Role-aware Interaction Generation from Textual Description

Mikihiro Tanaka[1], KentFujiwara[2]
LINE Corporation
[1]mikihiro.tanaka@linecorp.com, [2]kent.fujiwara@linecorp.com

Abstract

This research tackles the problem of generating interaction between two human actors corresponding to textual description. We claim that certain interactions, which we call asymmetric interactions, involve a relationship between an actor and a receiver, whose motions significantly differ depending on the assigned role. However, existing studies of interaction generation attempt to learn the correspondence between a single label and the motions of both actors combined, overlooking differences in individual roles. We consider a novel problem of role-aware interaction generation, where roles can be designated before generation. We translate the text of the asymmetric interactions into active and passive voice to ensure the textual context is consistent with each role. We propose a model that learns to generate motions of the designated role, which together form a mutually consistent interaction. As the model treats individual motions separately, it can be pretrained to derive knowledge from single-person motion data for more accurate interactions. Moreover, we introduce a method inspired by Permutation Invariant Training (PIT) that can automatically learn which of the two actions corresponds to an actor or a receiver without additional annotation. We further present cases where existing evaluation metrics fail to accurately assess the quality of generated interactions, and propose a novel metric, Mutual Consistency, to address such shortcomings. Experimental results demonstrate the efficacy of our method, as well as the necessity of the proposed metric. Our code is available at `https://github.com/line/Human-Interaction-Generation`.

1. Introduction

Modeling human motion is becoming an important element for creating high-quality 3D animation, with the rising demand in applications such as animating game characters and 3D online avatars. As human motion is complex and has a wide range of variations, animation of human characters has mostly been either hand-crafted or produced based on human actors through means such as 3D motion capture systems, both of which are generally very expensive.

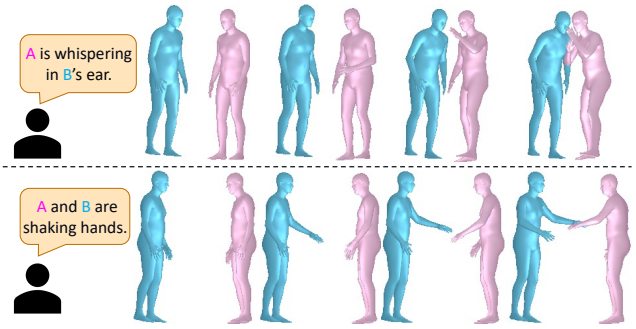

Figure 1. Examples of role-aware interactions generated by the proposed method, with language as input and colored roles assigned to corresponding human actors. Top: Asymmetric interaction, in which there is a speaker and a listener. Bottom: Symmetric interaction, in which both share the same action of shaking hands.

In recent years, generating single human motion from language has made significant progress [37, 9, 51, 59]. Especially with the recent development of diffusion models [16, 47, 20], it has become possible to generate more faithful and diverse motions of a single character from language. When we consider multiple characters, more complex modeling is required compared to animating a single character, as their relative positions and interactions need to be taken into account, in addition to modeling individual motions. Some attempts have been made to combine two human motions and learn the correspondence between a single label and the pair of actions [29, 11]. However, viewing from the perspective of each actor, this approach is particularly problematic in interactions where there is an initiator and a receiver of an action, as there is a contradiction between the action label, which is generally in the active voice, and the receiver's motion, which should be passive.

In this paper, we propose a role-aware interaction generation model that can designate individual roles while achieving consistency between two human motions by extending a single motion generation method based on the diffusion model [59]. We base our proposal on the insight that among interactions between two humans, there are asymmetric interactions, in which there is an actor and a receiver,

and symmetric interactions, where both human actors perform a common action, as shown in Fig. 1. In the asymmetric example, the sentence describing the interaction between two humans encompasses a relationship where one motion should be described as "whispering" in the active voice, and the other as "being whispered to" in the passive voice.

We assign appropriate textual description for each role in the interaction by providing passive and active voice languages to the corresponding motions when the interaction is asymmetric, and the same language description to both motions when the interaction is symmetric. To capture the characteristics of the different roles, we propose a model based on two Transformer units with shared parameters. We introduce a cross attention module between the two Transformers, to take into account the relationship between two human positions and motions, and to capture temporal correspondence between motions of the two roles. As the proposal handles individual motions separately, we further show that the Transformer units can be pretrained on a single-person motion data, which is a powerful prior knowledge for generating individual motions faithful to texts.

To train our model, we need to indicate which of the two humans is an actor or a receiver. To avoid introducing additional annotation costs, we propose a method that can automatically learn to differentiate the roles by adopting Permutation Invariant Training (PIT) [22], which is a commonly used technique in multi-talker speech separation tasks. We train a label-based version of our interaction generation model to separate the two roles and assign pseudo-labels for each role. Which of the pseudo-labels corresponds to an active or a passive role can be determined by simple inspection of the generated results. The obtained roles are assigned to the training data to generate interactions that conform to the textual description of each role.

We evaluate our method on the interaction subset of the NTU-RGB+D 120 [43] dataset. As the conventional evaluation metrics were insufficient to identify the flaws and inconsistency in the generated interactions, we additionally propose a novel metric, Mutual Consistency, to further assess the accuracy of interactions. The experimental results, along with the proposed metric, demonstrate that the proposed method is able to effectively generate realistic interactions between two human characters.

