and HumanML3D [20], numerous works [12, 14, 32, 47, 62, 66, 73, 80] have emerged. TEMOS [47] proposes a transformer-based VAE to encode motion and an additional DistillBert [55] to encode texts. MDM [62] adopts the diffusion model [23] for motion generation. MLDM [12] denoises the low-dimensional latent code of motion instead of the raw motion sequences. T2M-GPT [73] utilizes VQ-VAE [43] to compress motion sequences into a learned codebook and then generate human motion in an auto-regressive fashion. GMD [27] further adds keyframe conditions to synthesize controllable human motion. In this work, we base our model on the widely adopted MDM [62] framework.

2.3   Human-Object Interaction Generation

Generating plausible and realistic human-object interaction sequences has attracted much attention in recent years [11, 13, 15, 19, 29, 32, 33, 46, 58–61, 64, 68, 71, 78]. GRAB [59] generates plausible hand poses grasping a given object. GOAL [58] and SAGA [64] extend it to full-body grasping motion generation. However, these methods convert the task into the generation of the final pose of a human interacting with the *static* object. Recent works like InterDiff [68] tackle the problem of synthesizing comprehensive full-body interactions with dynamic objects, which leverages a diffusion model [23] to predict future HOI with realistic contacts. IMoS [19] synthesizes vivid HOIs conditioned on an intent label, such as: "use", "lift" and "pass". Notice that none of the aforementioned methods can generate full-body HOI involving multiple objects.

2.4   Temporal Action Composition

Human motion composition [5, 7, 30, 34, 48, 53, 56, 77] with text timeline control aims to generate arbitrary long human motion sequences with smooth and realistic transitions, which can be separated as two categories. First, several auto-regressive methods [5, 30, 34, 53, 77] are developed relying on the availability of motion transition annotations. These methods iteratively generate the subsequent motion based on the already generated motion. Second, a few diffusion-based methods [6, 7, 48, 56, 75] are proposed to modify the sampling process of diffusion models with overlaps of a predefined transition duration. In our setting of `HIMO-SegGen`, we also need to generate composite HOI sequences with smooth transitions. Thus, we adopt a simple yet effective auto-regressive pipeline to iteratively synthesize the HOI sequences.

3   The HIMO Dataset

We present HIMO, a large-scale 4D-HOI dataset of full-body human interacting with multiple objects, comprising accurate and diverse full-body human motion, object motions and mutual contacts. Next, we will introduce the data acquisition process in Sec. 3.1 and then describe the SMPL-X fitting process and the annotation of textual descriptions and the temporal segments in Sec. 3.2.

3.1   Data Acquisition

Hybrid Capture System. To obtain the accurate movement of the human body against occlusions, we follow [59] to adopt the optical MoCap scheme instead of multi-view RGB camera system [8, 24], which guarantees much lower error [59, 65]. We believe that high-quality moving sequences are more important than natural images for HOI motion data. An overview of our capturing setup can be seen in Fig. 2(a,c). We deploy Optitrack [3] MoCap system with 20 PrimeX-22 infrared cameras to capture the body motion, where each subject
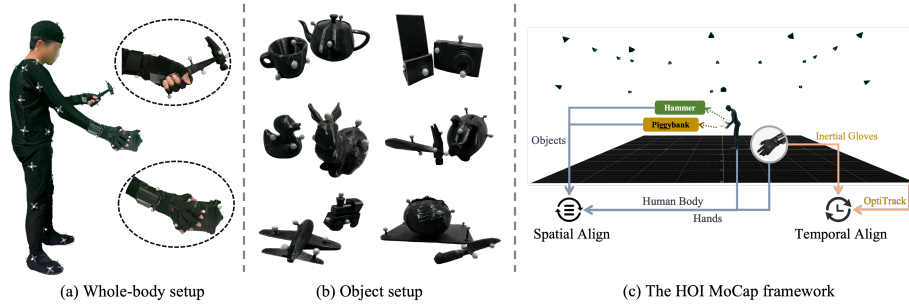
(a) Whole-body setup   (b) Object setup   (c) The HOI MoCap framework

Fig. 2: Overview of the HIMO capture system. (a). We combine the Optitrack MoCap framework and the inertial gloves as a hybrid MoCap system. (b). Some examples of the 3D printed objects and the attached reflective markers. (c). The whole HOI MoCap framework, where the objects, the human body and hands of the subject are spatially aligned, and the Optitrack and inertial gloves are temporally synchronized.

wearing MoCap suits with 41 reflective markers. Each camera can capture 2K frames at 120fps. To capture the dexterous finger movements, we select the inertial Noitom Perception Neuron Studio (PNS) gloves [2], which can precisely record the subtle finger poses in real-time. The inertial gloves can still work well under severe occlusions of human-object and object-object. We frequently re-calibrate the PNS gloves to ensure the capture quality. Spatially, the locating boards attached to the back of the hands provide the rotation information of the wrists, thus the human body movement can be integrated with the finger movements. Temporally, we utilize the Tentacle Sync device to provide timecodes for Optitrack and PNS gloves to synchronize them. Apart from the hybrid MoCap system for full-body motion capture, we also set up a Kinect camera to record the RGB videos from the front view with the resolution of 1080×768.

