Interactive Humanoid: Online Full Body Human Motion Reaction Synthesis With Social Affordance Forecasting and Canonicalization

Yunze Liu
Tsinghua University
Shanghai Qi Zhi Institute
liuyzchina@gmail.com

Changxi Chen
Tsinghua University
knightleyaa@gmail.com

Li Yi
Tsinghua University
Shanghai AI Laboratory
Shanghai Qi Zhi Institute
ericyi0124@gmail.com

Abstract

We focus on the human-humanoid interaction problem optionally with an object. We propose a new task named online full-body motion reaction synthesis, which generates humanoid reactions based on the human actor's motions. The previous work only focuses on human interaction without objects and generates body reactions without hand. Besides, they also do not consider the task as an online setting, which means the reactor can only see the current information and cannot perceive the future actions of the actor. To support the task of online full-body motion reaction synthesis, we construct two datasets named HHI and CoChair and propose a unified method. Specifically, we encode the motion of human actors and objects from an interaction-centric view through a social affordance representation. Then we leverage a social affordance forecasting scheme to enable the reactor to predict based on the imagined future. We also use SE(3)-Equivariant Neural Networks to learn the local frame to canonicalize the social affordance. Experiments demonstrate that our approach effectively generates high-quality reactions on HHI and CoChair. Furthermore, we also validate our method on existing human interaction datasets Interhuman and Chi3D in real-time at 25 fps. Website: https://yunzeliu.github.io/iHuman/

1. Introduction

In various applications including VR/AR, games, and human-robot interaction, there is a strong demand for generating reactive humanoid characters or robots based on the actions of human actors. Such a reaction needs to occur in real-time, dynamically responding to the movements of the human actor. Furthermore, in many cases, these interactions involve objects (e.g., a human and a humanoid collaboratively carrying a chair) and call for an emphasis on the precise movements of humanoid hands in addition to the overall body motion. Addressing the challenge of synthesizing humanoid[1] reactions in these contexts can significantly enhance the social experience of humans interacting with humanoids.

Previous research on humanoid motion synthesis has mainly focused on single humanoid movements [4, 15, 19, 23, 40, 49, 56] or interactions with objects [13, 51, 54, 55, 57, 59]. Some recent studies [31, 49], have explored the synthesis of social interactions between two humanoids. However, these studies have limitations. Firstly, they primarily focus on the offline generation between two humanoids, which is not suitable for the asymmetric reaction synthesis setting where the humanoid continuously responds to the dynamic human actor in an online manner. Secondly, they overlook the fact that humans often interact with objects. Thirdly, recent studies [31, 49] do not consider synthesizing full-body motions involving both body and hand motions, which are crucial for various interactions such as handshakes or collaborations. Therefore, synthesizing full-body humanoid reactions online considering both human actors and the possible objects goes beyond the scope of existing works, presenting three major challenges: 1) representing complex motions of a human actor and optionally an object, 2) interpreting the human actor's intentions for prompt reactions by the humanoid, and 3) supporting detailed reactions involving both coarse-grained body movements and fine-grained hand movements.

To address the challenges mentioned, we draw inspiration from affordance learning. Our approach involves encoding the motion of the human actor (possibly with an object) as a social affordance representation, capturing supported and expected social interactions at both the body and hand levels. Subsequently, we learn the humanoid's reaction based on the social affordance representation. We introduce an online social affordance forecasting scheme to

[1] In this paper, we use *human* to denote real people initiating interactions and *humanoid* to indicate the virtual character reacting in response.

