Inter-X: Towards Versatile Human-Human Interaction Analysis

Liang Xu[1,2] Xintao Lv[1] Yichao Yan[1†] Xin Jin[2†] Shuwen Wu[1] Congsheng Xu[1] Yifan Liu[1]
Yizhou Zhou[3] Fengyun Rao[3] Xingdong Sheng[4] Yunhui Liu[4] Wenjun Zeng[2] Xiaokang Yang[1]

[1]Shanghai Jiao Tong University [2]Eastern Institute of Technology, Ningbo
[3]WeChat, Tencent Inc. [4]Lenovo

https://liangxuy.github.io/inter-x/

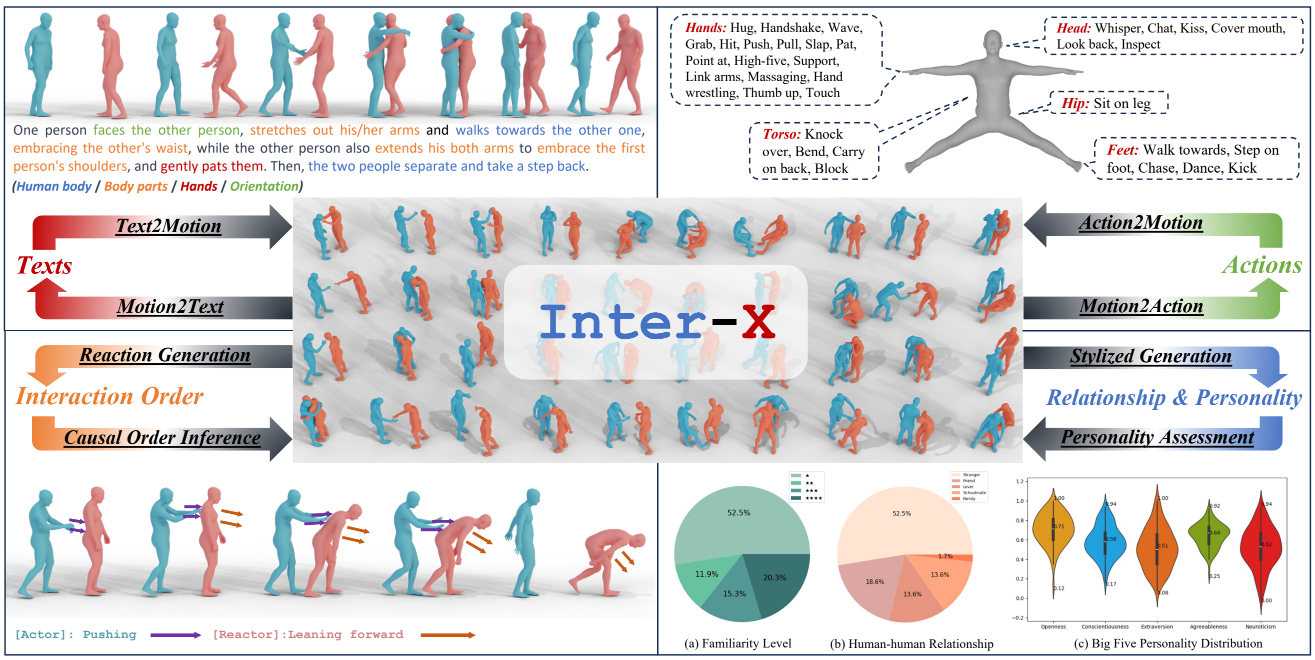


Figure 1. An overview of the data and task taxonomy of our proposed Inter-X dataset, which is a large-scale human-human interaction MoCap dataset with ∼**11K** interaction sequences and more than **8.1M** frames. The fine-grained textual descriptions, semantic action categories, interaction order, and relationship and personality annotations allow for 4 categories of downstream tasks.

Abstract

The analysis of the ubiquitous human-human interactions is pivotal for understanding humans as social beings. Existing human-human interaction datasets typically suffer from inaccurate body motions, lack of hand gestures and fine-grained textual descriptions. To better perceive and generate human-human interactions, we propose Inter-X, a currently largest human-human interaction dataset with accurate body movements and diverse interaction patterns, together with detailed hand gestures. The dataset includes ∼***11K** interaction sequences and more than **8.1M** frames. We also equip Inter-X with versatile annotations of more than **34K** fine-grained human part-level textual descriptions, semantic interaction categories, interaction order, and the relationship and personality of the subjects. Based on the elaborate annotations, we propose a unified benchmark composed of 4 categories of downstream tasks from both the perceptual and generative directions. Extensive experiments and comprehensive analysis show that Inter-X serves as a testbed for promoting the development of versatile human-human interaction analysis. Our dataset and benchmark will be publicly available for research purposes.*

†Corresponding authors

1. Introduction

The ability to perceive and generate human-human interactions is fundamental in constructing intelligent digital hu-

Dataset	Year	Motions	Frames	Texts	Scheme	Modality	Hands	Asyn.	Rel.& Pst.
UMPM [81]	2011	36	400K	✗	MoCap	Skel.	✗	✗	✗
SBU Kinect [94]	2012	300	7.5K	✗	RGB+D	Skel.	✗	✗	✗
You2Me [61]	2020	42	77K	✗	RGB+D	Skel.	✗	✗	✗
NTU120 [56]	2019	8,276	462K	✗	RGB+D	Skel.	✗	✗	✗
Chi3D [29]	2020	373	63K	✗	MoCap	SMPL-X	✓	✗	✗
ExPI [37]	2022	115	30K	✗	mRGB	Skel.	✗	✗	✗
Hi4D [93]	2023	100	11K	✗	mRGB	SMPL	✗	✗	✗
InterHuman [54]	2023	6,022	1.7M	16,756	mRGB	SMPL	✗	✗	✗
Inter-X	2023	**11,388**	**8.1M**	**34,164**	MoCap	SMPL-X	✓	✓	✓

Table 1. **Dataset comparisons.** We compare our Inter-X dataset with the existing human-human interaction datasets. **Motions**: The number of the motion clips; **Frames**: The frame number of the 3D human motions; **Texts**: The number of the textual descriptions; **Scheme**: The strategy to obtain the motion data; **Modality**: The representation of the motion data and "Skel." denotes skeleton; **Hands**, **Asyn.** and **Rel.&Pst.** refer to the components of hand gestures, asymmetry annotations, human-human relationships and personalities.

man systems, which have numerous applications in surveillance, AR/VR, games, and robotics. However, this task is challenging due to the complex and diverse interaction patterns, as well as self-occlusions. Although impressive progress has been made in the perception tasks, *i.e.*, skeleton-based interaction recognition [26, 44, 62, 64, 79], and the generation tasks, *i.e.*, action/text-conditioned interaction generation [34, 54, 65, 78, 87], they remain sub-optimal due to the lack of a comprehensive dataset to cover all the aspects of this task.

