ActFormer: A GAN-based Transformer towards General Action-Conditioned 3D Human Motion Generation

Liang Xu[2,3*] Ziyang Song[5*] Dongliang Wang[1,6] Jing Su[1] Zhicheng Fang[1] Chenjing Ding[1]
Weihao Gan[7] Yichao Yan[2] Xin Jin[3,4] Xiaokang Yang[2] Wenjun Zeng[3,4] Wei Wu[1,6]
[1]SenseTime Research [2]Shanghai Jiao Tong University [3]Eastern Institute of Technology, Ningbo
[4]Ningbo Institute of Digital Twin [5]The Hong Kong Polytechnic University [6]Shanghai AI Laboratory
[7]Mashang Consumer Finance Co., Ltd.

Abstract

We present a GAN-based Transformer for general action-conditioned 3D human motion generation, including not only single-person actions but also multi-person interactive actions. Our approach consists of a powerful Action-conditioned motion TransFormer (ActFormer) under a GAN training scheme, equipped with a Gaussian Process latent prior. Such a design combines the strong spatio-temporal representation capacity of Transformer, superiority in generative modeling of GAN, and inherent temporal correlations from the latent prior. Furthermore, ActFormer can be naturally extended to multi-person motions by alternately modeling temporal correlations and human interactions with Transformer encoders. To further facilitate research on multi-person motion generation, we introduce a new synthetic dataset of complex multi-person combat behaviors. Extensive experiments on NTU-13, NTU RGB+D 120, BABEL and the proposed combat dataset show that our method can adapt to various human motion representations and achieve superior performance over the state-of-the-art methods on both single-person and multi-person motion generation tasks, demonstrating a promising step towards a general human motion generator. The project website can be found at https://liangxuy.github.io/actformer/.

1. Introduction

This work aims to tackle the action-conditioned motion generation task. Specifically, given a semantic action label as input and generate corresponding 3D human motions. The technique is key to applications like character animation creation, humanoid robots interaction and data synthesis for computer vision tasks related to human actions.

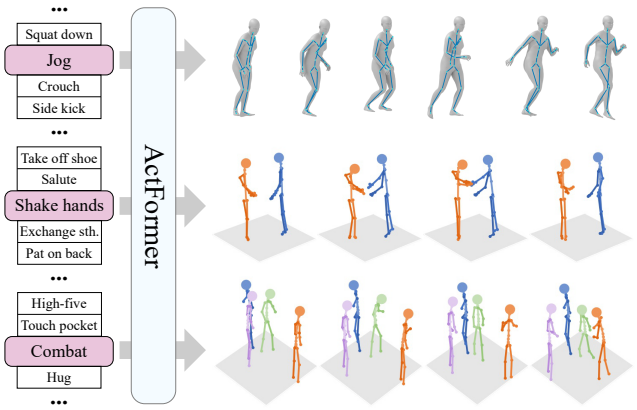

Figure 1. **Towards general action-conditioned 3D human motion generation.** Our framework adapts to more action categories, various human motion representations (*e.g.*, SMPL body models, skeleton joint coordinates), and multi-person interactive actions.

Human motion synthesis has been a long-standing research topic. However, most of the prior works are closer to a prediction task, in which future motions are generated from previous motions [16, 26, 41, 8, 22, 11, 55, 34]. In recent years, some works started to focus on motion generation from action labels [20, 45, 49]. Despite some impressive generation results, these works are still limited in the following two aspects. Firstly, most of these works bias towards motion data of SMPL pose parameters while performing poorly on data of skeleton joint coordinates, which limits the generalization. A solution adaptable to various human motion representations is thus expected. Secondly, prior works only focus on single-person motion generation while neglecting multi-person interactive actions, which are integral parts of daily human motions. In general, prior works fail to cover a complete domain of human motions and stand far from a general human motion generator.

*Denotes equal contribution. Work done when Liang and Ziyang were at SenseTime. Corresponding authors: Dongliang Wang (wangdongliang@senseauto.com), Yichao Yan (yanyichao@sjtu.edu.cn) and Xin Jin (jinxin@eias.ac.cn).

This paper explores a solution towards general action-conditioned human motion generation, as shown in Fig. 1. The very first challenge lies in generating long motion sequences with realism and diversity. Many prior works assume a Markovian dependency in temporal motions and adopt an auto-regressive model [58, 20, 5, 6, 31]. However, these methods are subject to the "mean-pose" problem, in which the model starts to generate the mean pose continuously after a few frames. In contrast, CSGN [53] and ACTOR [45] sample from a sequence-level latent prior and produce the whole sequence altogether. Specifically, CSGN samples from a Gaussian Process (GP) latent prior and stacks convolutions in the generator to enforce temporal correlations. On the other hand, ACTOR samples a single vector as the sequence-level embedding and produces multiple frames by querying through different positional encodings. We argue that both are sub-optimal solutions, and we seek a better trade-off between inductive bias and representation capacity. Our proposed **Act**ion-Conditioned Motion Trans**Former** (**ActFormer**) leverages the GP prior for the inherent temporal correlations. Meanwhile, we adopt a Transformer architecture for its simple structure and strong power in encoding non-local correlations proved in many other tasks. The Transformer model naturally regards a latent vector sequence from the GP prior as a sequence of tokens, leading to a seamless conjunction. We incorporate the Transformer-based motion generator into GAN, known for high-quality generative modeling. These designs jointly contribute to significant advantages of our framework in the single-person motion generation task.