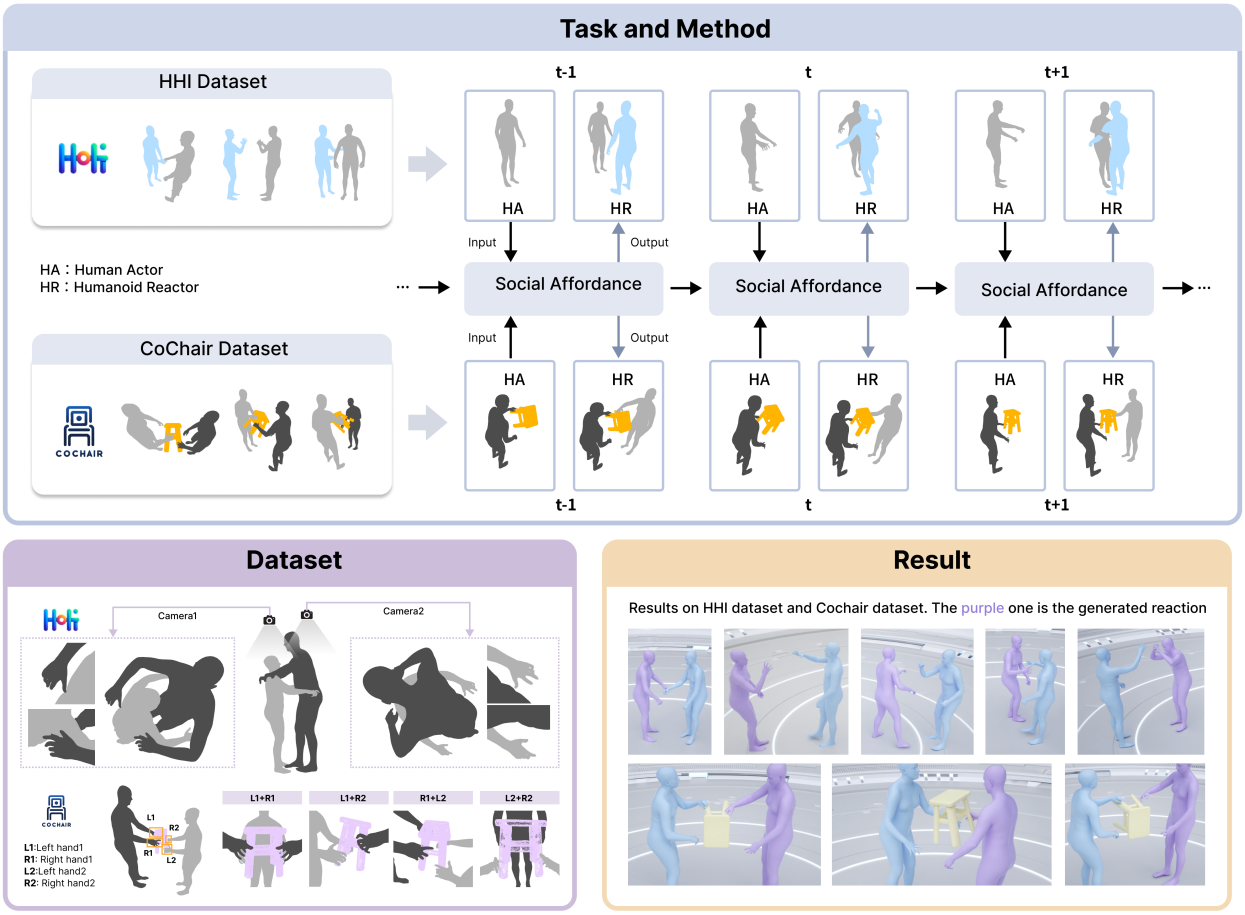

Figure 1. We propose a new task named online full-body motion reaction synthesis optionally with an object. We construct two datasets HHI and CoChair to support the task. We propose Social Affordance Forecasting and Canonicalization techniques to generate realistic reactions and establish benchmarks.

enable the humanoid to react promptly. Furthermore, we employ a canonicalization technique to simplify the distribution of social affordance and facilitate learning.

Specifically, our method handles a sequential input data stream comprising the human actor's pose at each time step. When objects are involved, the input also includes a stream of 6D-transforming 3D shapes. To unify the input with and without objects, we introduce the concept of "affordance carrier", which can be either the real object in human-object-humanoid interactions or just the reactor in a rest state in human-humanoid interaction. Centering around the affordance carrier, we propose a social affordance representation encompassing the actor's motion, the carrier's dynamic geometry, and the actor-carrier relationship up until each time step. In an online setup, the humanoid reactor can only access present and past observations, restricting its social affordance to short-sighted information and making prompt reactions challenging. To overcome this, we first propose a social affordance forecasting scheme to enable the reactor to imagine the future and react accordingly. However, we find directly predicting the reaction based on the forecasted social affordance representation not satisfactory enough. It is because the actor's motion exhibits diverse patterns, complicating the social affordance and increasing the learning difficulty. To address this, we observe that the patterns of the actor's motion become more compact when viewed from the carrier's local coordinate system. We, therefore, learn such local frames to canonicalize the social affordance through SE(3)-Equivariant network design. Finally, we learn to predict the humanoid's online reaction through a 4D motion transformer.

To validate the effectiveness of our design and also to address the lack of large-scale full-body reaction-synthesis benchmarks, we have gathered two large-scale full-body social interaction datasets named HHI and CoChair. HHI covers a diverse range of human-human interactions with a clear actor and reactor while CoChair focuses on human-object-human interaction. Our method consistently outperforms previous methods in all metrics. Moreover, our method can provide more reasonable and prompt full-body

reactions and better collaboration with human actors. We also validate our method on previous datasets Interhuman and Chi3D.

The key contributions of this paper are threefold: i) we propose a new task named online full-body motion reaction synthesis optionally with an object and establish benchmarks; ii) we propose a unified solution to reaction synthesis with or without objects by social affordance canonicalization and forecasting, significantly outperforming baselines; iii) we construct two datasets HHI and CoChair to support the research on full-body reaction synthesis tasks.

2. Related Work

Human Motion Generation. Human motion generation is to generate human motion conditioned on different signals. A line of works[4, 9, 11, 19, 33, 39, 49, 52] propose to generate human motion conditioned on action label. Some works[1, 15, 20, 23, 25, 40, 56] directly generate human motion conditioned on text description. There are some works[26, 28, 30] that generate human motion conditioned on music and speech. Recently, some works[31, 44, 46, 49] have started to focus the human-human interaction synthesis. [31] propose a new dataset with natural language descriptions and design a diffusion model to generate human-human interaction. However, this method cannot be directly applied to reaction synthesis because it uses CLIP[41] branch to extract text features. [49] presents a GAN-based Transformer for action-conditioned motion generation. However, it cannot generate full-body motions and handle the presence of objects.