Capturing Objects. We choose 53 common household objects in various scenes including the dining rooms, kitchens, living rooms and studies from ContactDB [9] and Sketchfab [4]. We further reshape them into appropriate sizes and 3D print them, so that the 3D object geometry is aligned with the real printed objects. To precisely track the object motion, we attach several (3-6) reflective markers on their surface with strong glue, as shown in Fig. 2(b), to track the rigid body composed of the attached markers. The coordinate system of the subject and the objects are naturally aligned. One critical rule of attaching the markers to objects is to mitigate the side effects of manipulating the objects. Empirical results show that 12.5 mm diameter spherical markers achieve more robust tracking performance than smaller markers. Since the Optitrack system tracks the 6-DoF poses of the centroid of the attached markers rather than the centroid of the object, thus we employ a post-calibration process to compensate for the bias between them. The object category definition is listed in the supplementary.

Recording Process. Each subject is asked to manipulate the pre-defined combinations of 2 or 3 objects, such as "pour tea from a *teapot* to a *teacup*" and "cut an *apple* on a *knifeboard* with a *knife*". Initially, the objects are randomly

placed on a table with the height of 74 cm and the subject keeps the rest pose. For each sequence, the subject is required to involve all the provided objects into the manipulation and perform 3 times of each interaction task with different operating patterns for variability. Casual operations yet may not be relevant to the prescribed action are welcome to enhance the complexity of each interaction sequence, such as "shaking the goblet before drinking from it" or "cleaning the knifeboard before cutting food on it". Finally, 34 subjects participated in the collection of HIMO, resulting in **3,376** HOI sequences, **9.44** hours and **4.08M** frames, where 2,486 of them interacting with 2 objects and the other 890 sequences interacting with 3 objects. More details of the setting of the interaction categories can be found in the supplementary materials.

3.2   Data Postprocessing

Fitting SMPL-X Parameters. The expressive SMPL-X [45] parametric model is widely adopted by recent works [18, 37, 41, 48, 72] for its generality and flexible body part editing ability. We also adopt the SMPL-X model to represent the human body with finger articulations. Formally, the SMPL-X parameters consist of global orient $g \in \mathbb{R}^3$, body pose $\theta_b \in \mathbb{R}^{21 \times 3}$, finger poses $\theta_h \in \mathbb{R}^{30 \times 3}$, root translation $t \in \mathbb{R}^3$ and shape parameter $\beta \in \mathbb{R}^{10}$. An optimization algorithm is adopted to obtain the SMPL-X parameters for each HOI sequence based on our full-body MoCap data. We initialize the subjects' shape β based on their height and weight following [51]. The joint energy term $\mathbb{E}_j$ optimizes body joints to our MoCap data as:

$$\mathbb{E}_j = \sum_{n=0}^{N} \sum_{j=0}^{J} \|P_n^j - \hat{P}_n^j\|_2^2, \tag{1}$$

where N is the frame number of the sequence, J is the number of human joints, P_n^j and $\hat{P}_n^j$ represent our MoCap joint position and the joint position regressed from the SMPL-X model, respectively. The smoothing term $\mathbb{E}_s$ smooths the motion between frames and mitigates pose jittering as:

$$\mathbb{E}_s = \sum_{n=0}^{N-1} \sum_{j=0}^{J} \|\hat{P}_{n+1}^j - \hat{P}_n^j\|_2^2. \tag{2}$$

Finally, a regularization term $\mathbb{E}_r$ is adopted to regularize the SMPL-X pose parameters from deviating as:

$$\mathbb{E}_r = \|\theta_b\|_2^2 + \|\theta_h\|_2^2. \tag{3}$$

In total, the whole optimization objective can be summed as follows:

$$\mathbb{E} = \alpha \mathbb{E}_j + \lambda \mathbb{E}_s + \gamma \mathbb{E}_r, \tag{4}$$

where $\alpha = 1, \lambda = 0.1, \gamma = 0.01$.
Textual Annotations. Text-driven motion generation thrives greatly thanks to the emergence of text-motion datasets [20, 49, 63]. To empower the research on
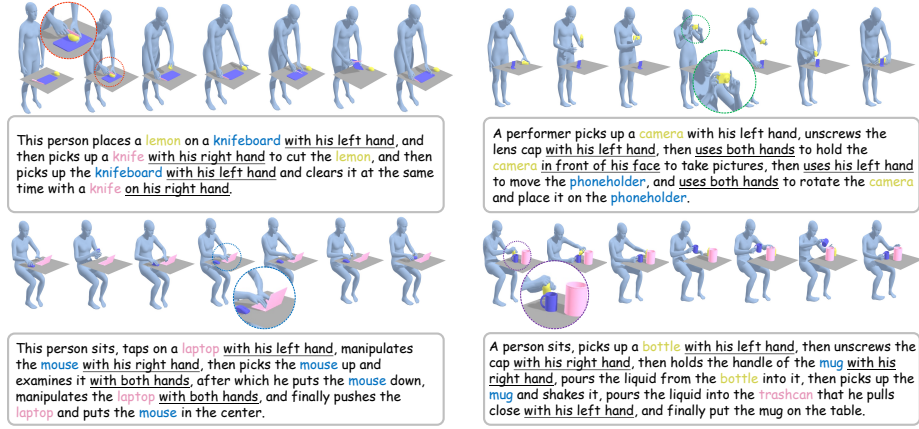
Fig. 3: More samples of the HIMO dataset. The subject is asked to interact with 2 or 3 objects based on their functionalities. The textual annotations comprise the detailed interaction order, the interaction pattern and the involved human body parts. Object names are depicted as the corresponding colors and the details are underlined.

text-driven HOI synthesis with multiple objects, *i.e.*, HIMO-Gen, we elaborately annotate fine-grained textual descriptions of each motion sequence. Different from simple HOI manipulations with a single object and a single prompt, our HIMO dataset includes complex transitions among objects and temporal arrangements among different objects. Thus, the annotators are asked to describe the entire interaction procedure, emphasizing the detailed manipulation order, *i.e.*, "First..., then..., finally...", the interaction pattern, *i.e.*, "picking up", "rotating" and "pouring", and the involved human body parts, *i.e.*, the left hand or right hand. We exploit an annotation tool based on [1] to enable the annotators to rotate and resize the view to observe the subtle HOIs patterns. More visualization results of the text-HOI pairs are presented in Fig. 3.