Overall, our contributions are as follows:

- We propose a novel role-aware interaction generation model that can assign individual roles while maintaining consistency between two human motions. For the asymmetric interactions, we translate the text into active and passive voice descriptions for the model input.
- We propose a method inspired by PIT, which automatically learns to separate interactions into motions of an actor and a receiver to reduce annotation costs.
- Experimental results, and further evaluation with our metric Mutual Consistency, show that our method can generate interactions in which two human motions are more mutually consistent and faithful to the input text than existing studies that do not consider roles.

2. Related Work

2.1. Motion Recognition

Human motion analysis has so far been mainly conducted using image-based videos [46, 7]. However, with the recent development of motion acquisition methods [14, 12, 13], skeletal motion data is attracting significant interest. Motion recognition for skeletal data is one of the topics under active research, as it allows us to focus on the motion itself, eliminating the influence of background [54, 6]. Some approaches convert the coordinates of skeletal joints into vector data [27] or two-dimensional grid data [19]. More recently, methods have been proposed to treat the connectivity of joints as a graph to be used as input for deep learning models [55, 45]. Furthermore, studies have been conducted to identify not only what actions are in the motion sequence, but also when in the sequence they take place [56].

While most research has focused on single human motion, human interaction research has also been conducted, similarly starting with videos [52, 60, 23, 41]. Since it is important to understand the intrinsic properties of each human body, as well as the relationships between multiple bodies in recognition of interactions, skeletal data has lately captured widespread attention [57, 18, 30, 35, 17, 33]. To promote research in the field, Yun *et al*. [57] created a relatively small scale skeletal interaction dataset. Recently, Shahroudy, *et al*. [43] created a large scale skeletal motion dataset, NTU-RGB+D. NTU-RGB+D contains both single person motions and interactions. Liu *et al*. [26] increased the number of data and classes and created NTU-RGB+D 120. Because of the scale, NTU-RGB+D and NTU-RGB+D 120 have been used as a general benchmark for interactions.

2.2. Motion Generation

Along with the aforementioned research on analysis of human motion based on skeletal data, there has also been growing interest in research aiming to generate new motion based on knowledge obtained from skeletal data analysis.

Traditionally, motion generation has been conducted by connecting the most consistent motion from various patterns [1, 3]. With the development of deep learning, learning from data is currently the mainstream option for motion generation. Various approaches have been proposed to learn to generate motion from data, including methods for generating appropriate actions given action labels [10, 36, 5], research on generating actions that match music [24, 25], and methods that focus on the periodicity of actions [48].

For generating random motion from input conditions, many works have relied on conditional Variational Autoencoders (VAE) [10, 36]. Petrovich *et al*. [36] used action classes as conditions to learn a latent space, from which samples of each action can be generated. The approach has been extended by conditioning the VAE to learn a joint embedding space of motion and language [37, 2]. Some studies generate motions autoregressively by learning a latent representation compressed by an Autoencoder [9, 58].

With the advent of Contrastive Language-Image Pre-Training (CLIP) [40], language has gained considerable prominence as an intuitive interface in the research of motion generation. MotionCLIP [50] maps the CLIP cross-modal feature representation between image and language, to the motion feature space, and generates motions that match the language. In addition, with the recent development of diffusion models and large scale text-motion paired datasets [38, 39, 9], it has become possible to generate diverse motions that are faithful to input descriptions [51, 59].

Some attempts have been made to generate interactions between human characters. There are methods that generate reactions to existing single-person actions [8, 31]. These require the motion of the active side, limiting the variety of interactions. There is also a study on generating facial motion of listeners that is appropriate as a response to the voice and motion of speakers [32]. Some attempt to generate pairs of human motions that constitute interactions [29, 11]. One concurrent research also try to generate interactions using human motion diffusion model as a prior [42]. However, these methods attempt to learn the correspondence between a single label and motions of both the actor and the receiver combined, overlooking the difference of individual roles. This limits the flexibility, as the roles of individual actors cannot be assigned during generation. We propose a method for generating role-aware interactions that learns the correct correspondence between texts and individual motions.

3. Method

In this study, we propose a role-aware interaction generation model that satisfies the following properties.

- The model can generate interactions in which individual roles can be assigned to the humans involved for both symmetric and asymmetric interactions.
- Generated interactions of two humans are mutually consistent, *i.e*. generated motions achieves temporal and positional consistency with respect to each other.

We assume that for an interaction i, a description y^i of an interaction category c^i is given. In the asymmetric interaction category, where there is an actor and a receiver, each motion in an interaction corresponds to one of the descriptions (y^i_{active} , $y^i_{passive}$), which are translated from y^i into active and passive voices respectively.

To avoid manual annotation of (y^i_{active} , $y^i_{passive}$), we propose a method inspired by PIT [22], which is able to separate interactions into motions of an actor and a receiver when only the category c^i is available. We assign corresponding descriptions (y^i_{active} , $y^i_{passive}$) to each motion and use the data to train our role-aware model to generate consistent interactions from text.