The advancement of human-human interaction analysis is accompanied by the construction of human-human interaction datasets [29, 37, 54, 56, 61, 81, 93, 94], as listed in Tab. 1. However, we believe that all the previous datasets remain unsatisfactory on the following aspects: 1) **Expressive ability,** *i.e.*, the dexterous hand gestures play important roles for human-human interactions, like "shaking hands", "grabbing", "waving", *etc.* However, to the best of our knowledge, there is no large-scale dataset providing high-fidelity finger movements for human-human interactions. 2) **Fine-grained text descriptions,** *i.e.*, text-driven generative tasks are promising for practical applications and have attracted much attention. Unlike coarse text annotations like "one person approaches the other and embraces her/him", fine-grained descriptions with human part-level semantics enable controllable interaction generation and better alignment [47] between motion and text modalities, spatiotemporally. 3) **Interaction order,** *i.e.*, during a causal human-human interaction period such as "kicking", the actor and reactor are asymmetric. However, the asymmetry property for human-human interactions is not considered in previous datasets. 4) **Relationship and personality,** *i.e.*, the intimacy level and social relationships between individuals together with their personalities intuitively affect the interaction patterns, which should be considered.

To address the aforementioned limitations of existing datasets, we thus build a large-scale human-human interaction dataset, called Inter-X, as depicted in Fig. 1, with precise, diverse human-human interaction sequences, and detailed hand gestures. To capture Inter-X, we first build a MoCap system with the combination of the optical scheme to capture accurate body movement and the inertial solution to record hand gestures against occlusion. Inter-X covers 40 daily interaction categories, $\sim$**11K** motion sequences with more than **8.1M** frames. We recruited 89 distinct subjects with different social relationships, *i.e.*, strangers, friends, lovers, schoolmates, and family members. We also collect their familiarity levels and their individual Big Five personalities [23, 82, 85].

With our proposed high-precision human-human interaction dataset and the versatile annotations, as illustrated in Fig. 1, we empower 4 categories of downstream tasks with half of them as generative tasks and the remaining as perceptive tasks. 1) **Texts** enable not only controllable human interaction generation from natural languages [54] but also the human interaction captioning tasks [35, 46]; 2) **Action categories** facilitate action-conditioned human interaction generation [87] together with the human interaction recognition tasks [26, 62]; 3) **Interaction order** enables the causal human reaction generation [21, 31, 57, 76] and the causal order inference tasks, *i.e.*, detecting the perpetrator in surveillance scenarios; 4) **Relationship and personality** make the stylized interaction generation [5, 43] and the personality assessment possible. We formulate our Inter-X dataset as a unified testing ground for all the downstream tasks. For the existing tasks, we extensively evaluate the state-of-the-art methods on the Inter-X's test set with extensive discussions. We also build up the baseline methods and evaluation metrics for the remaining tasks.

In summary, our contributions can be summarized as follows: 1) We collect the currently largest human-human interaction dataset with accurate human body movements, diverse interaction patterns, and expressive hand gestures; 2) We complement Inter-X with fine-grained human part-level textual descriptions, semantic action categories, causal interaction order annotations, relationship and personality information. 3) We propose a unified human-human interaction benchmark with 4 categories of downstream tasks to enable extensive research directions.

2. Related work

2.1. Human motion datasets

Compared to RGB videos, human motion representation is high-level, efficient, privacy-friendly and robust to illumination [56, 86]. Human motion datasets with action labels [45, 56, 69, 101] and text descriptions [34, 55, 67] facilitate the development for understanding human motions. Datasets accompanied with audio signals [53, 80] and scene/object conditions [9, 39, 40, 75, 84, 92, 98] are also produced for real-world human-centric tasks.

2.2. Human-human interaction datasets

Besides the single-human motion datasets, many human-human interaction datasets have been proposed [29, 37, 54, 56, 61, 81, 93, 94] as listed in Tab. 1 with various sizes, modalities and functionalities. Especially, InterHuman [54] was recently built as a large-scale human-human interaction dataset with textual annotations. However, as aforementioned, our Inter-X dataset still maintains advantages with respect to motion quality, fine-grained textual annotation, hand gestures, and comprehensive annotation modalities.

2.3. Perceptive tasks for human motion

Skeleton-based human action recognition has been a long-standing problem for years [18, 19, 30, 50, 51, 58, 72, 89, 96, 97, 99]. Compared to it, human interaction recognition [26, 44, 62, 64, 79] is a sub-field of it, relying on modeling the semantic correlations between humans. Besides human action recognition, human motions contain biometric cues about human subjects [24, 82]. Gait recognition [70, 83] aims to identify the individuals from human motions. Other works like[23, 27] regard the human movements as personality predictors. Our Inter-X dataset with large-scale action-motion and text-motion pairs will promote the development of human action recognition. We also take a significant step forward in assessing the human-human relationships and personalities from human motions.