Human Reaction Generation. We focus on motion generation conditioned on another human motion. Human reaction generation is conditioned on the actor's motion and requires the reactor to provide a reasonable response, which is very important in the fields of VR/AR and humanoid robots. [12] propose a Transformer network with both temporal and spatial attention to generate reactions. [6] propose to predict human intent in Human–Human interactions. However, they are only concerned with the generation of body motions and cannot generate hand motions. At the same time, they only focus on the interaction between humans and cannot generate reasonable reactions in the presence of objects. In addition, reaction synthesis should be in an online setting, meaning that the reactor cannot observe future information, which is also not discussed in previous work. There is no dataset providing full-body human-human interaction and human-object-human interaction with clear actor and reactor, so in this paper, we first construct two datasets and propose a novel method to generate realistic reactions.

Human Motion Prediction Human motion prediction [8, 14, 22, 29, 36, 37, 45, 45, 47, 58] is a traditional task widespread attention. A line of works predicts human motions in an encoding-decoding way [7, 16, 34, 42, 50]. Some works carefully designed loss constraints [3, 21, 27] to generate diversity and realistic human motions. Without multi-stage training,[10] propose a human motion diffusion model to predict human motion in a masked completion fashion. Besides, [51] propose to predict human motion with the object as an HOI sequence and use interaction diffusion and interaction correction to predict the future state of human and object. In this paper, we focus on human-human and human-object-human interactions. We propose to use a motion forecasting module to improve the ability of the reactor, thus relying on both human motion prediction and human-object interaction prediction methods.

3. Constructed HHI and CoChair Datasets

To support our proposed task, we propose two datasets named HHI and CoChair. HHI is the first large-scale dataset with diverse interactions for whole-body reaction synthesis. It not only provides motion capture of the whole body but also designs interaction with clear initiators. Compared to the challenging free-form interaction provided by Interhuman, our dataset has explicit categories of interactive actions, which facilitates the evaluation of generated results. Compared to datasets such as SBU, K3HI, and Chi3D which fully or partially use image-based methods to estimate human poses, our dataset is completely captured by motion capture devices and meticulously annotated by human experts, which can provide higher-quality motions. CoChair is the first large-scale dataset for human-object-human collaborative carrying. It not only has clear motion initiators but also diverse object geometries and different carrying patterns. Compared to other datasets, we support a more challenging setting that involves not only human interaction but also the interaction between humans and objects. For more information on the dataset construction, annotation, visualization, and additional details, please refer to the supplementary materials.

4. Method

Our goal is to generate motions for the humanoid to interact or collaborate with the human in social scenarios or collaborative tasks, which requires the humanoid to not only understand the human's intentions and motions but also comprehend the state of the environment or objects. In this Section, we elaborate on the method in detail. We first introduce the concept of social affordance carrier in Section 4.1 and carrier-centric representation in Section 4.2. We also introduce an online social affordance forecasting scheme to enable the humanoid to react promptly in Section 4.3. Then we introduce the social affordance canonicalization technique to simplify the distribution of social affordance and facilitate learning in Section 4.4. Finally, we introduce the 4D motion transformer and objective function for the entire

Dataset	Object	Full-body	Actor&Reactor	Mocap	Motions	Verbs	Duration
SBU[53]	-	-	-	-	282	8	0.16h
K3HI[5]	-	-	-	-	312	8	0.21h
NTU120[32]	-	-	-	-	739	26	0.47h
You2me[38]	-	-	-	-	42	4	1.4h
Chi3D[18]	-	-	-	✓	373	8	0.41h
InterHuman[31]	-	-	-	✓	6022	5656	6.56h
HHI (Ours)	-	✓	✓	✓	5000	30	5.55h
CoChair (Ours)	✓	✓	✓	✓	3000	5	2.78h

Table 1. **Dataset comparisons.** We compare our iHuman dataset with existing multi-human interaction datasets. **Object** refers to human-object-human interaction. **Whole-body** refers to whole-body motion capture. **Actor&Reactor** refers to whether there is an obvious initiator of the action. **Motions** is the total number of motion clips. **Verbs** is the number of interaction categories.

framework in Section 4.5.

4.1. Social Affordance Carrier

Social affordance carrier refers to the entity having potentially rich contact with the human actor. Specifically, we use either the sparse point cloud of the interaction object or vertices from SMPL-X humanoid template mesh as the carrier since a human actor interacts with them either directly or indirectly. This is shown in the beginning of Fig.2. Encoding the human actor motion, the carrier motion, and the contact relationship between the human actor and the carrier simultaneously from the carrier's local point of view would form an interaction-centric representation describing the social affordance of the scene. To be more specific, we denote a sequence with L frames as $\boldsymbol{s} = [s^1, s^2, \ldots, s^L]$, where s^i consists of human actor h^i and Carrier c^i. Human actor $h^i \in \mathbb{R}^{J \times D_h}$ is defined by J joints(22 for the body and 32 for both hands) with a D_h-dimensional representation at each joint, which is joint position and its velocity. Carrier $c^i \in \mathbb{R}^{N \times 3}$ is defined by the position of N object points or humanoid vertices from SMPL-X in the world coordinate system. We denote the humanoid reactor sequence as $\boldsymbol{s}_r = [s_r^1, s_r^2, \ldots, s_r^L]$. Given the sequence of human actor and carrier $\boldsymbol{s}$, our goal is to predict a reasonable humanoid reactor sequence $\boldsymbol{s}_r$.