Another challenge lies in handling human interactions when multi-person interactive actions are included. Human interactions have been explored by some motion prediction algorithms, in which pooling or self-attention modules are adopted to encode the interactions [21, 3, 4, 56]. However, it has not been considered in the motion generation task. To our knowledge, our approach is the first to tackle multi-person motion generation. We share the same latent vector sequence from GP among multiple persons in a group to enforce their synchronization over time. Meanwhile, different persons are distinguished through positional encodings. Our ActFormer can be easily extended to the multi-person scenario by alternately modeling temporal correlations and human interactions. The generation results show impressive realism in both motions and multi-person interactions.

The strong demand for motion capture (MoCap) data with action labels also poses a challenge. Prior methods rely on datasets with $\sim$10 categories, which can hardly drive a general motion generator. MoCap datasets with multi-person interactive actions are even rarer. We leverage NTU RGB+D 120 [37] and the newly-released BABEL dataset [47], both including more than 100 action categories. To facilitate the research on multi-person motion generation, we further construct a GTA Combat dataset through the Grand Theft Auto V's (GTA-V) [1] gaming engine. We collect $\sim$7K motion sequences of combat behavior, which is one of the most complex types of human interactions. Experiments on these datasets verify the effectiveness of our approach.

Our three-fold contributions are summarized as follows: (i) We propose ActFormer, a GAN-based Transformer framework, which adapts to various human motion representations and achieves leading results in the single-person motion generation task; (ii) Our ActFormer takes a faithfully early step to solve the multi-person motion generation problem; (iii) We contribute a GTA Combat dataset with plentiful and complex multi-person interactive motions.

2. Related Work

We review the literature on human motion prediction and generation tasks and MoCap datasets. We also review Transformers in GANs which are relevant to our approach.

2.1. Motion Prediction

Motion prediction aims to predict motions of future frames, given one or several frames of past motions. Recurrent Neural Networks have been predominantly adopted to model sequence learning [16, 26, 41, 22, 11, 55, 8]. Especially, generative models like VAEs and GANs are incorporated in [22, 11, 55]. Recently, the powerful Transformer architecture is utilized by [34] to predict dance motions conditioned on music.

Multi-person interactions have also been considered in motion prediction. [21] and [3] adopts the pooling module to aggregate information across multiple persons. [4] proposes a graph-based message passing mechanism to model both human-human and human-object interactions. Recently, [56] uses the self-attention module to entangle multi-person motions. Unlike the works above, we aim to tackle the motion generation task without relying on past motions.

2.2. Motion Generation

Compared to future motion prediction, motion generation from scratch is a new and less explored field. CSGN [53] generates unconstrained motions with a graph-convolution-based GAN framework. [58] further explores generating ever-changing motions for unbounded durations.

More works tend to generate motions from various conditions. [31, 33, 24] generates dance motions corresponding to the given music. [35, 17, 5, 6, 13, 19, 10, 46, 29] synthesize motions from language descriptions. More recently, diffusion model is also proposed for text-driven human motion generation in [57, 49].

Our work is dedicated to the task of action-conditioned motion generation, which uses semantic action labels as the

condition. Action2Motion [20] and ACTOR [45] are the works most similar to ours. Action2Motion proposes a temporal VAE to generate motions frame by frame, based on GRU architecture. ACTOR also adopts a VAE framework while leveraging the Transformer architecture and learning a sequence-level latent distribution, which differs from Action2Motion. Our ActFormer also receives a sequence-level latent prior as input. Unlike ACTOR, our input is a latent vector sequence sampled from a GP prior, inherently enforcing temporal correlations and thus reducing the difficulty of generating realistic motions.

Another stream of works lies in simulation-based character control. For example, [36] learns an autoregressive conditional VAE and then applies task-specific control policies based on this model. [48] guides characters to achieve specific character-scene interaction goals with their motions. Recently, [52] and [51] seek general approaches for physics-based character control.

2.3. MoCap Dataset

Large-scale MoCap data is critical for driving a general motion generator. There have been many MoCap datasets [25, 40, 44, 7], and AMASS [39] contributes a large collection by unifying MoCap data from different sources into a common representation based on the SMPL body model [38]. Despite a large amount of high-quality and diverse human motion data, semantic action labels are missing in AMASS. Fortunately, the newly-released BABEL [47] provides fine-grained action labels on frame-level for AMASS, resulting in a challenging and practically valuable benchmark for our task.

Some other datasets provide action labels, while their motion data is represented by skeleton joint coordinates. Among them, NTU RGB+D 120 [37] is a large-scale and representative one. Owing to the scalability of our method, we can learn from MoCap data with various motion representations. The NTU RGB+D data is used for both single-person and multi-person motion generation. As a complement to the daily interactive actions in NTU RGB+D, we contribute another GTA Combat dataset with more complex multi-person motions and interactions.

2.4. Transformer in GANs

The success of Transformers [50] in visual recognition tasks inspires its application in generation tasks. TransGAN [27] is the first pure Transformer-based GAN architecture. Later ViTGAN [32] proposes novel regularization methods to improve the stability of Transformer-based GAN training. [14] combines a Transformer-based generator and a CNN-based discriminator to form a robust model without cumbersome design choices. We follow [14] to exempt from the tricky designs in the discriminator and pay more attention to the generator.