Temporal Segment Annotations. As aforementioned, the temporal segments of the long HOI sequences allow for fine-grained timeline control of decomposing the complex operation process into several simple ones. The granularity of the sequence splitting is small, as shown in Fig. 1. For example, the sequence of "pouring tea from the teapot to the teacup" can be split into "lifting the teapot", "pouring tea into the teacup" and "putting the teacup down". Similarly, we implement an annotation tool by [1] to better view the interaction details. The HOI sequence and its corresponding texts are simultaneously split into equal segments. Statistically, each sequence contains 2.60 segments on average.

4   Text-driven HOI Synthesis

In this section, we benchmark two *novel* tasks of HIMO-Gen in Sec. 4.1 and HIMO-SegGen in Sec. 4.2. For HIMO-Gen, we propose a dual-branch diffusion-based framework to synthesize the motion of human and objects separately, with
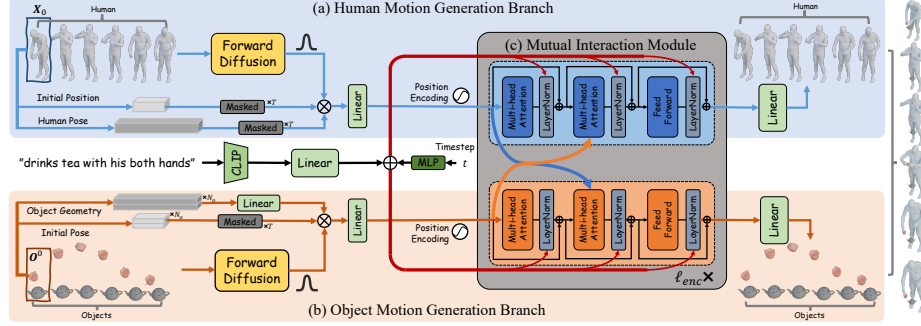
Fig. 4: Overview of the HOI synthesis framework. We propose a dual-branch conditional diffusion model to generate the human motion and object motions, respectively, conditioned on the textual description and the initial states of the human and objects. A mutual interaction module is also integrated for information fusion to generate coordinated HOI results.

meticulously designed mutual interaction module and optimization objectives. For `HIMO-SegGen`, we apply an auto-regressive generation scheme.

4.1 The `HIMO-Gen` Framework

Data Representation. We denote the human motion as $\boldsymbol{H} \in \mathbb{R}^{T \times D_h}$ and the multiple object motions as $\boldsymbol{O} = \{\boldsymbol{O}_i\}_{i=0}^{N_o} \in \mathbb{R}^{T \times D_o}$, where T represents the frame number of the HOI sequence, N_o is the number of the objects, D_h and D_o represent the dimension of the human and object motion, respectively. We adopt the SMPL-X [45] parametric model to represent the human movements. The pose state of human at frame i is denoted as $\boldsymbol{H}^i$, which consists of the global joint positions $P^i \in \mathbb{R}^{52 \times 3}$, global joint rotation $Q^i \in \mathbb{R}^{52 \times 6}$ represented by the continuous 6D rotation format [79] and translation $t^i \in \mathbb{R}^3$. We also denote the object motion at frame i as $\boldsymbol{O}^i$ (We omit the subscript of the i-th object/geometry for better expression), which comprises of the relative rotation $\mathbf{R}^i \in \mathbb{R}^6$ with respect to the input object's frame and its global translation $\mathbf{T}^i \in \mathbb{R}^3$. To encode the object geometries $\boldsymbol{G} = \{\boldsymbol{G}_i\}_{i=0}^{N_o}$, we follow prior works [31,59] to adopt the Basis Point Set (BPS) representation [50]. We sample 1,024 points from the surface of the object meshes, and then for each sampled point, we calculate the directional vector with its nearest neighbor from the basis points set, resulting in $\boldsymbol{G} \in \mathbb{R}^{1024 \times 3}$ for each object.

Problem Formulation. Given the textual description $\boldsymbol{L}$, the initial states of $\boldsymbol{H}^0$ and $\boldsymbol{O}^0$ as the *first* frame human motion and object motions and the object geometries $\boldsymbol{G}$. Our model aims to generate the spatio-temporally aligned human motion $\boldsymbol{H}$ together with the object motions $\boldsymbol{O}$ with plausible contacts.

Dual-branch Diffusion Models. Our proposed framework is demonstrated in Fig. 4, which consists of two parallel diffusion-based branches for human and object motion generation and a mutual interaction module for feature fusion.