2.4. Generative tasks for human motion

The goal of human motion generation is to generate plausible and diverse motion data based on different guidances. Human motion generation from action labels [16, 17, 32, 65, 78, 87], textual descriptions [6, 22, 33, 48, 55, 59, 66,

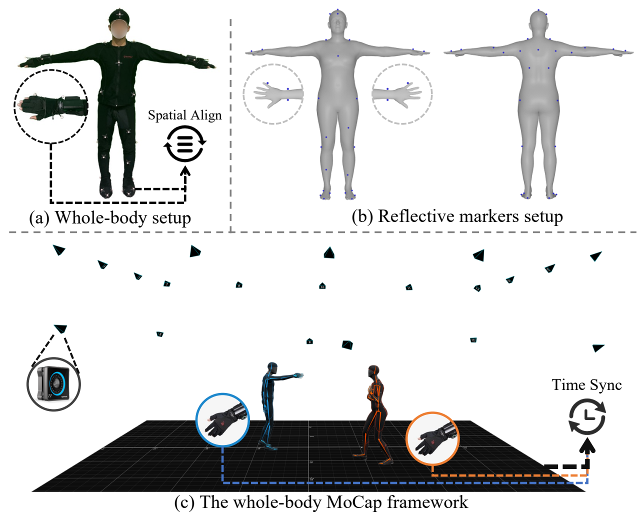


Figure 2. **An overview of the Inter-X capture system.** (a). The optical MoCap clothing together with the inertial gloves are spatially integrated via a triangular bracket of reflective markers; (b). The details of the markers setup; (c). The body and hands are temporally synchronized in the whole-body MoCap framework.

95] and audios [7, 8, 10, 38, 49, 52] have emerged in recent years. Besides single-person human motion generation, [54, 71, 87] attempt to generate multi-person interactions. Besides, a few works [21, 76] tackle the problem of generating the reaction between two interactions. To enhance the expressibility of the generated motions, [5, 43] manage to solve motion style transfer and stylized motion generation tasks. Our Inter-X dataset can be utilized for action or text-conditioned human interaction generation tasks. The explicit interaction order annotations greatly facilitate the reaction generation task. At the same time, personalities and relationships can serve as factors for stylized human interaction generation.

2.5. Multimodality in vision

The world surrounding us involves multiple modalities [12, 36, 88, 90, 91], so are the ubiquitous human-human interactions. Many multimodal datasets [54, 55, 77, 93] related to human motions emerged in recent years. Based on Inter-X, we unify several categories of downstream tasks towards a deeper understanding of human-human interactions.

3. The Inter-X Dataset

We present the large-scale Inter-X dataset towards versatile human-human interaction analysis, which consists of 11,388 interaction sequences and more than 8.1M frames, covering 40 daily interaction categories and 89 subjects.

3.1. Data Capturing System

Most of the previous datasets take the multi-view RGB-based technologies [54, 56], *i.e.*, extracting the human motion from RGB videos. Though the natural RGB images are captured, these datasets suffer from severe occlusions and penetrations, and the subtle finger movements are hard to obtain precisely. For the trade-off between accuracy and natural RGB images [75], we prioritize accuracy and thus choose the optical MoCap system for body movements. Additionally, we adopt inertial gloves to capture the finger gestures, which are robust to occlusions. The overview of our capturing system is illustrated in Fig. 2.

The length, width, and height of our MoCap venue are 8.5 meters, 5.4 meters, and 3.3 meters, which is capable of covering most daily human-human interactions. We deploy the OptiTrack MoCap system [3] with 20 PrimeX 22 infrared cameras. For each camera, we capture the resolution of 2048×1088 at 120 fps. The optical motion capture scheme ensures a ± 0.15mm error, much lower than the RGB camera scheme.

To capture the dexterous hand gestures without occlusion, we adopt the inertial solution of the commercial Noitom Perception Neuron Studio (PNS) gloves [2]. The subtle finger movements can be captured in real-time, disregarding the self-occlusion and occlusion with the other person during the interactions. We also re-calibrate the PNS gloves frequently to mitigate the error accumulation.

For each group of two volunteers, they wear the MoCap suits with 41 reflective markers and the inertial gloves as depicted in Fig. 2(a),(b). Both of them are carefully calibrated before they perform the interactions. We provide timecodes for the OptiTrack MoCap system and the PNS gloves so that the body and hands can be temporally synchronized. For each batch of the shoot, we arrange five action categories with five repetitions for variability, which improves efficiency and also ensures the continuity of the volunteers' actions. The volunteers pause for several seconds between two interaction snippets to ease the subsequent segmentation. More details of the data capturing processing can be found in the supplementary materials.

3.2. Data Postprocessing

The crux of the postprocessing is the alignment between the body poses from the OptiTrack MoCap system and the finger gestures from the inertial gloves. Temporally, we retrieve the intersection of the body pose and hand pose sequences. Spatially, they are naturally integrated through the shared wrist rotation from the triangular locating bracket. Given the spatiotemporally aligned motion sequences, the annotators should segment the start and end frames for each atomic interaction snippet. We collect, check the temporal segmentation results, and then trim the long recorded motion sequences into atomic segments.

4. Dataset Taxonomy

We enrich the high-precision human-human interaction sequences with multifaceted modalities, resulting in 13,888 pairs of SMPL-X [63] motion sequences, 273,312 synthetic multi-view RGB videos, 34,164 detailed text descriptions, 40 semantic action categories with diverse action/reaction patterns, interaction order labels, and the relationship for 59 groups and personality for 89 volunteers. Fig. 3 shows some characteristics of the Inter-X dataset.

4.1. Interaction data

MoCap Data. We adopt the SMPL-X parametric model for its expressivity for human body poses and articulated hand poses, and the generality for various downstream tasks. Formally, the SMPL-X parameter is composed of the body pose parameters $\theta \in \mathbb{R}^{N\times 55\times 3}$, shape parameters $\beta \in \mathbb{R}^{N\times 10}$ and the translation parameters $t \in \mathbb{R}^{N\times 3}$, where N is the number of the frames. We initialize the shape parameters β based on the height and the weight of the volunteer as [68]. Then an optimization algorithm is well-tuned to fit the SMPL-X parameters based on the captured key points:

$$E(\theta,t) = \lambda_1 \frac{1}{N} \sum_{j\in\mathcal{J}} \lambda_p ||\boldsymbol{J}_j(\mathbb{M}(\theta,t)) - \boldsymbol{g}_j||_2^2 + \lambda_2 ||\theta||_2^2,$$

where $\mathcal{J}$ denotes the joints set, $\mathbb{M}$ is the SMPL-X parametric model, $\boldsymbol{J}_j$ is the joint regressor function for joint j, $\boldsymbol{g}$ is the skeleton captured from the MoCap system. λ_1, λ_2 and λ_p are different weights and we apply different weights for different body parts. Please refer to the supplementary materials for more details.