We firstly introduce a state-of-the-art single motion generation method using diffusion models, which we base our interaction generation model on. Then, we explain the details of our interaction generation model when (y^i_{active} , $y^i_{passive}$) is available. Finally, we introduce a method that can automatically separate interactions into an actor and a receiver to assign (y^i_{active} , $y^i_{passive}$) when only c^i is given.

3.1. Diffusion-based Single Motion Generation

First, we describe a method that uses one of the state-of-the-art diffusion models for motion generation from language [59], which we base our interaction generation model on. Motion generation can be formulated as an inverse diffusion process from randomly sampled noise. The motion series $X^{(0)} = [x_1, ..., x_F]$, where F is the number of frames, and $[x_1, ..., x_F]$ are the pose representations at each time, is sampled from the real data $X^{(0)} \sim q(X^{(0)})$, and noise is added at T steps to obtain $X^{(1)}, ..., X^{(T)}$. The diffusion process can be expressed by adding Gaussian noise according to the Markov chain as

$$\begin{aligned} q(X^{(1:T)}|X^{(0)}) &= \prod_{t=1}^{t=T} q(X^{(t)}|X^{(t-1)}) \\ q(X^{(t)}|X^{(t-1)}) &= \mathcal{N}(X^{(t)}; \sqrt{1-\beta_t}X^{(t-1)}, \beta_t I)\,, \end{aligned} \tag{1}$$

where $X^{(T)}$ approximately follows $\mathcal{N}(0, I)$. β_t is a hyperparameter that determines the variance schedule.

The inverse diffusion process starts from random noise and then removes the noise according to the Markov chain by adding Gaussian noise according to the estimated values of $\mu_\theta(X^{(t)}, t)$ and $\Sigma_\theta(X^{(t)}, t)$

$$\begin{aligned} p_\theta(X^{(0:T)}) &= p_\theta(X^{(T)}) \prod_{t=1}^{t=T} p_\theta(X^{(t-1)}|X_t) \\ p_\theta(X^{(t-1)}|X^{(t)}) &= \mathcal{N}(X^{(t-1)}; \mu_\theta(X^{(t)}, t), \Sigma_\theta(X^{(t)}, t))\,, \end{aligned} \tag{2}$$

in order to generate the motion. Motiondiffuse [59] uses a constant value for $\Sigma_\theta(X^{(t)}, t)$ and only estimates $\mu_\theta(X^{(t)}, t)$ with a noise estimator model ϵ_θ based on the Transformer [53]. Since $\mu_\theta(X^{(t)}, t)$ can be calculated from the model output $\epsilon_\theta(X^{(t)}, t, text)$, the model parameters are optimized by minimizing the loss between randomly sampled noise ϵ and the estimated noise at step number t:

$$\mathcal{L} = E_{t\in[1,T], X^{(0)}\sim q(X^{(0)}), \epsilon\sim\mathcal{N}(0,I)}[\|\epsilon - \epsilon_\theta(X^{(t)}, t, text)\|]\,. \tag{3}$$

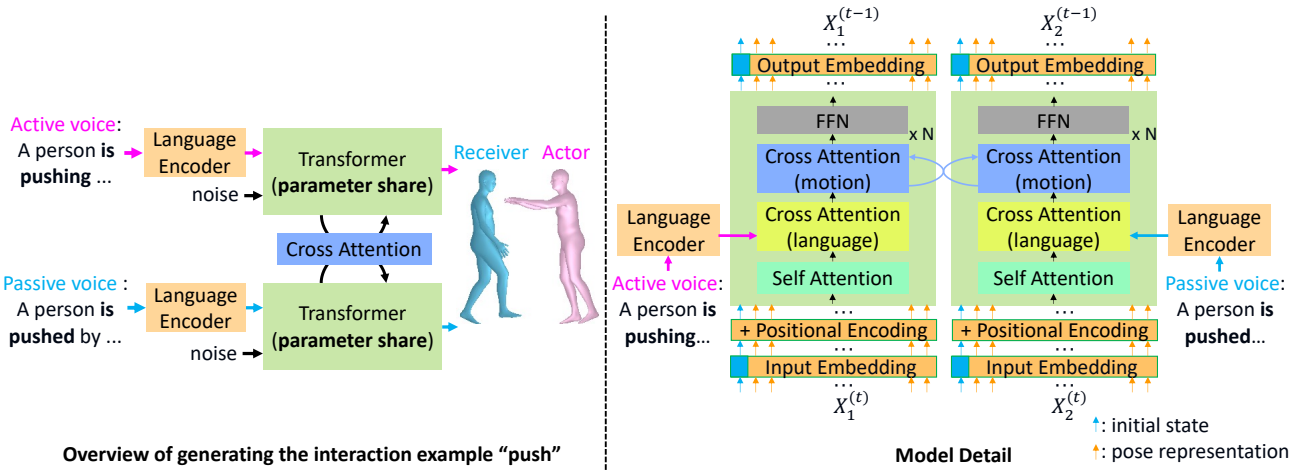

Figure 2. Our model generates interactions with two Transformers sharing parameters and a cross attention module connecting them. To instruct individual roles for asymmetric interactions, we input active and passive voice descriptions to the corresponding Transformer responsible for generating an actor and a receiver, respectively. The detail of our model is shown in the right side of the figure. The Transformer consists of N attention blocks containing three attention modules and one feedforward network (FFN). The generated pose sequence is rendered with the SMPL human model [28] for visualization.