3. Approach

The problem to be tackled is action-conditioned motion generation. Formally speaking, given a semantic action label a and a seed z sampled from the latent prior, Our Action-Conditioned Motion TransFormer (ActFormer) will generate a sequence of human motions $M = \{M_t | t \in \{1, ..., T\}\}$ corresponding to the action label. Each frame M_t contains the motions of P persons, *i.e.*, $M_t = \{M_t^p | p \in \{1, ..., P\}\}$. The motion of one person in a frame M_t^p is composed of a root translation l_t^p in a global coordinate frame, and local body poses θ_t^p. The latter can be either SMPL-based parameters or other formats like skeleton joint coordinates. In this section, we first introduce our approach to solve the single-person motion generation and then show how it can be extended to the multi-person setting.

3.1. Single-person Motion Generation

The generation starts with sampling a random seed from the latent prior. Since temporal correlation is critical to a realistic motion sequence, we select the Gaussian Process as our latent prior and sample a (T, C_0) latent vector sequence z for each generation. The time length T is the same as that of the motion sequence to be generated, and the latent vector at each time step has C_0 channels. A 1-d vector sequence with T time steps is sampled independently from a GP on each of C_0 channels, with the characteristic length-scales on different channels spanning a spectrum of values. As suggested by [53], this models a composition of correlations at various time scales into the latent vector sequence.

The ActFormer employs a Transformer-based generator to transform the latent vector sequence and the given action label, to a human motion sequence. As shown in Fig. 2, at the input stage, the (T, C_0) latent vector sequence z is regarded as a list of T tokens and passed through an MLP layer for input embedding. To incorporate the semantic condition, we append a class token embedding the action category a, resulting in totally $T + 1$ tokens. Since the Transformer encoder is data-dependent, we add learnable positional encoding (PE) to the input tokens to maintain location information. After that, L layers of Temporal-transFormer (T-Former) model the temporal correlations among various time steps represented by tokens. Finally, the class token is discarded while each of the rest T tokens is projected by an output layer into a C-d vector. The vector is regarded as a concatenation of a person's root translation and local body poses at a specific time step.

3.2. Multi-person Motion Generation

Transferring from the single-person to multi-person setting induces an additional *person-wise* dimension P. Fortunately, our approach introduced above can fit such an additional dimension P by slight adjustment, thus be scaled to the multi-person case.

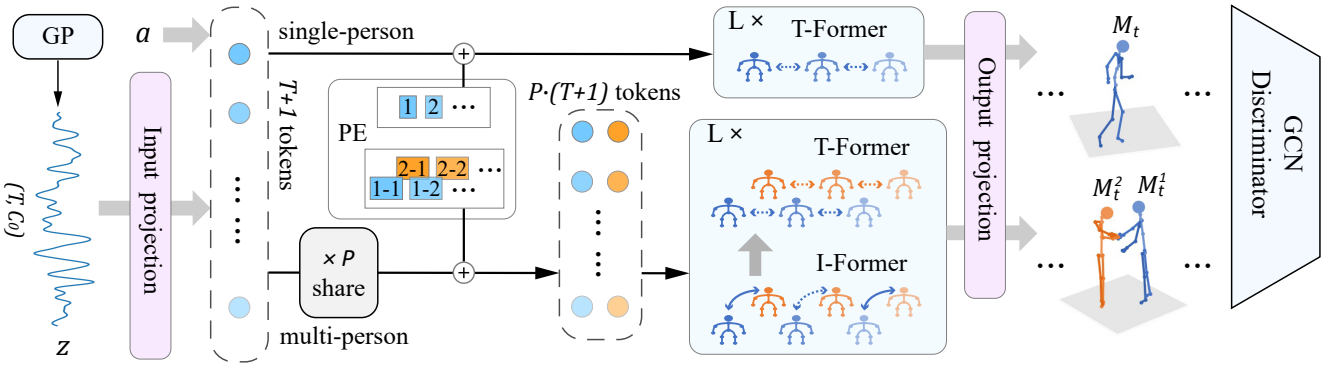

Figure 2. **Overview of the proposed ActFormer framework.** Given a latent vector sequence z sampled from Gaussian Process (GP) prior and an action label a, the model can synthesize either a single-person (top stream) or a multi-person (bottom stream) motion sequence. The model is trained under a GAN scheme.

In the single-person setting, a T-frame motion sequence is encoded from a list of T tokens. Therefore, another T-frame motion sequence with P persons requires P lists of T tokens. Considering that motions of these P persons are highly correlated and synchronized at each time step, we regard them as an entity and sample only one latent vector sequence to share among them. Specifically, the T input embedding tokens along with the class token are shared P times to produce the input. This strategy enforces the synchronization among multiple persons in a group from the input stage. Meanwhile, different persons need to be distinguished from each other. We turn to learnable positional encoding (PE) again to achieve this goal. Here $T+1$ temporal positional encodings (TPE) and P person-wise positional encodings (PPE) are separately learned. Then they are tiled to generate $P \cdot (T+1)$ positional encodings to match input tokens. Specifically, the PE for a (t, p)-indexed token $PE(t, p) = concat(TPE(t), PPE(p))$. Completely independent PE for each token is also feasible, while such a 2D combination provides better results.

The ActFormer generator can be extended to a multi-person setting with slight adjustments, owing to the flexibility of Transformer encoders to adapt to modeling correlations along various dimensions. As Fig. 2 shows, an Interaction-transFormer (I-Former) firstly models the interactions among different persons at each time step independently. Then a Temporal-transFormer (T-Former) follows to model the temporal correlations for each person independently. Such a module, which alternately encodes human interactions and temporal correlations, is stacked L layers. Similar to the single-person case, all class tokens are discarded finally. The remaining $P \cdot T$ tokens are projected into the final (P, T, C) output, in which the C-d vector from each token represents the motion of a specific person at a particular time step.