Given a textual description $\boldsymbol{L}$ like "drinks tea with his both hands", we utilize the CLIP [54] as text encoder followed by a linear layer to obtain the semantic embeddings and then concatenate them with the features of the sampling timestep t extracted by an MLP process. Take the human motion generation branch as an example, the initial state $\boldsymbol{H}^0$ including the initial position and the human pose serves as the conditions. Following [31], we adopt the masked representation by zero padding $\boldsymbol{H}_0$ to T frames. The masked representations are then concatenated with the noised human motion representation $\boldsymbol{H}_t$, where t is the timestep of the forward diffusion process. For the object motion generation branch, the object geometry $\boldsymbol{G}$ is embedded by a linear layer and concatenated with the masked representation of the initial pose of objects and the noised representation of object motions. We separately feed the obtained human and object representations into a linear layer to obtain the motion embeddings and then pass them into the denoising mutual interaction module with positional encodings. After that, two linear layers are adopted to obtain the denoised motion representations of human and objects, respectively.

Mutual Interaction Module. To model the interaction between human and objects, we fuse the features of the two branches via a mutual interaction module as depicted in Fig. 4(c). The two branches share the same Transformer architecture with ℓ_{enc} blocks without sharing weights. Each Transformer block is comprised of two attention layers and one feed-forward layer. The first self-attention layer embeds the aggregated human/object features $\boldsymbol{H}^{(i)}$ and $\boldsymbol{O}^{(i)}$ into embeddings $\mathbf{e}_h^{(i)}$ or $\mathbf{e}_o^{(i)}$. The second attention layer functions as a *mutual attention* layer, where the key-value pair of the human branch Transformer is provided by the object hidden embedding $\mathbf{e}_o^{(i)}$, and vice versa, which can be formulated as:

$$
\boldsymbol{H}^{(i+1)} = FF(softmax(\frac{\mathbf{Q}_h \mathbf{K}_o^T}{\sqrt{C}} \mathbf{V}_o)); \boldsymbol{O}^{(i+1)} = FF(softmax(\frac{\mathbf{Q}_o \mathbf{K}_h^T}{\sqrt{C}} \mathbf{V}_h)),
$$
$$
\mathbf{Q}_h = \mathbf{e}_h^{(i)} \mathbf{W}_h^{Q_h}, \mathbf{K}_h = \boldsymbol{H}^{(i)} \mathbf{W}_h^{K_h}, \mathbf{V}_h = \boldsymbol{H}^{(i)} \mathbf{W}_h^{V_h},
$$
$$
\mathbf{Q}_o = \mathbf{e}_o^{(i)} \mathbf{W}_o^{Q_o}, \mathbf{K}_o = \boldsymbol{O}^{(i)} \mathbf{W}_o^{K_o}, \mathbf{W}_o = \boldsymbol{O}^{(i)} \mathbf{W}_o^{V_o},
$$

$$(5)$$

where the subscript h and o denote the human branch and the object branch, respectively, $\mathbf{W}_h$ and $\mathbf{W}_o$ denote the trainable parameters of the two branches.

Losses Formulation. To model the spatial relation between objects, we designed a novel object-pairwise loss. Our insight is that the relative distance between the interacted objects changes constantly with certain patterns, and explicitly adding constraints on the relative distance between the interacted objects is beneficial. For example, when cutting an apple with a knife, the knife gradually approaches and stays on the surface of the apple for a while, then moves away. We adopt the $\mathbb{L}2$ loss as $\mathcal{L}_{dis}$ to keep the consistency between the generated results and the ground truth.

Besides, we also adopt the widely adopted geometric loss in the field of human motion, including joint position loss $\mathcal{L}_{pos}$ and the joint velocity loss $\mathcal{L}_{vel}$. To ensure the authenticity of our synthesized motion, we additionally utilized the interpenetration loss $\mathcal{L}_{pen}$ between human and objects. Here, we summarize our
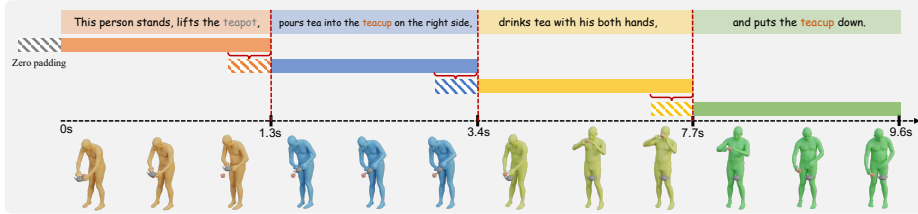
Fig. 5: The auto-regressive generation pipeline. The auto-regressive pipeline to iteratively generate the HOI synthesis results. We obtain the subsequent motion conditioned on the last few frames of the previously generated motion.

ultimate optimization objective as:

$$\mathcal{L} = \lambda_{vel}\mathcal{L}_{vel} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{pen}\mathcal{L}_{pen} + \lambda_{dis}\mathcal{L}_{dis}, \tag{6}$$

and we empirically set $\lambda_{vel} = \lambda_{pos} = \lambda_{pen} = 1$, and $\lambda_{dis} = 0.1$. More details of the formulations of the losses will be illustrated in the supplementary materials.

4.2   The `HIMO-SegGen` Framework

Problem Formulation. The only difference with `HIMO-Gen` lies in the input text prompts, where `HIMO-Gen` takes one single text description L as a condition, yet `HIMO-SegGen` is conditioned on a series of several atomic text prompts as $L = \{l_1, l_2, \cdots, l_m\}$. The key of `HIMO-SegGen` is to ensure smooth and realistic transitions between the generated HOI clips.