3.2. Extension to Interaction Generation

We now extend the base method to generate interactions consisting of two motions. Two human motions are denoted as $X_1 = [x_0^1, ..., x_F^1]$, $X_2 = [x_0^2, ..., x_F^2]$., where x_0 is the initial position and orientation of the body represented in a coordinate space determined by the relationship of two actors, indicated as "initial state" in Fig. 2, and $[x_1, ..., x_F]$ are the pose representations at each time, corresponding to "pose representation" in Fig. 2. As natural interactions require accurate relationship between two bodies, x_0 is a crucial element for interactions. x_0 is omitted in single motions by setting the initial pose to face forwards. We describe the details in Sec. 4.1. x_0 and $[x_1, ..., x_F]$ are processed with different embedding layers at the input and output.

We consider the case where we have a dataset $((X_1, y_1), (X_2, y_2))$ with a sentence assigned to each motion. A simple extension of the previous work to generate interactions would be to use a single Transformer to generate two human actions at once, in other words, treating X_1 and X_2 as single data with only one description y_1. However, this method cannot assign different roles for each human. Therefore, we use two Transformers that share parameters as in Fig. 2, and for asymmetric interactions, we input the active and passive texts, respectively. For symmetrical interactions, we input active voice text to both of them.

Generating interactions requires matching the positional relationships and timing of each other's actions. To achieve this goal, we introduce cross attention between two human motions. We denote query, key, value features calculated from the features of X_1 and X_2 as $Q_1, Q_2 \in \mathbb{R}^{(F+1)\times d}$, $K_1, K_2 \in \mathbb{R}^{(F+1)\times d}$, and $V_1, V_2 \in \mathbb{R}^{(F+1)\times d}$, respectively, where d is the feature dimension. Following the base model, we employ an efficient strategy [44] to obtain

$$\begin{aligned} R_1 &= \rho(Q_1)(\rho(K_2^\top)V_2) \\ R_2 &= \rho(Q_2)(\rho(K_1^\top)V_1) \end{aligned} \tag{4}$$

corresponding to X_1 and X_2 respectively, where ρ is the softmax function. By adding R_1 and R_2 to respective features of X_1 and X_2, the influence from each other is injected into each motion feature, and the model learns the relationship between actions and reactions, as well as cooperative motions that are tailored to each other. The final loss function can be expressed by modifying Eqn. 3 into

$$\begin{aligned} \mathcal{L} =& E_{t\in[1,T],(X_1^{(0)},X_2^{(0)})\sim q(X^{(0)}),\epsilon_1\sim\mathcal{N}(0,I),\epsilon_2\sim\mathcal{N}(0,I)} \\ & \left[\|\epsilon_1 - \epsilon_\theta(X_1^{(t)}, t, y_1)\| + \|\epsilon_2 - \epsilon_\theta(X_2^{(t)}, t, y_2)\|\right]. \end{aligned} \tag{5}$$

Because the proposed method is based on two Transformer branches to treat each human actor separately, parameters of the pretrained single motion generation model [59] can be used for initializing our model. The knowledge from single-person motion further enhances our role-aware interaction generation, as our method focuses on learning the correct correspondence between the active and passive motions and their accurate descriptions.

3.3. Actor and Receiver Separation

In this section, we consider the case where we do not have the complete annotated dataset $((X_1, y_1), (X_2, y_2))$ as described in Sec. 3.2, but instead a dataset (X_1, X_2, c^i), with only an interaction category assigned to each pair of motions. To employ the method in Sec. 3.2, we require

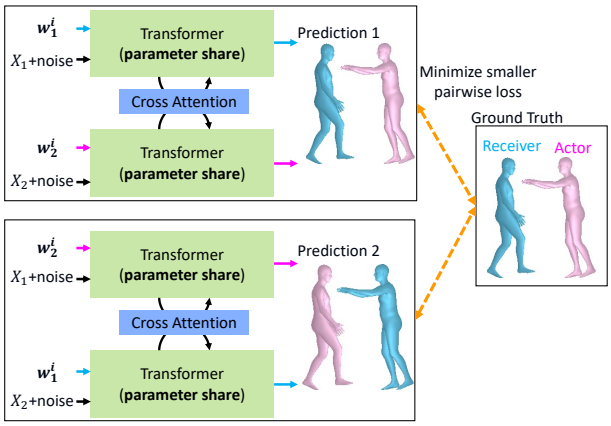

Figure 3. The overview of our method for actor and receiver separation. Two motion pairs are generated with different guidance using w_1^i, w_2^i, and adopt the smaller loss for back propagation.

the knowledge about which of the two humans is an actor or a receiver. Manual annotation of such information requires annotators to view every sequence and assign the roles to the human characters, which is a costly operation. We propose to avoid such effort by providing pseudo labels to create $((X_1, y_1), (X_2, y_2))$ using a methodthat can automatically learn to generate interactions, while also learning to **differentiate** between motions of an actor and a receiver.