3.3. Generative Adversarial Training

Our ActFormer is learned under the conditional generative adversarial training framework [18, 42]. During training, the ActFormer generator synthesizes human motion sequences conditioned on given action labels. Besides, a discriminator receives human motion sequences and action labels as inputs, trying to discriminate the generated human motions from real ones of specified actions. The generator learns from the discriminator's feedback to improve its generation results to be close to real ones. Conditional Wasserstein GAN loss functions [42, 9] are adopted for training, formatted as,

$$\begin{aligned} L_D &= \mathbb{E}[D(G(z,a),a) - D(\tilde{M},a)], \\ L_G &= \mathbb{E}[-D(G(z,a),a)], \end{aligned} \tag{1}$$

where $\tilde{M}$ represents the sequence sampled from real motion data belonging to the action category a. D denotes the discriminator and G denotes the ActFormer generator.

An ST-GCN [54] is adopted as the discriminator. The C-d vector of a person's motion at each time step is decomposed into a (K, D) part-wise representation. The K-node Graph is constructed according to the skeleton topology behind the local body poses θ_t^p, with the root translation l_t^p also connected. Therefore, a (T, C) motion sequence is re-organized into a (T, K, D) spatial-temporal graph. In the multi-person setting, the part-wise motions of multiple persons are directly concatenated on the D dimension. In other words, the (P, T, C) multi-person motion sequence becomes a $(T, K, P \cdot D)$ graph. Finally, the graph-based motion sequence is input to the GCN discriminator to compute a score, with the semantic condition integrated by projection [43].

By concatenating the part-wise motions of multiple persons, the GCN can model their interactions at various spatial and temporal scales. However, the concatenation operator is not permutation-invariant. In other words, it cannot ensure to output the same score for the same motion sequence with various person-wise permutations. We adopt a simple data augmentation strategy to compensate for this. For each sample from the MoCap dataset, the persons inside are randomly permuted in every training iteration. In

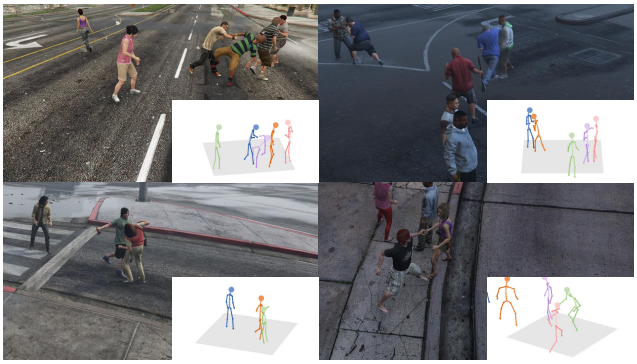

Figure 3. Sample RGB images and human pose annotations in the GTA Combat dataset.

this way, we encourage the ActFormer to regard the same sample with various permutations as different samples and model all of them into the learned distribution. We find such a simple data augmentation more robust than importing symmetric functions into the discriminator network, since the latter often destabilizes the GAN training.

4. Experiments

In this section, we evaluate the proposed ActFormer on both single-person and multi-person motion generation tasks. We firstly introduce the datasets and quantitative metrics used for evaluation. Next, the ActFormer is compared with baseline methods from prior works. Then we conduct an ablation study to investigate various components in ActFormer. Finally, qualitative results are shown.

4.1. Datasets and Evaluation Metrics

NTU-13 [37, 20] is a subset of NTU RGB+D 120 with only 13 action categories and we choose it to be able to compare to previous works [20, 45]. The SMPL-based motion data are obtained through VIBE [30], as adopted in [20, 45].

NTU RGB+D 120 [37] contains 114,480 motion clips belonging to 120 action categories. Among them, 94 categories are single-person actions, and the rest 26 categories are two-person interactive actions. The dataset provides MoCap data in the format of skeleton joint coordinates, which are captured by Kinect [59] and severely noisy. We fill missing detections and perform temporal smoothing to improve its quality. Different parts of the dataset are used for both single-person and multi-person motion generation (referred to as NTU-1P and NTU-2P respectively in the following context). Moreover, to evaluate the methods' adaptability to different motion representations, we apply a motion reconstruction pipeline similar to [34] onto multi-view images from NTU-1P. This results in a new version of NTU-1P data with SMPL parameters, named NTURecon-1P.

BABEL [47] is another large-scale motion dataset with semantic action labels. The SMPL-based motion data is derived from AMASS with higher quality. The whole dataset contains more than 250 action categories but is remarkably long-tailed. Therefore, we follow the BABEL-120 benchmark in [47] and select motion data belonging to the most frequent 120 action categories. The data is used for single-person motion generation.