Generation Pipeline. Our generation pipeline is demonstrated in Fig. 5, which is simple yet effective. We segment the long HOI sequences into smaller motion-text pairs so that the natural transitions of consecutive HOIs are reserved. We first train the `HIMO-Gen` model to empower the ability to generate more fine-granularity HOIs. Different from the vanilla `HIMO-Gen` model which is only conditioned on the initial state of the first frame, we modify it to conditioned generation on the past *few* frames. During the inference time, we take the last few frames of the previous generation result as the condition for the next generation, to iteratively obtain the composite HOI results.

5   Experiments

In this section, we elaborate on the dataset and implementation details, the baseline methods and the evaluation metrics. Then we present extensive experimental results with ablation studies to show the effectiveness of our benchmark.
Dataset Details. We follow [20] to split our dataset into train, test, and validation sets with the ratio of 0.8, 0.15, and 0.05. The maximum motion length is set to 300 for `HIMO-Gen` and 100 for `HIMO-SegGen`. The maximum text length is set to 40 and 15, respectively. To mitigate the influence of the order of several objects, we randomly shuffle the objects when loading the data.

Table 2: Quantitative baseline comparisons on the **2-objects** partition of HIMO. $\pm$ indicates the 95% confidence interval and $\rightarrow$ means the closer the better. **Bold** indicates best result and <u>underline</u> indicates second best.

Method	R-Precision(Top 3)↑	FID ↓	MM-Dist↓	Diversity→	MModality↑
Real	$0.7988^{\pm0.0081}$	$0.0176^{\pm0.0065}$	$3.5659^{\pm0.0109}$	$11.3973^{\pm0.2577}$	–
IMoS [19]	$0.5013^{\pm0.0120}$	$7.5890^{\pm0.1121}$	$8.7402^{\pm0.0310}$	$7.0033^{\pm0.3205}$	$0.9920^{\pm0.2004}$
MDM [62]	$0.6052^{\pm0.0099}$	$6.8457^{\pm0.3315}$	$8.0187^{\pm0.0500}$	$\mathbf{11.3891^{\pm0.2342}}$	$1.2880^{\pm0.2110}$
priorMDM [56]	$0.5891^{\pm0.0031}$	$7.8517^{\pm0.2516}$	$7.2509^{\pm0.0065}$	$12.5799^{\pm0.1460}$	$1.5911^{\pm0.1449}$
HIMO-Gen	$\underline{0.6369^{\pm0.0032}}$	$\mathbf{1.4811^{\pm0.0427}}$	$\underline{3.6491^{\pm0.0101}}$	$\underline{11.6603^{\pm0.2043}}$	$\underline{1.7863^{\pm0.0570}}$
HIMO-SegGen	$\mathbf{0.6404^{\pm0.0449}}$	$\underline{4.2004^{\pm0.4729}}$	$\mathbf{3.6077^{\pm0.0230}}$	$15.7317^{\pm0.4090}$	$\mathbf{2.0495^{\pm0.0030}}$

Table 3: Quantitative baseline comparisons on the **3-objects** partition of HIMO.

Method	R-Precision(Top 3)↑	FID ↓	MM-Dist↓	Diversity→	MModality↑
Real	$0.6988^{\pm0.0054}$	$0.1811^{\pm0.0442}$	$3.7696^{\pm0.0279}$	$9.7674^{\pm0.2180}$	–
IMoS [19]	$0.4662^{\pm0.1010}$	$4.9902^{\pm0.1772}$	$7.7702^{\pm0.0585}$	$\underline{9.2310^{\pm0.1130}}$	$0.6254^{\pm0.1102}$
MDM [62]	$0.5025^{\pm0.0136}$	$\mathbf{4.5713^{\pm0.1100}}$	$6.3144^{\pm0.0268}$	$8.8953^{\pm0.2851}$	$\underline{0.7893^{\pm0.5715}}$
priorMDM [56]	$\underline{0.5137^{\pm0.0257}}$	$4.8210^{\pm0.2030}$	$\underline{5.8900^{\pm0.0232}}$	$\mathbf{9.3402^{\pm0.0230}}$	$\mathbf{0.8120^{\pm0.3280}}$
HIMO-Gen	$\mathbf{0.5350^{\pm0.0185}}$	$\underline{4.7712^{\pm0.1638}}$	$\mathbf{5.0866^{\pm0.0412}}$	$8.9460^{\pm0.1378}$	$0.7561^{\pm0.0546}$

Implementation Details. For both `HIMO-Gen` and `HIMO-SegGen`, the denoising network consists of $\ell_{enc} = 8$ layers of mutual interaction modules with 4 attention heads. We train the network using Adam [28] optimizer with a learning rate of 0.0001 and weight decay of 0.99. The training of `HIMO-Gen` and `HIMO-SegGen` takes about 9 and 12 hours on a single A100 GPU with a batch size of 128.

Baseline Methods. We adopt the prominent text-to-motion methods MDM [62], priorMDM [56] and recent intent-driven HOI synthesis method IMoS [19] as our baselines. To be specific, we re-implement each model to support the condition input of object geometry and the initial states of human and objects. More details of the baselines are provided in the supplementary materials.