4.2. Social Affordance Representation

We define the social affordance representation centered on the carrier as shown at the top center of Fig.2. Specifically, given a carrier, we use a Graph Neural Network(GNN) to encode each human pose and concatenate it with each carrier's points to obtain a carrier-centric representation. With this representation, we propose a social affordance that contains the motion of the human actor, the carrier's dynamic geometry, and the actor-carrier relationship up until each time step. The advantage of the social affordance representation is that it tightly associates the local region of the carrier with the human actor's motion, forming a strong representation for network learning. Note that the carrier-centric representation refers to the spatial representation based on the current time step, while the social affordance representation refers to the spatiotemporal representation from the initial time step to the current time step.

Carrier-centric actor representation. Given a human actor h^i and a carrier c^i at time step i, we first define the carrier-centric actor representation R^i as a collection of point-wise vectors on a set $\{x_j^i\}_{j=1}^N$ of N points from carrier c. Note that this is a dense interaction representation for a single time step.

$$R^i(h^i, c^i) = \{Concat(x_j^i, \epsilon_\theta(h^i))\}_{j=1}^N, \tag{1}$$

where R^i is carrier-centric actor representation, x_j^i is the position of the point or joints from the carrier at time step i, h^i is the human actor at time step i and ϵ_θ is the GNN network to encode human actor's pose to an embedding. *Concat* means the concatenation operation.

Social Affordance Representation. Given the carrier-centric actor representation at each time step, we define the social affordance representation A^i as a collection of $\{R^t\}_{t=1}^i$ up until each time step. Note that the social affordance representation is a data stream from the beginning to a certain time step.

$$A^i = \{R^t\}_{t=1}^i = \{\{Concat(x_j^t, \epsilon_\theta(h_j^t))\}_{j=1}^N\}_{t=1}^i, \tag{2}$$

where A^i refers to the social affordance representation at time step i and R^t refers to carrier-centric actor representation at time step i.

4.3. Social Affordance Forecasting with Human Motion Forecasting Module

The definition reveals that Social Affordance Representations vary in length over time, hindering network learning. We propose to employ Social Affordance Forecasting to transform these representations into a fixed length for each time step.

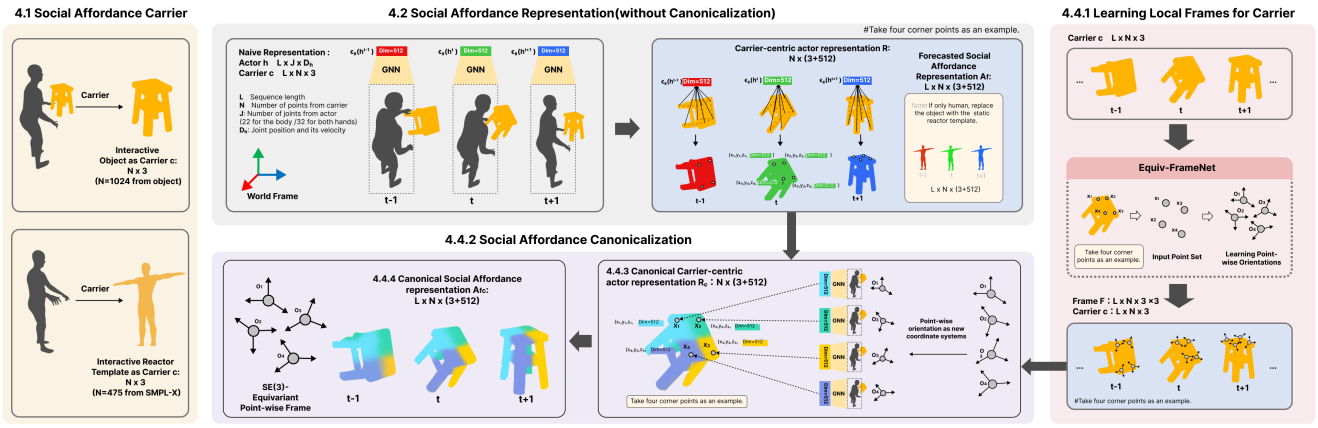

Figure 2. **Social Affordance Representation and Canonicalization.** Given a sequence, we first select a social affordance carrier and build the carrier-centric representation. Then we can compute the social affordance representation. Here, the carrier-centric representation refers to the spatial representation based on the current time step, while the social affordance representation refers to the spatiotemporal representation from the initial time step to the current time step. It is worth noting that the social affordance representation varies in duration at different time steps. Therefore, we use Social Affordance Forecasting to transform these representations into a fixed length for each time step which is described in Section 4.3. We propose to learn the local frame for carrier and canonicalize social affordance to simplify the distribution. Then a motion encoder and decoder are used to generate reactions which are described in Section 4.5.

Additionally, Social Affordance Forecasting is to anticipate human actors' behavior so that the humanoid reactor can provide more reasonable responses. In real situations, the humanoid reactor can only observe the historical motions of the human actor. The humanoid reactor should have the ability to predict the motions of the human actor to better plan its motions. For example, when someone raises their hand and walks towards you, you might instinctively think they are going to shake hands with you and be prepared for a handshake. Here, we introduce how to enable the reactor to make motion forecasting during prediction.

The t observed motions of human actor and carrier are noted as $\mathbf{s}^{1:M} = [\mathbf{s}^1; \mathbf{s}^2; \ldots; \mathbf{s}^M]$. Therefore, the problem of online reaction synthesis is modeled as predicting $\mathbf{s_r}^M$ given $\mathbf{s}^{1:M}$. Given the observed motion $\mathbf{s}^{1:M}$, the objective of the motion forecasting problem is to predict the following motions $\mathbf{s}^{M+1:L} = [\mathbf{s}^{M+1}; \mathbf{s}^{M+2}; \ldots; \mathbf{s}^L]$.