Rendered RGB. The synthetic data has broad applications for human motions [13, 15, 28, 87]. To enrich our Inter-X dataset with RGB modality, we utilize the Unreal Engine to render multi-view 2D videos similar to [77]. We download the free character models from Renderpeople [4], and then retarget our full-body interaction data to the rigged characters. We select the realistic scene models from the Unreal Engine Store and then place the Renderpeople models into them. We capture multi-view videos with 6 rounded cameras, with a resolution of 1920×1080 and a frame rate of 30 fps. Ultimately, 273,312 synthesized RGB videos with 11,388 interaction sequences, 4 different scenes and 6 viewpoints are generated.

4.2. Action categories

We choose the action categories referring to the existing human-human interaction datasets [29, 54, 56] and large language models [14]. Finally, we figure out 40 daily human-human interaction categories, which cover the most interaction categories to the best of our knowledge. We instruct each volunteer to perform *naturally* and *diversely*. For diversity, the volunteers can perform 1) Diverse actions,

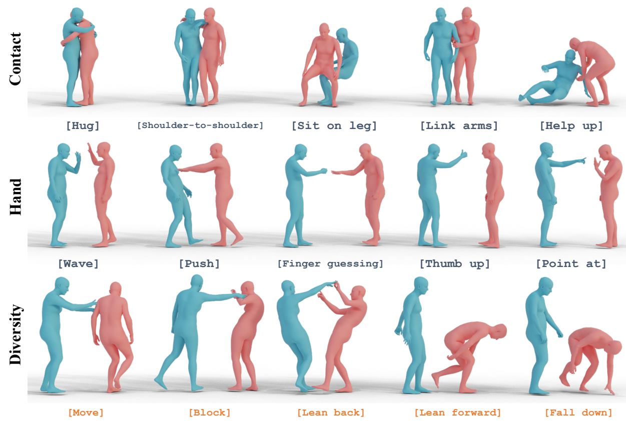


Figure 3. **More examples of the Inter-X dataset.** Our proposed Inter-X dataset for human-human interaction analysis is highly accurate, hand gestures incorporated, with diverse actions and reactions. Please zoom in for the details.

i.e., raising left hand, right hand, or both hands when "raising hands"; 2) Diverse reactions, *i.e.*, rebelling, taking a few steps back or falling down when being "pushed"; 3) Diverse human boy states, *i.e.*, standing, sitting, crouching or even lying on the ground. Each interaction is repeated five times for variability.

4.3. Text descriptions

Textual descriptions, especially fine-grained ones, empower various practical applications for better perception and generation. We implement an annotation tool based on [1], so that the annotators can scale and rotate the view for 360 degrees to observe the details of the interactions. For each interaction sequence, we ask 3 distinct annotators to describe it from human part levels with 1) the coarse body movements, 2) the finger movements, and 3) the relative orientations. We correct the typos of the collected textual descriptions with GPT-3.5 [14] and then spot-check the results. Upon analysis, the average length of our textual descriptions is ∼35, which significantly surpasses existing action datasets, reflecting the fine-grained nature of our texts.

4.4. Interaction Order

The study of causal relationships, where one person acts and the other one reacts, could help extend the understanding of human-human interactions [94]. We ask the volunteers to explicitly annotate the order of the actors and reactors for each atomic interaction sequence.

4.5. Relationship & Personality

Exploring the correspondence between human motion and personality is a niche [23, 27], and the essence lies in the disentanglement of the personality factors from motions. We adopt the dominant paradigm of the Big-Five Personality Model [23, 82, 85]. The participants are asked to fill out the NEO Five-Factor Inventory [60] to measure their personalities of openness, conscientiousness, extraversion, agreeableness and neuroticism. The volunteers also fill out the questionnaire to rank their familiarity level from levels 1 to 4, and declare their social relationships of 5 categories, *i.e.*, strangers, friends, lovers, schoolmates, and family.

5. Task Taxonomy

Our high-precision human-human interaction MoCap data with dexterous hand details bring vitality and challenge to existing tasks. Moreover, we also propose different downstream tasks with practical applications tailored to the versatile annotations. Formally, we denote each human-human interaction sequence as $\boldsymbol{m}{=}{<}\boldsymbol{x},\boldsymbol{y}{>}$, and the annotations as action category $\boldsymbol{l}_a$, text description $\boldsymbol{l}_t$, causal interaction order $\boldsymbol{l}_c$, relationship $\boldsymbol{l}_r$ and personalities $\boldsymbol{l}_p{=}{<}\boldsymbol{l}_{p_x},\boldsymbol{l}_{p_y}{>}$.

5.1. Texts related Tasks

Text-conditioned human interaction generation. Text-conditioned single-person human motion generation has been widely explored with various datasets [34, 55, 67] and models. We pose opportunities for controllable human-human interaction generation [47, 55] with fine-grained textual annotations and challenges to synthesize the subtle hand gestures and the alignment between human part-level textual descriptions and interactions. The task can be represented as learning a function F_{t2m}:

$$F_{t2m}(\boldsymbol{l}_t) \mapsto \boldsymbol{m}. \tag{1}$$

Human interaction captioning. Human interaction captioning is a newly proposed task [35, 46], to generate corresponding textual descriptions rather than recognizing the action category given a human-human interaction sequence, which can boost the alignment between texts and motion data and automatically generate diverse and reasonable textual descriptions. This task can be formulated as:

$$F_{m2t}(\boldsymbol{m}) \mapsto \boldsymbol{l}_t. \tag{2}$$

5.2. Actions related Tasks

Action-conditioned human interaction generation. Given an action label, $F_{a2m}(\cdot)$ aims to generate diverse and plausible human-human interaction sequences [65, 78, 87]. With our proposed Inter-X, we can generate more realistic and detailed interactions with fingers:

$$F_{a2m}(\boldsymbol{l}_a) \mapsto \boldsymbol{m}. \tag{3}$$

Human interaction recognition. Human interaction recognition has practical applications for visual surveillance [26, 62]. We believe that integrating the fine hand movements will enhance the recognition ability of current models. We formulate this task as:

$$F_{m2a}(\boldsymbol{m}) \mapsto \boldsymbol{l}_a. \tag{4}$$

5.3. Interaction-order related Tasks

Human reaction generation. Human reaction generation [21, 31, 57, 76] is less explored yet with broad applications in AR/VR and gaming. Explicit annotations of the actor-reactor order will advance the research on the asymmetry of different roles with human-human interactions:

$$F_{c2m}(\boldsymbol{l}_c, \boldsymbol{x}) \mapsto \boldsymbol{y}. \tag{5}$$

Causal order inference. $F_{m2c}(\cdot)$ aims to differentiate the actor and reactor given a human interaction sequence, which will benefit intelligent surveillance and sports:

$$F_{m2c}(\boldsymbol{m}) \mapsto \boldsymbol{l}_c. \tag{6}$$

5.4. Relationship & Personality related Tasks

Stylized human interaction generation. The relationship between two participants and their personalities can serve as stylization factors for customized human interaction generation. The large number of participants with each having a long sequence of motion data enable us to accomplish this task. We formulate this task as:

$$F_{s2m}(\boldsymbol{l}_a, \boldsymbol{l}_r, \boldsymbol{l}_p) \mapsto \boldsymbol{m}. \tag{7}$$

Personality assessment. Previous works [23, 27] regard the body movements of participants as personality predictors. Leveraging our Inter-X dataset, we propose a new task of personality and relationship assessment, which is vital for education, medicine, sports, *etc*. Specifically,

$$F_{m2s}(\boldsymbol{m}) \mapsto \{\boldsymbol{l}_r, \boldsymbol{l}_p\}. \tag{8}$$

6. Experiments

We extensively evaluate the state-of-the-art methods on the Inter-X dataset for the proposed downstream tasks with detailed discussion and analysis. In the main manuscript, we present four appealing tasks: 1) text-conditioned human interaction generation; 2) action-conditioned human interaction generation; 3) human reaction generation; and 4) human interaction recognition. The remaining experiments are presented in the supplementary materials.

6.1. Text-conditioned Interaction Generation

The detailed textual annotations combined with the human-human interaction sequences allow for human interaction generation. We extensively evaluate 6 state-of-the-art text to motion models, *i.e.*, TEMOS [66], T2M [34], MDM [78], MDM-GRU [20, 78], ComMDM [71] and InterGen [54]. We modify the input and output dimensions to extend the single-person models to two-person settings and change the motion representation to SMPL-X [63] parameters.

Experiment setup. We adopt the same protocol of [34, 54] to split our dataset into training, test, and validation sets with a ratio of 0.8, 0.15, and 0.05. Following [11], we directly borrow the SMPL-X parameters of Inter-X rather than the manually designed motion representation as in [34, 54]. Different from single-person motion sequences that are canonicalized to the first frame, we keep the global translation of the interacted persons so that their relative positions are reserved. For all the methods, we adopt the 6D continuous rotation representation [100] as previous works [34, 54, 65, 78]. For the diffusion-based models [42, 73], we train them with 1,000 noising timesteps and run 5 DDIM [74] sampling steps. Each model is trained on 4 NVIDIA A100 GPUs.

Evaluation metrics. We follow [34] to adopt the Frechet Inception Distance (FID) [41] to measure the latent distance between real and generated samples, diversity to measure latent variance, multimodality (MModality) to measure the diversity of the generated results for the same text, R Precision to measure the top-1, top-2 and top-3 accuracy of retrieving the ground-truth description from 31 randomly mismatched descriptions, and MultiModal distance (MM Dist) to calculate the latent distance between generated motions and texts. We train a motion feature extractor together with a text feature extractor in a contrastive manner to better align the features of texts and motions. We run all the evaluations 20 times (except MModality for 5 times) and report the averaged results with the confidence interval at 95%.

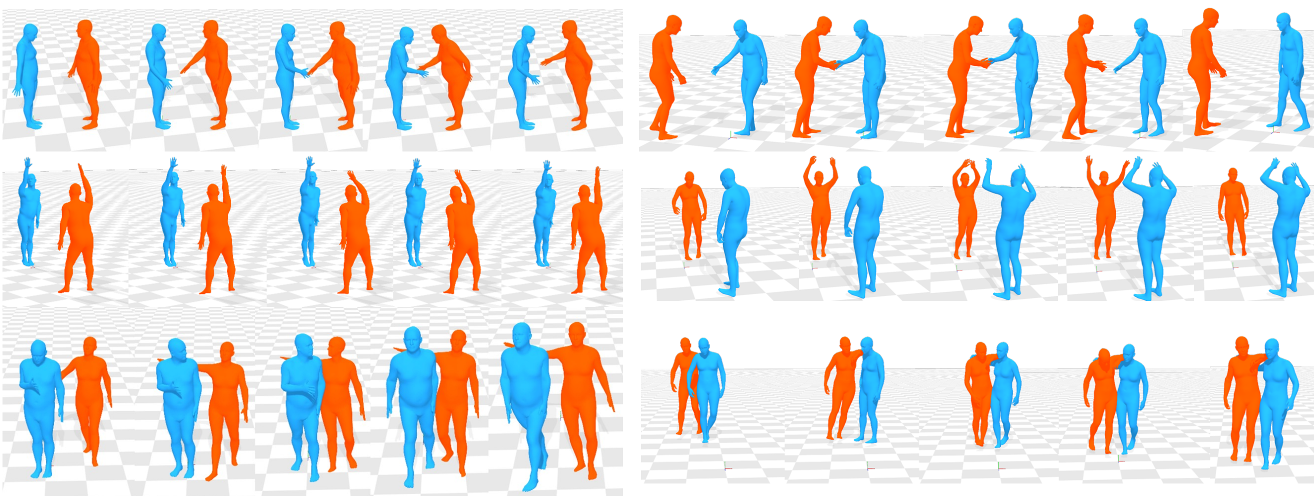
Quantitative results. The experimental results are depicted in Tab. 2. We can derive that InterGen [54] achieves state-of-the-art performance except for the MM Dist metric while ComMDM [71] achieves the worst R Precision scores. One possible explanation could be that ComMDM requires extra pre-training. From the results, we derive that our Inter-X dataset has the potential for further explorations.