We herein prepare two learnable parameters $w_1^i \in \mathbb{R}^{1\times d}$,$w_2^i \in \mathbb{R}^{1\times d}$ per an interaction category c^i. As shown in Fig. 3, the noise is estimated by two different guidance using w_1^i, w_2^i instead of language features in language cross attention in Fig. 2, and the one with the smaller loss is selected for back propagation by modifying Eqn. 5 into

$$\begin{aligned}\mathcal{L} =& E_{t\in[1,T],(X_1^{(0)},X_2^{(0)})\sim q(X^{(0)}),\epsilon_1\sim\mathcal{N}(0,I),\epsilon_2\sim\mathcal{N}(0,I)} \\ &[\min(\|\epsilon_1 - \epsilon_\theta(X_1^{(t)}, t, w_1^i)\| + \|\epsilon_2 - \epsilon_\theta(X_2^{(t)}, t, w_2^i)\|, \\ &\quad \|\epsilon_1 - \epsilon_\theta(X_1^{(t)}, t, w_2^i)\| + \|\epsilon_2 - \epsilon_\theta(X_2^{(t)}, t, w_1^i)\|)] \,. \end{aligned} \tag{6}$$

This is a method similar to PIT [22] used in the field of multi-talker speech separation, whereby w_1^i, w_2^i are expected to learn to obtain features for the motion of either the actor or the receiver. We can **identify** respective roles of w_1^i and w_2^i, by either checking the generated results or matching them with a small number of labeled data automatically. We note that this role separation model itself can be used as a label-based interaction generation model.

To learn to generate more precise interactions corresponding to descriptions, we assign the obtained roles to the training data by adding random noise to each of the training data and calculating which of X_1, X_2 corresponds to w_1^i, w_2^i using Eqn. 6. This allows us to create pseudo labels that identify which human is an actor or a receiver, and assign descriptions y^i_{active} , $y^i_{passive}$ to each motion. By using this automatically generated annotation, we can generate interactions from texts as described in Sec. 3.2.

4. Experiments

In this section, we evaluate the proposed role-aware interaction generation model. We firstly introduce datasets, comparison methods, implementation details, and evaluation metrics. Then, we evaluate the effectiveness of our method for actor and receiver separation. Finally, we quantitatively and qualitatively evaluate the generated interactions by comparing with the baseline and existing methods.

4.1. Experimental Settings

Datasets: We conduct experiments with a large scale skeletal motion dataset, NTU-RGB+D 120 [26]. The dataset contains 94 single motion classes and 26 interaction classes, and we use the latter classes for this evaluation. Out of 26 classes, 9 categories are symmetric interactions such as hugging and shaking hands, while 17 categories are asymmetric interactions such as kicking and pushing. We list all the interaction categories in the Appendix. We translate the categories into sentences that describe the corresponding interaction. We note that we prepare both the active and passive forms of descriptions for the asymmetric interactions.

As existing research has pointed out [10, 29], the original dataset is severely noisy for training motion generation. While Maheshwari *et al*. [29] mapped estimated 3D pose by VIBE [21] to the global trajectory of each human in the dataset by calculating similarity of each subject's orientation, we observed that the global trajectory as well as the similarity-based mapping contained noise. To address this issue and obtain cleaner interactions, we adopted BEV [49], which can estimate 3D pose and global trajectory simultaneously. After preprocessing, we split data by subject IDs to prepare training, validation, and test sets, which contain 14669, 2651, and 3259 sequences, respectively.

Comparison Methods: First, we assign language labels to each motion using the proposed method in Sec. 3.3. We apply the following methods to this data.

As a baseline, we extend the single motion generation method [59] based on diffusion models in a way that allows individual roles to be directed by using two separate Transformers, denoted as "Baseline (Two Trans.)". "Ours" is the proposed method that introduces cross attention between the human motions. Furthermore, as described in Sec. 3.2, the parameters of pretrained single-person motion generation model [59] can be used for initializing our model. We refer to this method as "Ours+pretrained".

We also conduct experiments using existing interaction generation models from interaction labels [29, 11]. As DSAG [11] is designed to control hands in addition to the body, which is out of scope of this research, we only use part of the model for body motion generation. For comparison,

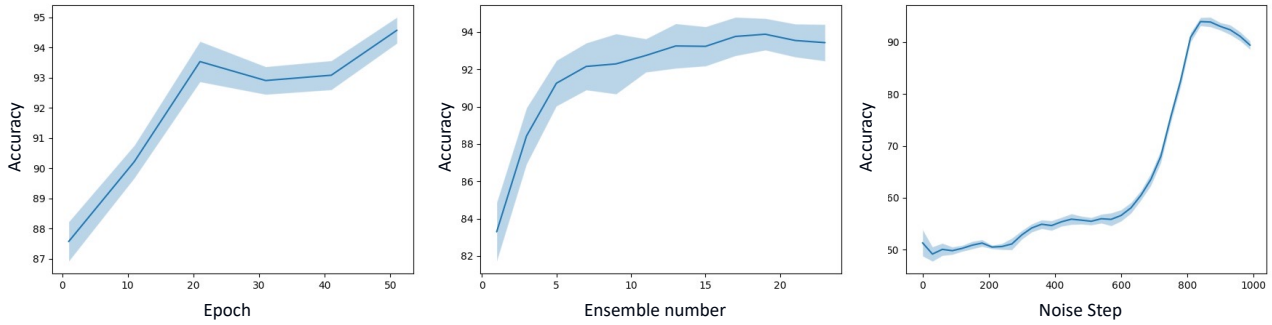

Figure 4. Actor and receiver separation results. From left to right, shifts in accuracy when changing the training epoch, the ensemble number, and the step number of noise, are shown respectively.

we also show the result of the model in Sec. 3.3, which generates role-aware interactions from labels, denoted as "Ours (PIT)". In addition to these models, we generate interactions with one Transformer by simply extending [59], denoted as "Single Transformer". In this method, positional encoding is applied to the two motions separately, and are then concatenated to be used as the input. Because MUGL [29], DSAG [11], and "Single Transformer" cannot control individual roles, we compare with these methods as reference.