We use a motion forecasting module to predict the human actor's motion and the object's motion(if available). For the human-humanoid interaction setting, we use HumanMAC[10] as the forecasting module. For the human-object-humanoid interaction setting, we build our motion forecasting module based on InterDiff[51] and add a prior that human-object contact is stable to simplify the difficulty in predicting the object's motion. Finally, with the predicted result, we can obtain the forecasted carrier-centric actor representation $\{R^t\}_{t=L-M}^{L}$, we define the Fix-length Social Affordance Representation A_f^i as:

$$A_f^i = \{R^t\}_{t=1}^{L} = \{\{Concat(x_j^t, \epsilon_\theta(h_j^t))\}_{j=1}^{N}\}_{t=1}^{L}, \quad (3)$$

where A_f^i refers to the Fix-length social Affordance Representation at time step i.

4.4. Social Affordance Canonicalization with SE(3)-Equivariant Local Frame Learning

We find directly predicting the reaction based on the forecasted Social Affordance Representation is not satisfied. It is because the actor's motion exhibits diverse patterns, complicating the social affordance and increasing the learning difficulty. To address this, we observe that the patterns within the actor's motion become more compact when viewed from the carrier's local coordinate system. We aim to learn a local frame that can transform correspondingly with the rigid transformation of the carrier, to canonicalize Social Affordance into a local coordinate system, thereby simplifying the distribution of the interaction patterns. The process of learning such local frames requires the use of SE(3)-Equivariant Neural Network to ensure that the local frame rigidly transforms together with the carrier. Therefore, we present a Social Affordance Canonicalization strategy enabled by SE(3)-Equivariant local frame learning as shown in the bottom center of Fig.2.

Learning Local Frames for Carrier using SE(3)-Equivariant Neural Network. The local frame, defining a new coordinate system, is determined by the point-wise orientation of its local region as shown in the right of Fig.2. We believe that a local frame can reflect the geometric information of the carrier and can reflect the contact information between the carrier's local area and the human actor. Please refer to [24] for more details on SE(3)-Equivariant Local Frame Learning.

Let c denote the carrier and $\{x_j\}_{j=1}^{N}$ as each point from carriers. Let H and $\mathbf{V}$ denote per-point invariant scalar fea-

tures and equivariant vector features of c, respectively.

We use $c, H_{in}, \mathbf{V}_{in}$ to denote the carrier, invariant scalar, and equivariant vector features, where $H_{in}, \mathbf{V}_{in}$ are all zeros to ensure strict SE(3) equivariance. We pass the carrier to an Equivariant network that aims to extract invariant and equivariant features, denoted as EquivLayer. Our EquivLayer is adapted from the GVP-GNN layer[24].

$$(H_{out}, \mathbf{V}_{out}) \leftarrow \text{EquivLayer}(c, H_{in}, \mathbf{V}_{in}). \tag{4}$$

Since EquivLayer is equivariant at all the layers, and the inputs $H_{in}, \mathbf{V}_{in}$ are invariant and equivariant features, the output $H_{out}, \mathbf{V}_{out}$ of each layer are also invariant and equivariant features, respectively.

To obtain local frames of each point from invariant and equivariant features, we use another set of equivariant networks adapted from the GVP layers[24]. We use FrameNet to denote the network.

$$\mathbf{V}_{out} \leftarrow \text{FrameNet}(H_{out}, \mathbf{V}_{out}), \tag{5}$$

where each frames will be constructed from the equivariant features $\mathbf{V}_{out} = (\mathbf{v}_{out,1}, \cdots, \mathbf{v}_{out,N})(\mathbf{v}_{out,j} \in \mathbb{R}^{2\times 3})$.

We orthonormalize the two vectors $\mathbf{v}_{out,j,1}, \mathbf{v}_{out,j,2}$ for each point to get $\mathbf{u}_{j,1}, \mathbf{u}_{j,2}$ using the Gram-Schmidt method and the third direction can be derived. Then we get the local frame $\mathbf{F} = \{\mathbf{F}\}_{j=1}^{N} = \{\mathbf{u}_{j,1}, \mathbf{u}_{j,2}, \mathbf{u}_{j,1} \times \mathbf{u}_{j,2}\}_{j=1}^{N} \in \mathbb{R}^{3\times 3}$. Since $\mathbf{V}_{out}$ is rotation equivariant, the constructed frames are also rotation equivariant. We refer to the whole module to generate local frames as Equiv-FrameNet:

$$\mathbf{F} \leftarrow \text{Equiv-FrameNet}(c, H_{in}, \mathbf{V}_{in}), \tag{6}$$

where $\mathbf{F} = \{\mathbf{F}_j\}_{j=1}^{N}$ denotes the local frames of each point from the carrier.

For a local region, the equivariant network can output any orientation while ensuring equivariance. We expect the network to learn an orientation that is optimal for downstream tasks and can be generalized across carriers.

Social Affordance Canonicalization. We propose a Social Affordance Canonicalization technique to simplify the distribution. We will explain in detail how to canonicalize social affordance using learned local frame $\mathbf{F}$.