Evaluation Metrics. Following [20], we train a motion feature extractor and a text feature extractor in a contrastive manner for evaluation. Since we are evaluating the generation quality of HOIs, the motion features here include both human and object motion. Then we evaluate over the set of metrics proposed by [20]: *R-precision* and *MM-Dist* determine the relevance between the generated motion and the input prompts, *FID* measures the dissimilarity of the generated and ground-truth motion in latent feature space, *Diversity* measures the variability of the generated motions, and *MultiModality* shows the mean-variance of generated motion conditioned on a single text prompt.

5.1   Results and Analysis

Quantitative Comparisons. Quantitative results on the 2-objects and 3-objects partitions of HIMO are demonstrated in Tab. 2 and Tab. 3. In the 2-objects setting, our method HIMO-Gen outperforms the baseline methods on
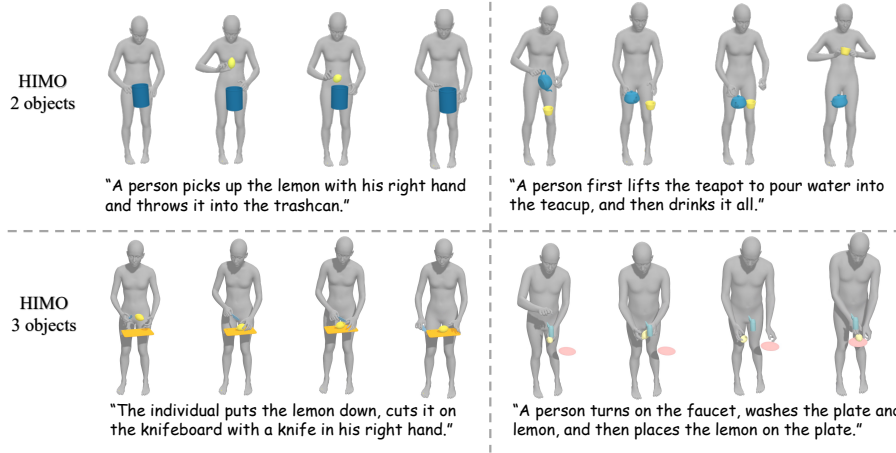
Fig. 6: Visualization results. Our HIMO-Gen framework can generate plausible and realistic sequences of human interacting with multiple objects.

all metrics, which indicates our method has better generation quality over baseline methods. The quantitative results of HIMO-SegGen outperform HIMO-Gen in MM-Dist and MModality, which may be attributed to its fine-grained control of generated motion since HIMO-SegGen takes more concise descriptions as input conditions. In the 3-objects setting of HIMO, our method also outperforms baselines on the R-precision and MM-Dist metrics, which further shows the effectiveness of our method.

Qualitative Results. Qualitative results of our method are depicted in Fig. 6, which show that our method can synthesize plausible human-object interaction sequences. More generation results of `HIMO-SegGen` and visual comparisons will be presented in the supplementary materials.

Generalization Experiments. We find that our trained model can also apply to unseen object meshes. Besides, with our elaborate annotate temporal segments, our trained `HIMO-SegGen` model can generate *novel* HOI compositions. Due to the space limit, more results are listed in the supplementary materials.

5.2 Ablation Study

We conduct extensive ablation studies on `HIMO-Gen` and `HIMO-SegGen`, as depicted in Tab. 4 and Tab. 5. For `HIMO-Gen`, we experiment with three settings: 1) Replacing the **mutual interaction module** with the original transformer layer (w/o IM); 2) Removing the **object-pairwise loss** (w/o $\mathcal{L}_{dis}$); 3) Generating human motion conditioned on the generated objects motion as [32]. From Tab. 4, we can derive that the removal of the mutual interaction module results in a huge drop of R-precision, which verifies the effectiveness of the feature fusion between human and objects. The object-pairwise loss helps a lot in constraining the generated motions. Consecutive generation scheme brings huge performance drop

Table 4: Ablation studies of several model designs on **2-objects** partition of HIMO.

Method	R-Precision(Top 3)↑	FID ↓	MM-Dist↓	Diversity→	MModality↑
Real	$0.7988^{\pm 0.0081}$	$0.0176^{\pm 0.0065}$	$3.5659^{\pm 0.0109}$	$11.3973^{\pm 0.2577}$	—
Ours-w/o IM	$0.4710^{\pm 0.0509}$	$3.5113^{\pm 0.0029}$	$6.7135^{\pm 0.1029}$	$13.0256^{\pm 0.1207}$	$1.2086^{\pm 0.0492}$
Ours-w/o $\mathcal{L}_{dis}$	$\mathbf{0.6426}^{\pm \mathbf{0.0085}}$	$\underline{2.8694}^{\pm 0.0488}$	$\underline{4.9342}^{\pm 0.0191}$	$12.0789^{\pm 0.1153}$	$\underline{1.4394}^{\pm 0.0549}$
Ours-consec.	$0.3902^{\pm 0.3046}$	$8.5024^{\pm 0.2310}$	$8.9982^{\pm 0.0120}$	$10.2788^{\pm 0.1113}$	$1.1209^{\pm 0.0890}$
HIMO-Gen	$\underline{0.6369}^{\pm 0.0032}$	$\mathbf{1.4811}^{\pm \mathbf{0.0427}}$	$\mathbf{3.6491}^{\pm \mathbf{0.0101}}$	$\mathbf{11.6603}^{\pm \mathbf{0.2043}}$	$\mathbf{1.7863}^{\pm \mathbf{0.0570}}$

Table 5: Ablation studies of the frame number of conditioned frames for `HIMO-SegGen` on **2-objects** partition of the HIMO dataset.