Qualitative results. We demonstrate the human-human interaction results generated from InterGen [54] together with the generated results for the InterHuman dataset for visual comparisons in Fig. 4. The visualization results show that with our Inter-X, the expressibility of the human-human interaction is highly enhanced with detailed hand movements. Since InterHuman does not provide dexterous hand gestures, the generated results for “Handshake”, “Wave” and “Shoulder to shoulder” are unplausible. Besides, the synthesized results of InterHuman contain occlusions and penetrations, while ours are much more precise.

Please refer to the supplementary materials for more visual comparisons and **video** results.

Methods	R Precision↑			FID ↓	MM Dist↓	Diversity→	MModality ↑
	Top 1	Top 2	Top 3				
Real	$0.429^{\pm0.004}$	$0.626^{\pm0.003}$	$0.736^{\pm0.003}$	$0.002^{\pm0.0002}$	$3.536^{\pm0.013}$	$9.734^{\pm0.078}$	-
TEMOS [66]	$0.092^{\pm0.003}$	$0.171^{\pm0.003}$	$0.238^{\pm0.002}$	$29.258^{\pm0.0694}$	$6.867^{\pm0.013}$	$4.738^{\pm0.078}$	$0.672^{\pm0.041}$
T2M [34]	$0.184^{\pm0.010}$	$0.298^{\pm0.006}$	$0.396^{\pm0.005}$	$5.481^{\pm0.3820}$	$9.576^{\pm0.006}$	$5.771^{\pm0.151}$	$2.761^{\pm0.042}$
MDM [78]	$0.203^{\pm0.009}$	$0.329^{\pm0.007}$	$0.426^{\pm0.005}$	$23.701^{\pm0.0569}$	$9.548^{\pm0.014}$	$5.856^{\pm0.077}$	$3.490^{\pm0.061}$
MDM(GRU) [78]	$0.179^{\pm0.006}$	$0.299^{\pm0.005}$	$0.387^{\pm0.007}$	$32.617^{\pm0.1221}$	$9.557^{\pm0.019}$	$7.003^{\pm0.134}$	$3.430^{\pm0.035}$
ComMDM [71]	$0.090^{\pm0.002}$	$0.165^{\pm0.004}$	$0.236^{\pm0.004}$	$29.266^{\pm0.0668}$	$\mathbf{6.870^{\pm0.017}}$	$4.734^{\pm0.067}$	$0.771^{\pm0.053}$
InterGen [54]	$\mathbf{0.207^{\pm0.004}}$	$\mathbf{0.335^{\pm0.005}}$	$\mathbf{0.429^{\pm0.005}}$	$\mathbf{5.207^{\pm0.2160}}$	$9.580^{\pm0.011}$	$\mathbf{7.788^{\pm0.208}}$	$\mathbf{3.686^{\pm0.052}}$

Table 2. Experimental results of text-conditioned interaction generation on the Inter-X dataset, where $\pm$ indicates 95% confidence interval and $\rightarrow$ means the closer the better. **Bold** indicates best results.



(1) InterHuman (2) Inter-X

Figure 4. **Visualization results** of the generated results on the InterHuman [54] and Inter-X dataset via ait-viewer [1]. From top to bottom, the action categories are "Handshake", "Wave" and "Shoulder to shoulder", respectively. Please zoom in for the details.

6.2. Action-conditioned Interaction Generation

Inter-X contains 40 semantic action categories, which are currently the largest compared to other human-human interaction datasets. We conduct experiments of action-conditioned human interaction generation with the state-of-the-art methods, *i.e*., Action2Motion [32], ACTOR [65], MDM [78], MDM-GRU [20, 78] and Actformer [87]. Same as the text-conditioned methods, we re-implement these methods to adapt to our dataset format. We adopt the same dataset split protocol and pose representation as the text-conditioned methods.

Evaluation metrics. Similar to the previous works [32, 65, 78] for human motion generation, we also adopt the Frechet Inception Distance (FID) [41], action recognition accuracy, diversity, and multi-modality for evaluation. For all these metrics, we train an action recognition model [89] for feature extraction as in previous works. We generate 1,000 samples 20 times and report the average score with a confidence score of 95%.

Quantitative results. From the experimental results in Tab. 3, Actformer [88] achieves the best FID and action recognition accuracy, MDM [78] achieves the best Multimod. score and MDM-GRU [20, 78] yields the best diversity score. Although the interaction transformer is designed to model the interaction between persons, there is still substantial potential for further improvements.

6.3. Human Reaction Generation

We explicitly annotate the interaction order for causal human interactions, *i.e*., human reaction generation. We select the MDM [78], MDM-GRU [20, 78], RAIG [76] and AGRoL [25] models for evaluation. We modify the architecture of all these methods so that the motion of the actor serves as the input conditions into the model, and the output is the human reaction.

Quantitative results. We demonstrate the quantitative results in Tab. 4. We observe that AGRoL [25] yields the best

Method	FID↓	Acc.↑	Div.→	Multimod.→
Real	$0.281^{\pm 0.002}$	$0.990^{\pm 0.0000}$	$12.890^{\pm 0.028}$	$22.391^{\pm 0.195}$
Action2Motion [32]	$20.295^{\pm 12.081}$	$0.766^{\pm 0.0003}$	$11.581^{\pm 0.024}$	$15.345^{\pm 0.245}$
ACTOR [65]	$9.392^{\pm 0.816}$	$0.855^{\pm 0.0003}$	$11.594^{\pm 0.029}$	$15.327^{\pm 0.195}$
MDM [78]	$12.426^{\pm 2.584}$	$0.896^{\pm 0.0004}$	$13.492^{\pm 0.033}$	$\mathbf{22.042^{\pm 0.153}}$
MDM(GRU) [78]	$35.003^{\pm 7.876}$	$0.716^{\pm 0.0006}$	$\mathbf{12.579^{\pm 0.038}}$	$16.456^{\pm 0.100}$
Actformer [87]	$\mathbf{8.067^{\pm 0.653}}$	$\mathbf{0.945^{\pm 0.0007}}$	$12.512^{\pm 0.05}$	$16.187^{\pm 0.189}$

Table 3. Experimental results of action-conditioned interaction generation on the Inter-X dataset. **Bold** for best results.