Since we have learned an equivariant local frame $\mathbf{F}$ for every point from the carrier, instead of directly encoding the human actor's pose using GNN in a unified world coordinate system, we first transform the motions of the human actor into each learning frame(a new coordinate system of each point from the carrier. Next, we encode the human actor's pose in these new frames respectively using GNN to obtain a frame-aware dense object-centric HOI representation. This can be seen as binding an 'observer' to each point on the carrier, and each 'observer' encodes the actor's motions from a first-person view. The advantage is that it tightly associates the object's motion with the actor's motion, simplifying the distribution of social affordance and facilitating network learning.

Given a human actor h^i and a carrier c^i at time step i, we define the canonical carrier-centric actor representation R_c^i as:

$$R_c^i(h^i, c^i) = \{Concat(x_j^i, \epsilon_\theta(F_j^i(h_j^i)))\}_{j=1}^{N}, \tag{7}$$

where F_j^i is the local frame of point j at time step i, $F_j^i(h_j^i)$ is the transformed human actor's pose in learned frame F_j^i. Based upon canonical carrier-centric actor representation R_c^i and the Social Affordance Forecasting process, we define the forecasted canonical social affordance representation A_c^i as:

$$A_{fc}^i = \{R^t\}_{t=1}^{L} = \{\{Concat(x_j^t, \epsilon_\theta(F_j^i(h_j^i)))\}_{j=1}^{N}\}_{t=1}^{L}, \tag{8}$$

where A_{fc}^i refers to the forecasted canonical Social Affordance Representation at time step i and F_j^i is the local frame of point j at time step i.

4.5. Network and Objective Design

Our network consists of a GNN and a 4D Transformer autoencoder. The GNN[43] transforms the human actor pose into a feature, efficiently modeling the relative motion between different joints. The 4D Transformer autoencoder[48] is composed of a motion encoder and a motion decoder. The motion encoder takes the forecasted canonical Social Affordance as input and generates a latent embedding of it. The motion decoder uses the latent embedding as a condition and takes the previously taken motions of the reactor as input to generate new reaction motions autoregressively.

Specifically, given a sequence s, we can compute the A_{fc}, and we can obtain the motions of the reactor s_r by a 4D Transformer network.

$$\hat{s_r} = 4DNet(A_{fc}), \tag{9}$$

where $4DNet$ denotes the whole 4D backbone to generate the motions of the humanoid reactor.

We use two loss functions to train our model. The first one is the sequence loss which compares the generated position of joints with the ground truth using the Mean Square Error. The second one is the velocities of each joint.

$$Loss = MSE(s_r - \hat{s_r}) + MSE(ds_r - \hat{ds_r}), \tag{10}$$

where s_r refers to the GT position of each joint and ds_r refers to the velocities.

Method	FVD ↓			Diversity →			Accuracy ↑			User Preference↑		
	HHI	InterHuman[31]	Chi3D[18]	HHI	InterHuman[31]	Chi3D[18]	HHI	InterHuman	Chi3D[18]	HHI	InterHuman[31]	Chi3D[18]
Real	0.21	0.17	0.05	10.8	12.4	14.0	88.2	-	80.4	-	-	-
PGBIG[35]	56.7	87.2	67.2	13.9	17.1	17.8	34.1	-	61.6	4.4	1.0	8.3
SS-Transformer[2]	77.8	107.3	54.9	16.2	18.5	19.2	51.9	-	57.1	2.7	4.6	18.4
InterFormer[12]	54.3	73.1	20.8	14.1	14.2	14.8	77.9	-	62.2	6.0	2.1	13.7
InterGen-Revised[31]	19.8	25.7	17.7	11.6	13.3	14.2	80.2	-	71.9	19.7	41.7	15.4
Ours	**13.3**	**14.7**	**12.8**	**11.1**	**13.3**	**14.1**	**85.4**	-	**77.6**	**67.2**	**50.6**	**44.2**

Table 2. Quantitative results on HHI, InterHuman, and Chi3D. Our method consistently outperforms the previous method in all metrics.

Method	FVD ↓	Diversity →	Penetration depth↓	User Preference↑
Real	0.07	16.4	0.5	-
PGBIG[35]	47.6	14.8	7.2	3.5
SS-Transformer[2]	51.2	15.7	3.7	3.3
InterFormer[12]	44.2	15.5	4.2	6.4
InterGen-Revised[31]	26.7	17.4	2.2	28.0
Ours	**7.8**	**16.9**	**0.9**	**58.8**

Table 3. Quantitative results on CoChair dataset.

Method	FVD ↓	Diversity →	Accuracy ↑	User Preference↑
Real	0.21	10.8	88.2	-
w/o canonicalization	34.5	12.5	78.4	13.4
w/o forecasting	16.7	11.4	82.1	19.4
w global frame	28.4	8.9	79.6	20.1
Ours	**13.3**	**11.1**	**85.4**	**47.1**

Table 4. Ablation study to justify each design.