GTA Combat is a synthetic multi-person MoCap dataset collected by us. As there are few full-body MoCap datasets qualified for multi-person interactive action generation, we tackle the game GTA-V [1] for synthetic data collection. We find combat behavior with 2 or more persons is a partial procedure generation process in GTA-V. The combat actions are randomly triggered by combing more than 10 atomic actions in real-time with more variations when equipped with different combating arms. And the attacked one can react randomly, triggered by the ragdoll physics [2] of the game engine. Our policy of combating behavior is that each character pick a random opponent to attack without order. This setting is more like the chaos fighting scene in some fighting games. So it simulates parts of real-world human social interactions but the complexity and diversity are beyond real-world fightings, which is suitable to serve as a benchmark to confirm that our framework can work well on more complex interactive actions (more than 2 persons). We sample 2 to 5 actors in each run and make each actor fight with one of the others randomly picked. We extend JTA [15] which helps to extract 3D human skeletons with this random combat logic for the data collection. Thereby, we get a high-quality multi-person interaction MoCap dataset ranging from 2 to even 5 persons. Fig. 3 shows some samples of GTA Combat dataset. More samples are provided in the supplementary. According to the number of persons in each sequence, the dataset is divided into 4 splits. Each split contains $\sim$2.3/1.9/1.5/1.2K sequences with motions of 2/3/4/5 persons.

Evaluation Metrics. We use action recognition accuracy and FID [23] score as quantitative metrics. Note that these metrics require a pre-trained action recognition model. We adopt ST-GCN [54], sharing a similar configuration to the discriminator in the ActFormer's training. Different from previous works [20, 45], the **root translation is considered** as an extra node when training the action recognition model since the root translation results are significant to measure the quality of the generated motions especially for the multi-person scenarios.

Previous works [20, 45, 53] consider statistics of the whole real/generated sample set when computing FID, regardless of action categories. We denote this as whole FID (FID_w), while we find this not reasonable for the action-conditional generation case. As a supplement, we adopt another mean FID (FID_m), in which FID is independently measured on samples of each action category and then averaged. FID_m assumes the conditional distributions of per-

Method	NTU-13			NTURecon-1P			NTU-1P			BABEL		
	Acc.↑	FID_m↓	FID_w↓	Acc.↑	FID_m↓	FID_w↓	Acc.↑	FID_m↓	FID_w↓	Acc.↑	FID_m↓	FID_w↓
Action2Motion [20]	94.9	4.40	2.01	41.97	23.86	18.35	8.34	56.13	46.13	6.04	17.03	6.05
CSGN [53]	85.9	8.07	3.64	20.02	36.02	27.51	38.99	17.01	7.14	7.02	20.11	9.17
ACTOR [45]	97.1	5.35	1.18	39.69	37.87	19.58	35.56	32.89	9.63	3.44	55.66	36.27
Ours	**99.9**	**4.28**	**1.11**	**49.12**	**20.35**	**8.86**	**42.01**	**12.53**	**3.57**	**13.49**	**11.21**	**2.55**

Table 1. **State-of-the-art comparison on single-person motion generation.** Higher Acc. and lower FIDs are better.

Method	Acc.↑	FID_m↓	FID_w↓	FID^a_m↓	FID^a_w↓
Action2Motion [20]	16.85	15.07	10.12	26.25	20.31
CSGN [53]	55.12	6.10	3.88	8.96	3.20
ACTOR [45]	63.04	13.56	6.72	19.14	8.16
Ours	**69.65**	**3.77**	**3.27**	**7.12**	**2.46**

Split	CSGN [53]	Ours
2p	1.15	**1.04**
3p	1.73	**1.13**
4p	2.09	**1.58**
5p	4.02	**2.23**

Table 2. **State-of-the-art comparison on multi-person motion generation. Left:** Performance on NTU-2P dataset. **Right:** FID_a performance on different splits of the GTA Combat dataset.

class samples to be individual Gaussians, thus better describing the similarity between two *mixture* distributions in this task. We adopt FID_m and FID_w for all the experiments.

When extending to the multi-person setting, we measure the FID in two ways to observe generation results from different perspectives. The first is to regard a group of persons as a whole. Specifically, the recognition model receives concatenated multi-person motions as input and extracts features for calculating FID directly from it. Data augmentation by random person-wise permutation is also applied in the recognition model's training to compensate for permutation invariance. The second is to construct multi-person features by aggregating single-person features. Towards this end, we train a single-person action recognition model. During the evaluation, we extract features of each person in a group independently and aggregate them by channel-wise max pooling. Such feature extraction is applied on both real and generated samples for alignment. FID calculated in this way is called *Aggregated FID* (FID^a).

Implementation Details. For each dataset, we use different data splits to train ActFormer and action recognition models. In NTU RGB+D 120, we follow its cross-subject split. In BABEL, we also follow the provided split as [47]. The data distribution of different action categories in BABEL is remarkably long-tailed. Therefore, we adopt a square-root sampling [28] strategy in training the ActFormer, the action recognition models, and all the compared baselines. GTA Combat contains only one action category, making it infeasible to train an action recognition model on it. Also, to our knowledge, there are no public datasets whose motions contain 2∼5 participants and meanwhile with action categories annotated. Therefore, we leverage an action recognition model trained on NTU-1P by aligning motion data in NTU and GTA Combat into a unified skeleton topology. In this way, FID^a can be measured on GTA Combat. Please refer to the supplementary for more details about our network architecture and training/evaluation configurations.

4.2. Comparison to State-of-the-Arts

We compare the proposed ActFormer with the following baseline methods: Action2Motion [20], ACTOR [45], and CSGN [53], on both single- and multi-person motion generation tasks. Action2Motion and ACTOR can be directly evaluated on the specified datasets by simply adapting motion representations if needed. CSGN is an unconditional motion generation method, and we extend it to conditional generation by incorporating conditional BatchNorm [12] in generator and projection [43] in discriminator. All the methods above are for the single-person setting, and there is no prior works tackling the multi-person case. Therefore, we scale these methods to the multi-person case, again by concatenating multi-person motions as a whole in each frame.