Method	FID↓	Acc.↑	Div.→	Multimod.→
Real	$0.260^{\pm 0.0021}$	$0.988^{\pm 0.0000}$	$12.115^{\pm 0.031}$	$21.498^{\pm 0.131}$
MDM [78]	$6.747^{\pm 0.3153}$	$0.903^{\pm 0.0001}$	$12.264^{\pm 0.051}$	$19.681^{\pm 0.234}$
MDM(GRU) [78]	$19.968^{\pm 1.1700}$	$0.752^{\pm 0.0003}$	$12.351^{\pm 0.049}$	$18.056^{\pm 0.156}$
RAIG [76]	$6.372^{\pm 0.2154}$	$0.908^{\pm 0.0001}$	$12.330^{\pm 0.060}$	$20.071^{\pm 0.299}$
AGRoL [25]	$\mathbf{4.386^{\pm 0.2186}}$	$\mathbf{0.925^{\pm 0.0001}}$	$\mathbf{12.204^{\pm 0.042}}$	$\mathbf{20.199^{\pm 0.226}}$

Table 4. Experimental results of human reaction generation based on action labels on the Inter-X dataset. **Bold** for best results.

Method	Top-1 (%)	Top-5 (%)
ST-GCN [89]	64.62	90.16
2s-AGCN [72]	75.22	93.73
HD-GCN [50]	77.40	94.73
CTR-GCN [18]	82.19	96.72
MS-G3D [58]	**83.30**	**97.09**

Table 5. Experimental results of skeleton-based human interaction recognition on the Inter-X dataset. **Bold** for best results.

performance for all the evaluation metrics, while the GRU architecture achieves the worst results.

6.4. Human Interaction Recognition

Inter-X is built from the MoCap system with accurate 3D skeleton data. We evaluate five state-of-the-art skeleton-based action recognition models as ST-GCN [89], 2s-AGCN [72], HD-GCN [50], CTR-GCN [18] and MS-G3D [58] and report the results of Top-1 and Top-5 recognition accuracy in Tab. 5. Note that for simplicity, we only employed the skeleton joint stream without ensembling with bone stream and motion streams [58, 72].

Quantitative results. From the results, we can observe that MS-G3D [58] achieves the best Top-1 accuracy of 83.30%, which is not satisfactory. One possible reason is that Inter-X contains dexterous hand gestures and action/reaction diversities, which would pose new challenges and opportunities for further research works.

7. Conclusion and Limitation

In this paper, we propose Inter-X, a large-scale human-human interaction dataset with high-precision human body movements, diverse interaction patterns, and subtle hand gestures. We also annotate Inter-X with human-part level textual descriptions from different perspectives, the semantic interaction categories, the interaction order, and the relationship and personalities of the subjects to facilitate 4 categories of downstream tasks. The qualitative and quantitative results show that Inter-X poses challenges for human-human interaction related perceptual and generative tasks.

Limitations. Our work has some limitations in the following aspects: 1) **Facial expressions:** Inter-X dataset is created through an indoor MoCap venue and non-professional actors. Thus facial expressions are not involved since the correlation between expression and motion is unreliable. A possible alternative is referring to natural outdoor scenes or professional actors to explore the correlation between emotion and interactions; 2) **Atomic interactions:** The Inter-X dataset contains 11,388 atomic human-human interaction sequences, rather than long human-human interaction sequences. We acknowledge that real-world interactions are much more complicated with longer durations and frequent transitions. However, we believe that our dataset with high precision and diversity can still serve as a cornerstone for more complicated human-human interaction analysis.

Acknowledgments: This work is supported by NSFC (62201342, 62101325), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), NSFC under Grant 62302246 and ZJNSFC under Grant LQ23F010008, and supported by High Performance Computing Center at Eastern Institute of Technology, Ningbo.

References

[1] Aitviewer. https://eth-ait.github.io/aitviewer/. 5, 7

[2] Noitom. https://noitom.com/. 4

[3] Optitrack. https://optitrack.com/. 4

[4] Renderpeople. https://renderpeople.com/. 4

[5] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64–1, 2020. 2, 3

[6] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *3DV*, pages 719–728. IEEE, 2019. 3

[7] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *TOG*, 41(6):1–19, 2022. 3

[8] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Trans. Graph.*, 2023. 3

[9] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *CVPR*, pages 21211–21221, 2023. 3

[10] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *arXiv preprint arXiv:2111.12159*, 2021. 3

[11] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. *arXiv preprint arXiv:2305.09662*, 2023. 6

[12] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *PAMI*, 41(2):423–443, 2018. 3

[13] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, 2023. 4

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 4, 5

[15] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 4

[16] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *ECCV*, pages 356–372. Springer, 2022. 3

[17] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. *arXiv preprint arXiv:2212.04048*, 2022. 3

[18] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 3, 8

[19] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *ECCV*, pages 536–553. Springer, 2020. 3

[20] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 6, 7

[21] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 2023. 2, 3, 6

[22] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. *arXiv preprint arXiv:2212.04495*, 2022. 3

[23] David Delgado-Gómez, Antonio Eduardo Masó-Besga, David Aguado, Victor J Rubio, Aaron Sujar, and Sofia Bayona. Automatic personality assessment through movement analysis. *Sensors*, 22(10):3949, 2022. 2, 3, 5, 6

[24] Fani Deligianni, Yao Guo, and Guang-Zhong Yang. From emotions to mood disorders: A survey on gait analysis methodology. *IEEE journal of biomedical and health informatics*, 23(6):2302–2316, 2019. 3

[25] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. *arXiv preprint arXiv:2304.08577*, 2023. 7, 8