#Frame	R-Precision(Top 3)↑	FID ↓	MM-Dist↓	Diversity→	MModality↑
Real	$0.7988^{\pm 0.0081}$	$0.0176^{\pm 0.0065}$	$3.5659^{\pm 0.0109}$	$11.3973^{\pm 0.2577}$	—
1	$0.5502^{\pm 0.0230}$	$8.7320^{\pm 0.3208}$	$6.2094^{\pm 0.3553}$	$\mathbf{13.2999}^{\pm \mathbf{0.5034}}$	$0.8814^{\pm 0.0394}$
5	$0.5448^{\pm 0.0493}$	$7.3029^{\pm 0.2677}$	$\underline{5.5053}^{\pm 0.0026}$	$17.8045^{\pm 0.0956}$	$0.9500^{\pm 0.2374}$
10	$\mathbf{0.6404}^{\pm \mathbf{0.0449}}$	$\mathbf{4.2004}^{\pm \mathbf{0.4729}}$	$\mathbf{3.6077}^{\pm \mathbf{0.0230}}$	$15.7317^{\pm 0.4090}$	$\underline{2.0495}^{\pm 0.0030}$
20	$\underline{0.5508}^{\pm 0.0891}$	$\underline{5.6749}^{\pm 0.1002}$	$6.7708^{\pm 0.0355}$	$\underline{15.2112}^{\pm 0.2210}$	$\mathbf{2.3863}^{\pm \mathbf{0.3399}}$

to generate the human and object motions simultaneously. For `HIMO-SegGen`, we experiment on the conditioned past 1, 5, 10 and 20 frames. Given the results in Tab. 5, "10-frames" achieves the best performance in R-precision, FID and *MM-Dist*. More past frames ("20 frames") provide no additional performance gain, which may result from less attention to the predicted frames.

6 Conclusion

In this paper, we propose HIMO, a dataset of full-body human interacting with multiple household objects. We collect a large amount of 4D HOI sequences with precise and diverse human motions, object motions and mutual contacts. Based on that, we annotate the detailed textual descriptions for each HOI sequence to enable the text-driven HOI synthesis. We also equip it with elaborative temporal segments to decompose the long sequences into several atomic HOI steps, which facilitates the HOI synthesis driven by consecutive text prompts. Extensive experiments show plausible results, verifying that the HIMO dataset is promising for downstream HOI generation tasks.

Limitations: Our work has the following limitations: 1) **Large objects:** Our HIMO dataset is built mainly based on small household objects, without human interacting with large objects such as chairs and suitcases. 2) **RGB modality:** Though we set up a Kinect camera to capture the monocular RGB sequences, the human and object appearances are unnatural since humans are wearing tight MoCap suits and objects are attached with reflective markers. 3) **Facial expressions:** We neglect the facial expressions since our dataset mainly focuses on the manipulation of multiple objects, *i.e.*, the body movements, finger gestures and the object motions, which have little correlation with facial expressions.

Acknowledgements

References

1. Aitviewer, `https://eth-ait.github.io/aitviewer/`
2. Noitom, `https://noitom.com/`
3. Optitrack, `https://optitrack.com/`
4. Sketchfab, `https://sketchfab.com/feed`
5. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 2022 International Conference on 3D Vision (3DV). pp. 414–423. IEEE (2022)
6. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
7. Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings. arXiv preprint arXiv:2402.15509 (2024)
8. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: CVPR. IEEE (jun 2022)
9. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: CVPR (6 2019)
10. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: ContactPose: A dataset of grasps with object contact and hand pose. In: ECCV (August 2020)
11. 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)
12. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR. pp. 18000–18010 (2023)
13. Christen, S., Kocabas, M., Aksan, E., Hwangbo, J., Song, J., Hilliges, O.: D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In: CVPR (2022)
14. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: CVPR (2023)
15. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. arXiv preprint arXiv:2311.16097 (2023)
16. Ding, G., Sener, F., Yao, A.: Temporal action segmentation: An analysis of modern techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
17. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR (2023)
18. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: 3DV (2021)
19. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Imos: Intent-driven full-body motion synthesis for human-object interactions. In: Eurographics (2023)

20. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR. pp. 5152–5161 (June 2022)
21. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
22. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: CVPR (2022)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
24. 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). Lecture Notes in Computer Science, vol. 13485, pp. 281–299. Springer (2022)
25. Jian, J., Liu, X., Li, M., Hu, R., Liu, J.: Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In: ICCV. pp. 14713–14724 (October 2023)
26. Jiang, N., Liu, T., Cao, Z., Cui, J., Zhang, Z., Chen, Y., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: ICCV. pp. 9365–9376 (October 2023)
27. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: ICCV. pp. 2151–2162 (2023)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
29. Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: Nifty: Neural object interaction fields for guided human motion synthesis (2023)
30. Lee, T., Moon, G., Lee, K.M.: Multiact: Long-term 3d human motion generation from multiple action labels. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1231–1239 (2023)
31. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. arXiv preprint arXiv:2312.03913 (2023)
32. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
33. Li, Q., Wang, J., Loy, C.C., Dai, B.: Task-oriented human-object interactions generation with implicit neural representations (2023)
34. Li, S., Zhuang, S., Song, W., Zhang, X., Chen, H., Hao, A.: Sequential texts driven cohesive motions synthesis with natural transitions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9498–9508 (2023)
35. Li, Y., Xue, Z., Xu, H.: Otas: Unsupervised boundary detection for object-centric temporal action segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6437–6446 (2024)
36. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. arXiv preprint arXiv:2304.05684 (2023)
37. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3d whole-body mesh recovery with component aware transformer. In: CVPR. pp. 21159–21168 (2023)
38. Liu, K., Li, Y., Liu, S., Tan, C., Shao, Z.: Reducing the label bias for timestamp supervised temporal action segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6503–6513 (2023)
39. Liu, Y., Yang, H., Si, X., Liu, L., Li, Z., Zhang, Y., Liu, Y., Yi, L.: Taco: Benchmarking generalizable bimanual tool-action-object understanding. arXiv preprint arXiv:2401.08399 (2024)

40. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: CVPR. pp. 21013–21022 (June 2022)
41. Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Human-tomato: Text-aligned whole-body motion generation (2023)
42. Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., Asfour, T.: Unifying representations and large-scale whole-body motion databases for studying human motion. IEEE Transactions on Robotics **32**(4), 796–809 (2016)
43. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018)
44. Pan, Z., Zeng, A., Li, Y., Yu, J., Hauser, K.: Algorithms and systems for manipulating multiple objects. IEEE Transactions on Robotics (2022)
45. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR. pp. 10975–10985 (2019)
46. 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)
47. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: ECCV (2022)
48. 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. arXiv preprint arXiv:2401.08559 (2024)
49. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
50. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets. In: ICCV. pp. 4332–4341 (2019)
51. Pujades, S., Mohler, B., Thaler, A., Tesch, J., Mahmood, N., Hesse, N., Bülthoff, H.H., Black, M.J.: The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements. IEEE transactions on visualization and computer graphics **25**(5), 1887–1897 (2019)
52. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: bodies, action and behavior with english labels. In: CVPR. pp. 722–731. Computer Vision Foundation / IEEE (2021)
53. Qian, Y., Urbanek, J., Hauptmann, A.G., Won, J.: Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2306–2316 (2023)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
55. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2020)
56. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
57. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866 (2023)
58. Taheri, O., Choutas, V., Black, M.J., Tzionas, D.: Goal: Generating 4d whole-body motion for hand-object grasping. In: CVPR. pp. 13263–13273 (June 2022)

59. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV (2020), `https://grab.is.tue.mpg.de`
60. Taheri, O., Zhou, Y., Tzionas, D., Zhou, Y., Ceylan, D., Pirk, S., Black, M.J.: Grip: Generating interaction poses using latent consistency and spatial cues (2023)
61. Tendulkar, P., Surís, D., Vondrick, C.: Flex: Full-body grasping without full-body grasps. In: CVPR (2023)
62. 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), `https://openreview.net/forum?id=SJ1kSyO2jwu`
63. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. In: NeurIPS (2022)
64. Wu, Y., Wang, J., Zhang, Y., Zhang, S., Hilliges, O., Yu, F., Tang, S.: Saga: Stochastic whole-body grasping with contact. In: ECCV (2022)
65. Xu, L., Lv, X., Yan, Y., Jin, X., Wu, S., Xu, C., Liu, Y., Zhou, Y., Rao, F., Sheng, X., et al.: Inter-x: Towards versatile human-human interaction analysis. arXiv preprint arXiv:2312.16051 (2023)
66. Xu, L., Song, Z., Wang, D., Su, J., Fang, Z., Ding, C., Gan, W., Yan, Y., Jin, X., Yang, X., et al.: Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In: ICCV. pp. 2228–2238 (2023)
67. Xu, L., Zhou, Y., Yan, Y., Jin, X., Zhu, W., Rao, F., Yang, X., Zeng, W.: Regennet: Towards human action-reaction synthesis. In: CVPR. pp. 1759–1769 (2024)
68. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: InterDiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023)
69. Xu, X., Joo, H., Mori, G., Savva, M.: D3d-hoi: Dynamic 3d human-object interactions from videos. arXiv preprint arXiv:2108.08420 (2021)
70. Yang, S., Zhang, W., Song, R., Cheng, J., Wang, H., Li, Y.: Watch and act: Learning robotic manipulation from visual demonstration. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2023)
71. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: Manipnet: Neural manipulation synthesis with a hand-object spatial representation. ACM Trans. Graph. **40**(4) (jul 2021). `https://doi.org/10.1145/3450626.3459830`, `https://doi.org/10.1145/3450626.3459830`
72. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: Pymaf-x: Towards well-aligned full-body model regression from monocular images. PAMI (2023)
73. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: CVPR. pp. 14730–14740 (June 2023)
74. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In: CVPR (2023)
75. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. arXiv preprint arXiv:2303.17076 (2023)
76. Zhang, X., Bhatnagar, B.L., Starke, S., Guzov, V., Pons-Moll, G.: Couch: Towards controllable human-chair interactions (October 2022)
77. Zhang, Y., Black, M.J., Tang, S.: Perpetual motion: Generating unbounded human motion. arXiv preprint arXiv:2007.13886 (2020)
78. Zheng, J., Zheng, Q., Fang, L., Liu, Y., Yi, L.: Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In: CVPR. pp. 585–594 (June 2023)

79. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR. pp. 5745–5753. Computer Vision Foundation / IEEE (2019)
80. Zhou, Z., Wang, B.: Ude: A unified driving engine for human motion generation. In: CVPR. pp. 5632–5641 (June 2023)