5. Experiment

5.1. Experiment Setting and Metric

We conduct experiments on CoChair, HHI, InterHuman[31] and Chi3D[18]. For detailed data pre-processing, data splits and baseline description, please refer to the supplementary materials. We use metrics commonly used in motion generation for quantitative results including action recognition accuracy, FVD, and diversity. Classification Accuracy measures how well our generated samples are classified by a motion classifier. FVD computes the distance between the ground truth and the generated data distribution. Diversity Score is the average deep feature distance between all the samples. For the human-object-humanoid setting, we also report the mean penetration depth(cm) when the distance between objects and generated reaction grasps is smaller than 0.2cm, which is commonly used for hand-object interaction[54, 57]. For the user study, 30 participants are presented with five videos with a label/description. Their task is to identify the most realistic one. Due to the lower quality of all current methods compared to real data, we decided not to include the real data. We utilize a 4D motion encoder comprising a 4D convolution and a Transformer[48] for feature extraction. We only train and test the classifier on the reaction part of the interaction, so the results are not influenced by the actor. Classifying human interactions solely based on reactor motions introduces greater ambiguity and challenge, leading to relatively lower accuracy in classification. Additionally, the definition of motion classes is dataset-specific. Due to CoChair and InterHuman not having a clear motion category, we use the feature extractor trained on the HHI dataset. We generate 1000 samples 10 times with different random seeds and report the average number. For the user study, we generate the reactions using baseline methods and ours. Users need to choose which one they think is the most reasonable.

5.2. Implementation Details

We train our model using a Nvidia A100 GPU. We use the Adam optimizer with α=0.0001, β_1=0.9, β_2=0.98, and ϵ=1×10^{-9}. The batch sizes are set to 64 for Human and CoChair, 128 for Interhuman, and 64 for CHI3D. We train 2000 epochs for all datasets. We use the pre-trained model provided by HumanMAC to forecast the actor's motion which is trained on the large-scale single human motion dataset Human3.6M and use the pre-trained model provided by InterDiff to forecast the motion of actor and object. For each prediction, we set the first frame of the reactor as known and predict the following reaction in an auto-regressive manner. We use the 22 body joints and 32 finger joints as the default joints set which is also can re-target to another skeleton easily. For the 4D Transformer used to predict reactor actions, we designed an encoder-decoder structured network based on P4Transformer[17] and PPTr[48].

5.3. Comparison to State-of-the-arts

For all baseline methods, we entirely used the author's code or made some modifications to adapt it to our task. The results on the CoChair dataset are shown in Tab.3. The FVD and Diversity score of real data are computed by randomly sampling 1000 real motions from the dataset and comparing it with the whole dataset. Our method consistently outperforms the previous method in all metrics. We compared the results with InterHuman and Fig.3, and we replace the CLIP branch with a spatiotemporal transformer to encode the actor's motion. It can be seen that our method can generate a more realistic and natural grasp(left) and collaboration(right). This indicates that through social affordance canonicalization, our approach can simplify the feature space, thus generating more complex and delicate motions. Through social affordance forecasting, we can anticipate the motions of human actors, thereby better planning cooperation with humans. As for the human-humanoid in-

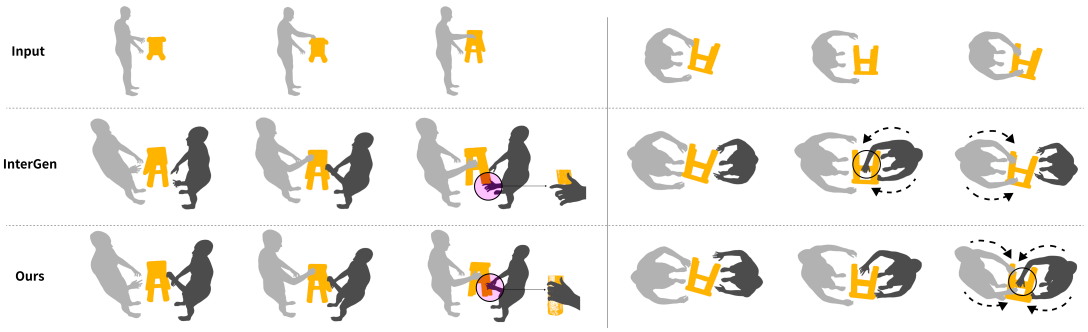

Figure 3. Visualization results on CoChair. Our method can provide a more reasonable grasp and better collaboration with the human actor.

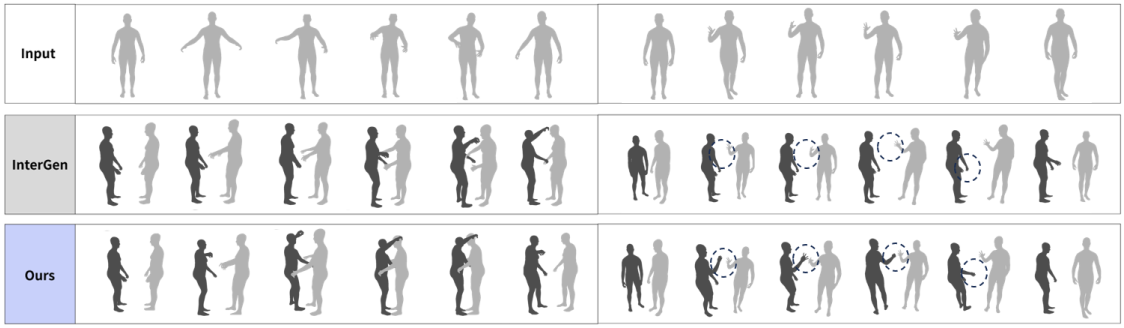

Figure 4. Visualization results on HHI. Our method can generate more prompt reactions and can better capture hand motion.

teraction setting, the results are shown in Tab.2, our method outperforms baselines in all metrics. Some visualization results are shown in Fig.4. Compared with InterGen, our method can generate prompt reaction(left), and can better capture the local hand motions(right), while InterGen fails.