The comparison on single-person motion generation is presented in Tab. 1. Action2Motion is sub-optimal in both latent prior (frame-level) and network architecture (GRU), causing it to lag behind ours on all datasets, especially for the NTU-1P dataset. CSGN achieves good performance on NTU-1P. However, this method, oriented for skeleton data, suffers a performance degradation when moving from NTU-1P to SMPL-based NTURecon-1P. ACTOR achieves excellent performance on NTU-13, but degrades significantly when faced with more challenging large-scale datasets and more strict evaluation metrics (considering root translation). Compared to them, our method demonstrates strong adaptability to various motion representations and achieves leading performance on all datasets. In particular, our ActFormer is the only workable method on the extremely challenging, long-tailed BABEL dataset.

As Tab. 2 shows, when extended to the multi-person case on NTU-2P, our advantages are more significant since no prior methods take specific designs to model human interactions. On GTA Combat, multi-person motions are fur-

Configuration	Acc.↑	FID_m↓	FID_w↓
(1) Gaussian latent prior	37.55	17.25	6.01
(2) GraphConv. in Gen.	32.97	18.68	6.09
(3) Fixed PE	37.57	18.38	5.55
(4) Full model	**42.01**	**12.53**	**3.57**

Table 3. **Ablation study:** Several design choices on NTU-1P.

Configuration	Acc.↑	FID_m↓	FID_w↓	FID^a_m↓	FID^a_w↓
(5) AvgPool. in Disc.	53.85	20.42	8.33	25.35	10.09
(6) MaxPool. in Disc.	60.19	10.71	4.01	17.26	5.42
(7) SelfAtt. in Disc.	51.53	19.96	8.27	24.91	8.75
(8) Fixed PE	61.42	4.86	3.64	7.71	2.78
(9) Independent PE	67.19	4.08	3.38	**7.00**	2.62
(10) Full model	**69.65**	**3.77**	**3.27**	7.12	**2.46**

Table 4. **Ablation study:** Several design choices on NTU-2P.

ther extended to at most 5 persons, making the generation task extremely challenging. According to the experimental results on NTU-2P, CSGN is the only baseline method showing some potential (lower FIDs) to be applied on the multi-person case. Therefore, here we compare our method with only CSGN on the GTA Combat dataset. With the increasing number of persons, both methods are prone to performance drops. By contrast, our ActFormer shows a more smooth decline and outperforms CSGN on each dataset split. Despite this, the interaction complexity in 4P/5P poses a significant challenge to our method. Specific failure cases will be discussed in Sec. 4.4.

4.3. Ablation Study

In this part, we conduct ablation experiments to study various components in our proposed framework.

Latent Prior. Here we further verify the importance of Gaussian Process latent prior in our framework. As in Tab. 3 (1), replacing GP with a Gaussian latent prior leads to 4.72 FID_m and 2.44 FID_w increases on NTU-1P.

Architecture Design. In the ActFormer, frame-wise human motions are encoded into vector-like token embeddings. In Tab. 3 (2), we evaluate another design choice: explicitly modeling skeleton topology with Graph Convolution and Graph Upsampling like in [53], meanwhile modeling temporal correlations with T-Former. It lags behind ActFormer, suggesting no need to explicitly model skeleton topology in this Transformer-based framework.

Positional Encoding. In the data-dependent Transformer architecture, positional encoding (PE) is a crucial component which brings additional positional dependencies. We find it better to make PE learnable rather than fixed in our method, given the results in Tab. 3 (3) and Tab. 4 (8). Moreover, for multi-person case, learning a 2D combination of temporal and person-wise PE is superior to Tab. 4 (9) of completely learnable independent PE, showing another tradeoff between inductive bias and representation capacity.

Discriminator. In the multi-person case, the GCN discriminator in the GAN training framework receives concatenated multi-person motions as input, equipped with a simple data augmentation strategy for permutation invariance. Here we investigate its effectiveness by comparing with other designs which embed permutation invariance into the discriminator architecture with symmetric functions. Specifically, these designs apply the same motion discriminator to each person independently and use the pooling/self-attention module to aggregate their features. Tab. 4 (5-7) shows that all of them fall behind our simple combination of concatenation with data augmentation.

Interaction Encoding. The leading performance of our method in multi-person motion generation is credited to two key designs. The first is to share the same latent vector sequence among multiple persons for inherent synchronization. The other is to model human interactions with I-Former. To verify their respective contributions, we experiment in Tab. 5 (1) to sample different latent vector sequences for different persons independently, and Tab. 5 (2) to adopt ActFormer for the single-person setting and output concatenated multi-person motions as a whole, without explicitly modeling interactions inside. From experiments on NTU-2P as shown in Tab. 5, we find both designs indispensable to synthesize high-quality interactive actions.

We further investigate the interaction encoding by experiments on GTA Combat. Despite the performance drop with the increasing number of persons, the complete model always maintains a leading position. When the number of persons reaches to 5, the ablation Tab. 5 (2) collapses with a 7.62 FID_a. We attribute this to the fact that human interactions become sparse correlations at this moment. Compared to position-dependent operation in concatenation-based interaction module, encoding human interactions with data-dependent self-attention in I-Former can better model such sparse correlations.

4.4. Qualitative results

We further evaluate the generation quality of ActFormer by visualization. Firstly, we visualize several generated samples for single-person actions from BABEL in Fig. 4. For each action label, diverse motions are synthesized. In the "Stretch" case, the persons are stretching different body parts in different samples. In the "Dance" case, various dancing styles are presented.