Implementation Details: Following the procedure for single-person motions, where the normal of the body plane, formed by two shoulders and two hip joints, is aligned to the Z axis, we normalize each interaction by setting the midpoint of two bodies as origin and forming a mean plane by averaging the coordinates of the two body planes. We transform both motions so that the normal of this mean plane is aligned to the Z axis. We define x_0 as a 4 dimensional vector consisting of 2-dimensional xz coordinates and 2-dimensional rotation of the initial pose. For pose representation $[x_1, ..., x_F]$, we use the same 263 dimensional vector consisting of (r^{va}, r^{vx}, r^{vz}, r_h, $\mathbf{j}^p$, $\mathbf{j}^v$, $\mathbf{j}^r$, $\mathbf{f}$) following [59]. r^{va}, r^{vx}, r^{vz} and r_h are the angular velocity around y axis, velocity in x direction, velocity in z direction and height of the global root on XZ-plane, respectively. $\mathbf{j}^p \in \mathbb{R}^{3(J-1)}$, $\mathbf{j}^v \in \mathbb{R}^{3J}$ denote the position and velocity of the J joints. $\mathbf{j}^r \in \mathbb{R}^{6(J-1)}$ denotes the six-dimensional rotation [61] of each joint. $\mathbf{f} \in \mathbb{R}^4$ denotes whether the four joint points of the feet touch the ground.

For all the diffusion-based models including ours, we set the latent dimension d to 512 and the number of attention blocks N to 8. Diffusion steps T is set to 1000 during which the variance β_t changes linearly from 0.0001 to 0.02. We used Adam optimizer with a learning rate of $2e^{-4}$. We used 4 NVIDIA Tesla V100 GPUs for training, and the total batch size is 480. As for text encoder, we used CLIP ViT B/32 [40] with its weights fixed, followed by four more transformer encoder layers. The latent dimension of the text encoder is 256. Our model is implemented in PyTorch [34].

Evaluation Metrics: We evaluate generated interactions in terms of naturality, diversity and fidelity to the input text. We can evaluate diversity and fidelity in a similar manner as single motion generation. Therefore, following [10], we adopt four evaluation criteria: (i) Frechet Inception Distance (FID) [15], (ii) Recognition Accuracy, (iii) Diversity, (iv) Multimodality. Brief explanation is as follows: (i) evaluates the agreement between the generated data and the true data distribution, (ii) evaluates the ratio of generated interactions that correctly correspond to the input category, (iii) evaluates the diversity of generated interaction, and (iv) evaluates the diversity of generated interaction within each category. Since there is no standard interaction recognition model, we trained a interaction recognition model based on the Transformer with the same training data.

However, for evaluating interactions, consistency between two motions should be considered in addition to naturalness of individual motions. Therefore, we propose a new metric, Mutual Consistency, which evaluates consistency by taking into account aspects unique to interactions, such as relative positions and action-reaction timings of the two motions. For evaluating Mutual Consistency, we trained a model to recognize whether the input motion pair is correct or not with the same training data. The correct pairs are unmodified interactions, and the incorrect pairs are prepared by randomly sampling pairs of motion within the same category. Taking "push" as an example, the incorrect pair is not consistent in terms of direction, timing, and positional relationships between an actor and a receiver. The correct and incorrect pairs were sampled at a ratio of 1 to 1. When given a pair of motions (X'_{i1}, X'_{i2}), the model f returns 1 if they are consistent, and 0 otherwise. Mutual Consistency is calculated as

$$\text{Mutual Consistency} = \frac{1}{M}\sum_{i}^{M} f(X'_{i1}, X'_{i2})\,, \qquad (7)$$

where M is the number of data. We further explain the details of each model used for the evaluation in the Appendix.

Table 1. Quantitative comparison of generation results. The length of the generated motion sequences is set to be the same as the test data. 4 evaluation experiments were conducted and the 95% confidence interval is shown by $\pm$. Entries "text","role","pretrain" indicate whether the method is able to generate interactions from text, to assign individual roles, and to initialize using single-person motion model. Metrics indicated by $\uparrow, \downarrow, \rightarrow$ are better when values are higher, lower, or closer to the Ground Truth, respectively. Multim. and Mut. Cons. are abbreviations of Multimodality and Mutual Consistency, respectively. Best results are in bold.