5.4. Ablation and Discussion

Ablation Study. To validate our method, we conducted ablative experiments on the HHI dataset to verify the effectiveness of each design as shown in Tab.4. Without canonicalization, our method drops significantly, indicating that the use of social affordance canonicalization to simplify feature space complexity is essential. Without social affordance forecasting, our method lost the ability to predict human actor motions, also leading to a performance drop. To verify the necessity of using the local frame, we also compared the effect of using a global frame, and it can be seen that our method is significantly superior. This also indicates that using a local frame to describe local geometry and potential contact is valuable.

Visualization of Learned Local Frame for Carrier. We visualized the frames on the rest-posed humanoid carrier as shown in Fig.5. It can be seen that the frames on the spine are consistent, and the joint frames on the left and right sides exhibit roughly symmetrical characteristics. We also visualize the local frames, and it can be seen that the frames on the chair legs are approximately the same and can be generalized to different chairs.

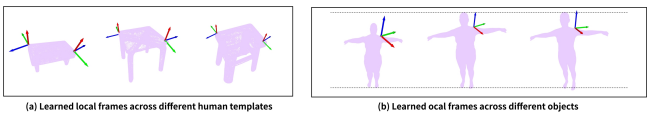

Figure 5. Visualization results of a sampled learned local frame. The local frames are roughly consistent across different chairs which can generalized within category.

Computation Overhead. Our method, compared to InterGen[31] on an 80G A100 graphics card, is more memory-efficient (38.12G vs. 22.28G) and requires fewer parameters (291.29M vs. 11.70M) due to the social affordance canonicalization. And our FrameNet achieves this with just 122B of parameters. It also provides real-time inference at 25fps, surpassing InterGen's 0.54fps.

Method	Memory	Parameter	FPS
InterGen-Revised[31]	38.12G	291.29M	0.54fps
Ours	**22.28G**	**11.70M**	**25fps**

Table 5. Computation overhead of our method.

6. Conclusion

We introduce a new task, online full-body motion reaction synthesis, aimed at generating humanoid reactions to human actors' motions. We develop two datasets and propose social affordance forecasting and canonicalization to produce realistic and natural humanoid reactions. Experiments show our method's efficacy in generating high-quality reactions across our and existing datasets.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 3

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 7

[3] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *CVPR*, pages 5223–5232, 2020. 3

[4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 1, 3

[5] Murchana Baruah and Bonny Banerjee. A multimodal predictive agent model for human interaction generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 1022–1023, 2020. 4

[6] Murchana Baruah, Bonny Banerjee, and Atulya K Nagar. Intent prediction in human–human interactions. *IEEE Transactions on Human-Machine Systems*, 53(2):458–463, 2023. 3

[7] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Behavior-driven synthesis of human dynamics. In *CVPR*, pages 12236–12246, 2021. 3

[8] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *IJCAI*, 2022. 3

[9] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *European Conference on Computer Vision*, pages 356–372. Springer, 2022. 3

[10] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. *arXiv preprint arXiv:2302.03665*, 2023. 3, 5

[11] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[12] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 2023. 3, 7

[13] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 1

[14] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, 2020. 3

[15] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 1, 3

[16] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *ACM MM*, pages 5162–5171, 2022. 3

[17] Hehe Fan, Yi Yang, and Mohan S. Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 14204–14213, 2021. 7

[18] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020. 4, 7

[19] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 1, 3

[20] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 3

[21] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018. 3

[22] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 3

[23] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 1, 3

[24] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020. 5, 6

[25] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 3

[26] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 3

[27] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017. 3

[28] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1279, 2022. 3

[29] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *CVPR*, pages 5226–5234, 2018. 3

[30] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 3

[31] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 1, 3, 4, 7, 8

[32] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 4

[33] Yunze Liu, Yun Liu, Che Jiang, Zhoujie Fu, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, He Wang, and Li Yi. HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. *arXiv e-prints*, 2022. 3

[34] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: quantization-based 3d human motion generation and forecasting. In *ECCV*, 2022. 3

[35] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6437–6446, 2022. 7

[36] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, pages 9489–9497, 2019. 3

[37] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, pages 2891–2900, 2017. 3

[38] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020. 4

[39] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 3

[40] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 1, 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[42] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. In *CVPR*, pages 6457–6466, 2022. 3

[43] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80, 2008. 6

[44] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3

[45] Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *ICCV*, 2021. 3

[46] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39 (4):54–1, 2020. 3

[47] Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamic. In *IJCAI*, 2018. 3

[48] Hao Wen, Yunze Liu, Jingwei Huang, Bo Duan, and Li Yi. Point primitive transformer for long-term 4d point cloud video understanding. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022. 6, 7

[49] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238, 2023. 1, 3

[50] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *ECCV*, 2022. 3

[51] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 1, 3, 5

[52] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019. 3

[53] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012. 4

[54] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 1, 7

[55] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. Artigrasp:

Physically plausible synthesis of bi-manual dexterous grasping and articulation. *arXiv preprint arXiv:2309.03891*, 2023. 1

[56] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 3

[57] Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, and Li Yi. Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2023. 1, 7

[58] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *CVPR*, 2022. 3

[59] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15741–15751, 2021. 1