In Fig. 5 we visualize multi-person motion generation results for actions from NTU-2P and GTA Combat. In the multi-person case, we focus more on the synchronization of human interactions. In our generated samples, the motions of different participants are well synchronized, making the interactions look natural and vivid. For example, in the "Cheers and drink" case, the two persons' toasting, cheering, and drinking actions are temporally well-matched. In

Configuration	NTU-2P Acc.↑	NTU-2P FID_m↓	NTU-2P FID_w↓	NTU-2P FID^a_m↓	NTU-2P FID^a_w↓	GTA Combat (2∼5 P) FID^a↓			
(1) w/o shared z	56.35	14.79	5.73	20.69	7.26	1.26	1.36	1.84	2.42
(2) Interaction by concatenation	56.46	7.63	4.33	11.51	3.57	1.15	1.37	2.23	7.62
(3) Complete model	**69.65**	**3.77**	**3.27**	**7.12**	**2.46**	**1.04**	**1.13**	**1.58**	**2.23**

Table 5. **Ablation study:** Interaction encoding on NTU-2P and GTA Combat.

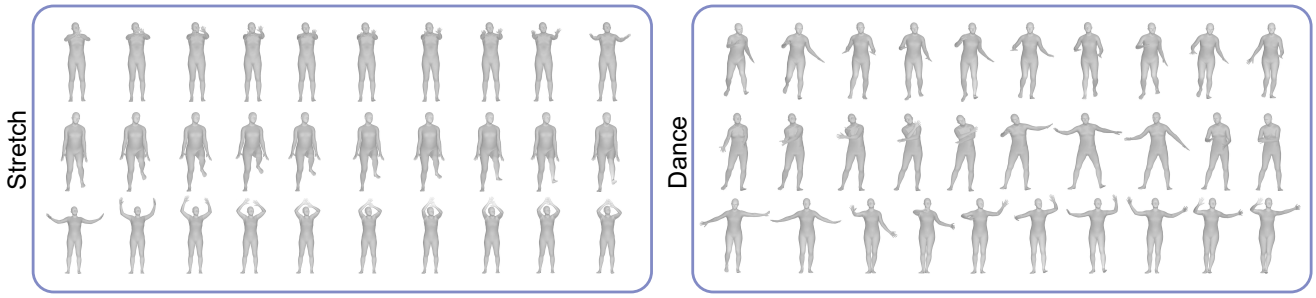

Figure 4. **Generated single-person motions.** The "Stretch" and "Dance" actions are both from BABEL.

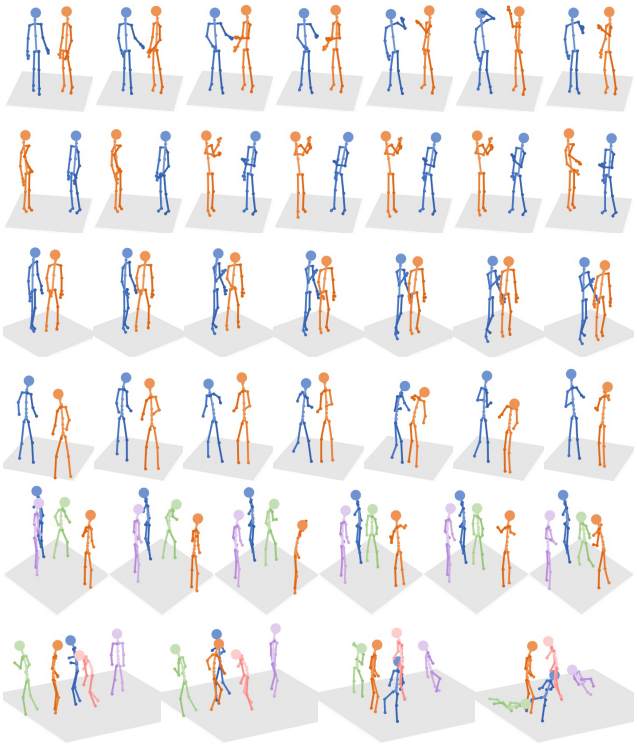

Figure 5. **Generated multi-person motions. Row 1∼3:** "Cheers and drink", "Take a photo", and "Support somebody" actions from NTU-2P. **Row 4∼6:** "Combat" actions from GTA Combat.

the "Take a photo" case, after the photographer on the left has prepared the camera, the person on the right poses up immediately (gesturing with one hand while slightly leaning back) and maintains the posture until the end of photographing. Synchronization is reflected in not only local poses but also global trajectories. As seen in the "Support somebody" case, the two persons are stumbling together. Their trajectories keep close since the sick one relies on the other to walk. In the first "Combat" case, we see the blue fighter suddenly attack the opponent and make him stagger backward. The second "Combat" case is more complex, in which four persons fight in pairs. The right two clash, while the left two, although not contacting, maintain fight postures and constantly move to find chances. More qualitative **video results** can be found in the supplementary.

The last sample in Fig. 5 shows a failure case. The green and purple persons seem to interact with nobody and then fall onto the ground without being attacked. This case reflects a limitation of our method: when the number of persons increases, human interactions become sparse and may form two separate groups. ActFormer has no mechanism to divide persons into different interaction groups, thus not sufficiently effective to learn from some GTA Combat data with 4 or 5 persons.