Methods	text	role	pretrain	Accuracy↑	FID ↓	Diversity→	Multim.→	Mut. Cons.↑
Ground Truth				$84.8^{\pm 0.1}$	$0.0^{\pm 0.0}$	$32.98^{\pm 0.40}$	$12.81^{\pm 0.57}$	$99.6^{\pm 0.0}$
MUGL [29]				$20.1^{\pm 0.5}$	$369.4^{\pm 5.1}$	$23.66^{\pm 0.39}$	$17.71^{\pm 0.67}$	$81.6^{\pm 0.3}$
DSAG [11]				$46.8^{\pm 0.6}$	$311.7^{\pm 2.2}$	$26.11^{\pm 0.33}$	$6.72^{\pm 0.64}$	$88.6^{\pm 1.4}$
Single Transformer	✓			$75.6^{\pm 1.7}$	$12.3^{\pm 1.17}$	$31.82^{\pm 1.04}$	$15.69^{\pm 1.48}$	$\mathbf{99.9^{\pm 0.0}}$
Baseline (Two Trans.)	✓	✓		$73.2^{\pm 0.6}$	$21.1^{\pm 1.0}$	$31.79^{\pm 0.58}$	$13.82^{\pm 1.79}$	$46.3^{\pm 0.7}$
Ours (PIT)		✓		$70.5^{\pm 1.5}$	$\mathbf{11.6^{\pm 0.85}}$	$31.92^{\pm 0.82}$	$16.40^{\pm 0.84}$	$99.3^{\pm 0.1}$
Ours	✓	✓		$76.1^{\pm 0.4}$	$13.1^{\pm 0.6}$	$\mathbf{32.47^{\pm 0.50}}$	$13.61^{\pm 1.25}$	$98.3^{\pm 0.3}$
Ours+pretrained	✓	✓	✓	$\mathbf{79.9^{\pm 0.7}}$	$12.8^{\pm 0.4}$	$31.92^{\pm 0.28}$	$\mathbf{13.45^{\pm 0.94}}$	$98.9^{\pm 0.2}$

4.2. Evaluation of actor and receiver separation

We first quantitatively evaluate the proposed actor-receiver separation. We annotated 15 samples regarding which motion is an actor or a receiver for each of the 17 asymmetric interaction categories to calculate the accuracy.

Fig. 4 shows the evaluation results. The left-most figure shows that actor and receiver separation is automatically conducted correctly as the training progresses. Compared to 50%, which is the chance rate, our model achieves high accuracy of approximately 94%.

The center figure shows that using ensemble method during inference can lead to higher accuracy. This is done by sampling different noise and repeating identification using Eqn. 6, and selecting the conclusion with the majority vote.

The right-most figure shows which time step provides the most accurate separation by changing t from 1 to 1000 in Eqn. 6. The accuracy improves as noise is added to interactions, and peaks at around $t = 800$. This indicates that at the early stages of the diffusion process, where only small amount of noise is added according to noise scheduling, the two estimated noises are very similar. The noise added to each motion in the interaction in the early stages is not indicative of the underlying motion, making the separation of roles from estimated noise difficult. Towards the end of the diffusion process, the accuracy slightly decreases. This is caused by the fact that the interactions are closer to random noise near the end. One concurrent work [4] applied a similar approach for classification, and reported a similar trend.

4.3. Quantitative Evaluation

We present the quantitative evaluation of interaction generation in Table 1. "Ours" performs better than the "Baseline (Two Trans.)" in all metrics, especially in Accuracy and Mutual Consistency. For "Baseline (Two Trans.)", Mutual Consistency is 46.3% while Accuracy is 73.2%. This indicates that the model can generate actions that are typical of the category to some extent, but more than half of the generated interactions show inconsistencies in the positions or timings of the two human motions. This demonstrates the importance of our proposal, Mutual Consistency, because high FID score can be obtained as long as humans act according to the input, even if two motions are not consistent.

"Ours" also outperforms previous interaction generation methods, MUGL [29] and DSAG [11]. This is because MUGL [29] and DSAG [11] compresses two human motions into one latent representation, making it difficult to generate two consistent human motions, also reducing Mutual Consistency. "Ours (PIT)" also achieves better performance than these methods although "Ours (PIT)" is trained with a noisier loss until the roles are accurately separated.

The best performance is achieved by "Ours+pretrained". Particularly, "Ours+pretrained" achieved the highest Accuracy. This is because "Ours+pretrained" uses prior knowledge of correspondence between single human motions and language, and this enables our model to generate interactions that are faithful to the input language. Even when compared to "Single Transformer", which is not affected by the pseudo label error, as the method is not able to freely determine the role of each actor, FID score was comparable.

4.4. Qualitative Evaluation

We qualitatively evaluate the methods by observing the generated results shown in Fig. 5. Compared to MUGL [29] and DSAG [11], "Ours" is able to generate more natural and diverse interactions. We observed many cases from MUGL [29] and DSAG [11] in which the body orientations were facing the opposite directions, leading to lower Mutual Consistency in Table 1. This is caused by the difficulty of generating 3D poses and relative trajectories of all sequences from a latent representation using two separate decoders. In contrast, our method generates them simultaneously to consider each other's information closely.