[26] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. c: Towards skeleton-based action recognition in the wild. In *ICCV*, pages 13634–13644, 2023. 2, 3, 5

[27] Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff, and Norman I Badler. Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics (TOG)*, 36(1):1–16, 2016. 3, 5, 6

[28] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, pages 430–446, 2018. 4

[29] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, 2020. 2, 3, 4

[30] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. Unified pose sequence modeling. In *CVPR*, pages 13019–13030, 2023. 3

[31] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: Reactive 3d motion synthesis for two-person interactions. *arXiv preprint arXiv:2311.17057*, 2023. 2, 6

[32] 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 *ACM Multimedia*, pages 2021–2029. ACM, 2020. 3, 7, 8

[33] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 3

[34] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 2, 3, 5, 6, 7

[35] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, pages 580–597. Springer, 2022. 2, 5

[36] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019. 3

[37] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *CVPR*, pages 13053–13064, 2022. 2, 3

[38] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *SIGGRAPH*, pages 1–9, 2022. 3

[39] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019. 3

[40] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, pages 11374–11384, 2021. 3

[41] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017. 6, 7

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 6

[43] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022. 2, 3

[44] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *2014 IEEE international conference on multimedia and expo workshops (ICMEW)*, pages 1–6. IEEE, 2014. 2, 3

[45] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *ACMMM*, page 1510–1518, 2018. 3

[46] 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. 2, 5

[47] Sai Shashank Kalakonda, Shubh Maheshwari, and Ravi Kiran Sarvadevabhatla. Action-gpt: Leveraging large-scale language models for improved and generalized action generation. In *ICME*, pages 31–36. IEEE, 2023. 2, 5

[48] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. *arXiv preprint arXiv:2209.00349*, 2022. 3

[49] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *NeurIPS*, 32, 2019. 3

[50] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*, pages 10444–10453, 2023. 3, 8

[51] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*, pages 10444–10453, 2023. 3

[52] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. *AAAI*, 36(2):1272–1279, 2022. 3

[53] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 3

[54] 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. 2, 3, 4, 6, 7

[55] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *arXiv preprint arXiv:2307.00818*, 2023. 3, 5

[56] 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. *T-PAMI*, 42(10):2684–2701, 2019. 2, 3, 4

[57] Yunze Liu, Changxi Chen, and Li Yi. Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. *arXiv preprint arXiv:2312.08983*, 2023. 2, 6

[58] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 3, 8

[59] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 3

[60] Robert R McCrae and Paul T Costa Jr. A contemplated revision of the neo five-factor inventory. *Personality and individual differences*, 36(3):587–596, 2004. 5

[61] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*, pages 9890–9900, 2020. 2, 3

[62] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *ECCV*, pages 605–622. Springer, 2022. 2, 3, 5

[63] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 4, 6

[64] Mauricio Perez, Jun Liu, and Alex C Kot. Interaction relational network for mutual action recognition. *IEEE Transactions on Multimedia*, 24:366–376, 2021. 2, 3

[65] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *CVPR*, pages 10985–10995, 2021. 2, 3, 5, 6, 7, 8

[66] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022. 3, 6, 7

[67] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 3, 5

[68] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements. *IEEE transactions on visualization and computer graphics*, 25 (5):1887–1897, 2019. 4

[69] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 3

[70] Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *PAMI*, 45(1):264–284, 2022. 3

[71] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3, 6, 7

[72] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 3, 8

[73] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 6

[74] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[75] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600, 2020. 3, 4

[76] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *ICCV*, pages 15999–16009, 2023. 2, 3, 6, 7, 8

[77] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. In *CVPR*, pages 22106–22117, 2023. 3, 4

[78] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3, 5, 6, 7, 8

[79] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *ICCV*, pages 1729–1736. IEEE, 2011. 2, 3

[80] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: Probabilistic Autoregressive Dance Generation With Multimodal Attention. *ACM Trans. Graph.*, 2021. 3

[81] NP Van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T Tan, and Remco C Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *ICCV Workshops*, pages 1264–1269. IEEE, 2011. 2, 3

[82] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014. 2, 3, 5

[83] Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5): 1–35, 2018. 3

[84] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *NIPS*, 2022. 3

[85] Jerry S Wiggins. *The five-factor model of personality: Theoretical perspectives*. Guilford Press, 1996. 2, 5

[86] Liang Xu, Cuiling Lan, Wenjun Zeng, and Cewu Lu. Skeleton-based mutually assisted interacted object localization and human action recognition. *IEEE Transactions on Multimedia*, 2022. 3

[87] 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 *ICCV*, pages 2228–2238, 2023. 2, 3, 4, 5, 7, 8

[88] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *PAMI*, 2023. 3, 7

[89] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452. AAAI Press, 2018. 3, 7, 8

[90] Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022. 3

[91] Xingyi Yang, Daquan Zhou, Songhua Liu, Jingwen Ye, and Xinchao Wang. Deep model reassembly. *Advances in neural information processing systems*, 35:25739–25753, 2022. 3

[92] Yan Yichao, Cheng Yuhao, Chen Zhuo, Peng Yicong, Wu Sijing, Zhang Weitian, Li Junjie, Li Yixuan, Gao Jingnan, Zhang Weixia, Zhai Guangtao, and Yang Xiaokang. A survey on generative 3d digital humans based on neural net-

works: representation, rendering, and learning. *SCIENTIA SINICA Informationis*, pages 1858–, 2023. 3

[93] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, 2023. 2, 3

[94] 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 *CVPR*, pages 28–35. IEEE, 2012. 2, 3, 5

[95] 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. 3

[96] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, pages 1112–1121, 2020. 3

[97] Xikun Zhang, Chang Xu, and Dacheng Tao. Context aware graph convolution for skeleton-based action recognition. In *CVPR*, pages 14333–14342, 2020. 3

[98] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: towards controllable human-chair interactions. In *ECCV*, pages 518–535, 2022. 3

[99] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *CVPR*, pages 10608–10617, 2023. 3

[100] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. 6

[101] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 3d human shape reconstruction from a polarization image. In *ECCV*, pages 351–368, 2020. 3