5. Conclusion and Discussion

This work explores a solution towards general action-conditioned human motion generation and proposes ActFormer, a GAN-based Transformer framework. The ActFormer is evaluated on several challenging benchmarks and achieves leading performance over prior methods on various human motion representations and both single-person and multi-person motion generation tasks. Detailed ablation studies are also conducted to investigate the various components in our approach. The ActFormer adapts to a more complete domain of human actions compared to prior works, while the general human motion generator is still not reached. Human-object interaction synthesis remains unex-

plored, and we leave this direction for future exploration.

Limitations. As discussed in the qualitative results section of the main manuscript, our method has no mechanism to divide persons into different interaction groups. Therefore, given MoCap samples in which multiple persons form separate interaction groups, the ActFormer cannot effectively learn from it. Besides, in the GAN training, multi-person motions are concatenated before input to the GCN discriminator. Thus we cannot learn a shared model for motions with a variable number of persons, despite the ActFormer being a Transformer-based generator.

Broader impacts. The proposed generative method can synthesize non-existing content. The community should be wary of the malicious uses of this feature. The collected GTA Combat dataset contains fighting scenes, while we do not promote violence. The dataset should only be used for research on modeling multi-person interactive actions.

Acknowledgement. This work is supported by ZJNSFC under Grant LQ23F010008, NSFC (62201342, 62101325, U19B2035), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). The authors would like to appreciate the High Performance Computing Center at Eastern Institute of Technology, Ningbo for the GPU support.

References

[1] Grand Theft Auto V. `https://www.rockstargames.com/V/`.

[2] Ragdoll Physics. `https://gta.fandom.com/wiki/Ragdoll_Physics`.

[3] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics Autom. Lett.*, 5(4):6033–6040, 2020.

[4] Vida Adeli, Mahsa Ehsanpour, Ian D. Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. *CoRR*, abs/2104.04029, 2021.

[5] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *ICRA*, pages 1–5. IEEE, 2018.

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

[7] Ijaz Akhter and Michael J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *CVPR*, pages 1446–1455. IEEE Computer Society, 2015.

[8] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *ICCV*, pages 7143–7152. IEEE, 2019.

[9] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[10] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. *arXiv preprint arXiv:2209.04066*, 2022.

[11] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *CVPR Workshops*, pages 1418–1427. Computer Vision Foundation / IEEE Computer Society, 2018.

[12] Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *NIPS*, pages 6594–6604, 2017.

[13] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV*, pages 346–362. Springer, 2022.

[14] Ricard Durall, Stanislav Frolov, Andreas Dengel, and Janis Keuper. Combining transformer generators with convolutional discriminators. *CoRR*, abs/2105.10189, 2021.

[15] 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*, 2018.

[16] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354. IEEE Computer Society, 2015.

[17] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, pages 1396–1406, 2021.

[18] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[19] 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.

[20] 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.

[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. Computer Vision Foundation / IEEE Computer Society, 2018.

[22] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *BMVC*. BMVA Press, 2017.

[23] 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.

[24] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *ICLR*. OpenReview.net, 2021.

[25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2014.

[26] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, pages 5308–5317. IEEE Computer Society, 2016.

[27] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong GAN. *CoRR*, abs/2102.07074, 2021.

[28] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*. OpenReview.net, 2020.

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

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *CVPR*, 2020.

[31] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *NeurIPS*, pages 3581–3591, 2019.

[32] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *CoRR*, abs/2107.04589, 2021.

[33] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *CoRR*, abs/2008.08171, 2020.

[34] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with AIST++: music conditioned 3d dance generation. *CoRR*, abs/2101.08779, 2021.

[35] Angela S. Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J. Mooney. Generating animated videos of human activities from natural language descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS 2018*, December 2018.

[36] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel van de Panne. Character controllers using motion vaes. *ACM Trans. Graph.*, 39(4):40, 2020.

[37] 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 Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, 2020.

[38] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015.

[39] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *ICCV*, pages 5441–5450. IEEE, 2019.

[40] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT whole-body human motion database. In *ICAR*, pages 329–336. IEEE, 2015.

[41] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, pages 4674–4683. IEEE Computer Society, 2017.

[42] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[43] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *ICLR (Poster)*. OpenReview.net, 2018.

[44] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. Efficient and robust annotation of motion capture data. In *Symposium on Computer Animation*, pages 17–26. ACM, 2009.

[45] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, 2021.

[46] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497. Springer, 2022.

[47] 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. Computer Vision Foundation / IEEE, 2021.

[48] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209:1–209:14, 2019.

[49] 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.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[51] Tingwu Wang, Yunrong Guo, Maria Shugrina, and Sanja Fidler. Unicon: Universal neural controller for physics-based character motion. *CoRR*, abs/2011.15119, 2020.

[52] Jungdam Won, Deepak Gopinath, and Jessica K. Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Trans. Graph.*, 39(4):33, 2020.

[53] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *ICCV*, pages 4393–4401. IEEE, 2019.

[54] 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.

[55] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. MT-VAE: learning motion transformations to generate multimodal human dynamics. In *ECCV (5)*, volume 11209 of *Lecture Notes in Computer Science*, pages 276–293. Springer, 2018.

[56] Mohammad Samin Yasar and Tariq Iqbal. A scalable approach to predict multi-agent motion for human-robot collaboration. *IEEE Robotics Autom. Lett.*, 6(2):1686–1693, 2021.

[57] 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.

[58] Yan Zhang, Michael J. Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *CoRR*, abs/2007.13886, 2020.

[59] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multim.*, 19(2):4–10, 2012.