"Baseline (Two Trans.)" is able to generate each motion

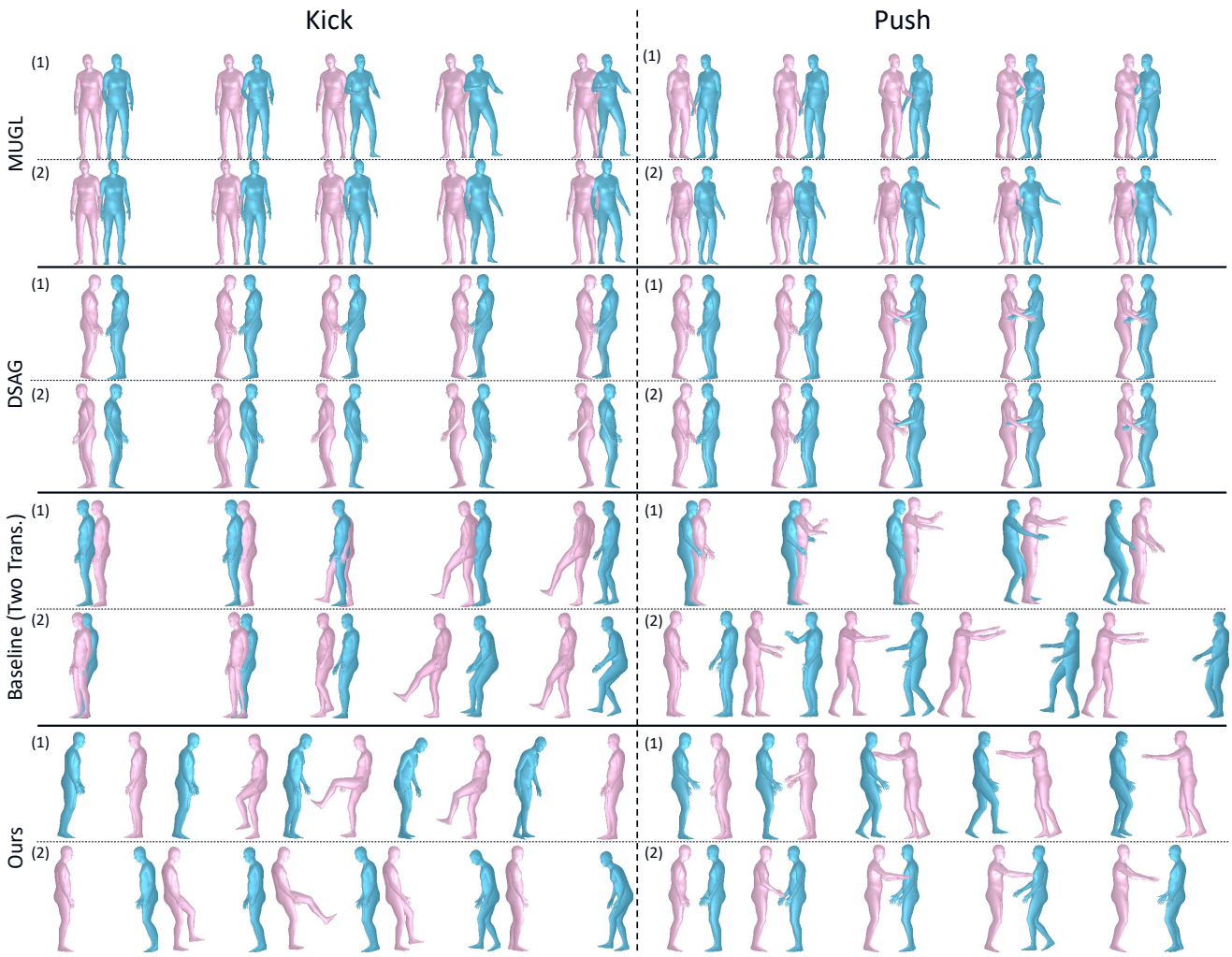

Figure 5. Qualitative comparison of generated results. We compare our method with MUGL [29], DSAG [11] and "Baseline (Two Trans.)". Two examples are visualized for each input. "Ours" generates more natural, consistent, and diverse interactions. For our method, we input active voice descriptions to pink colored human, and passive voice descriptions to blue colored human.

properly according to the input text. However, the consistency between two human motions is not maintained except in the example at the bottom right of Fig. 5. This is the reason why "Baseline (Two Trans.)" achieves low Mutual Consistency. If Mutual Consistency is not considered for evaluation, methods generating these inconsistent interactions can be regarded as acceptable models.

To observe why the proposal is able to achieve high consistency, we visualize one node in the cross attention module in Fig. 6. The pattern of attention along the diagonal shows that the model automatically learns to take into account the other human motion features in the temporal proximity, which is congruent with our instinct and experience.

We visualize the confusion matrix in Fig. 7 to compare the tendency of the methods. We used the original categories corresponding to the provided texts for visualization. DSAG was only able to generate interactions in limited categories. In particular, DSAG failed to capture the difference of similar categories such as interactions that involve reaching out of hands, including "exchange things", "whisper", "wield knife", "shoot with a gun" and "hit with something." Our method can generate almost all category interactions.

5. Failure Cases and Limitations

Fig. 8 shows the failure cases. Even though the proposal generated mostly accurate interactions, there were cases where the relativity between the human actors were not fully accurate, resulting in interactions that are either too close, or too far. There are also some artefacts such as inaccurate foot contact and sliding. We would like to address such issues by introducing physical aspects in the future work.

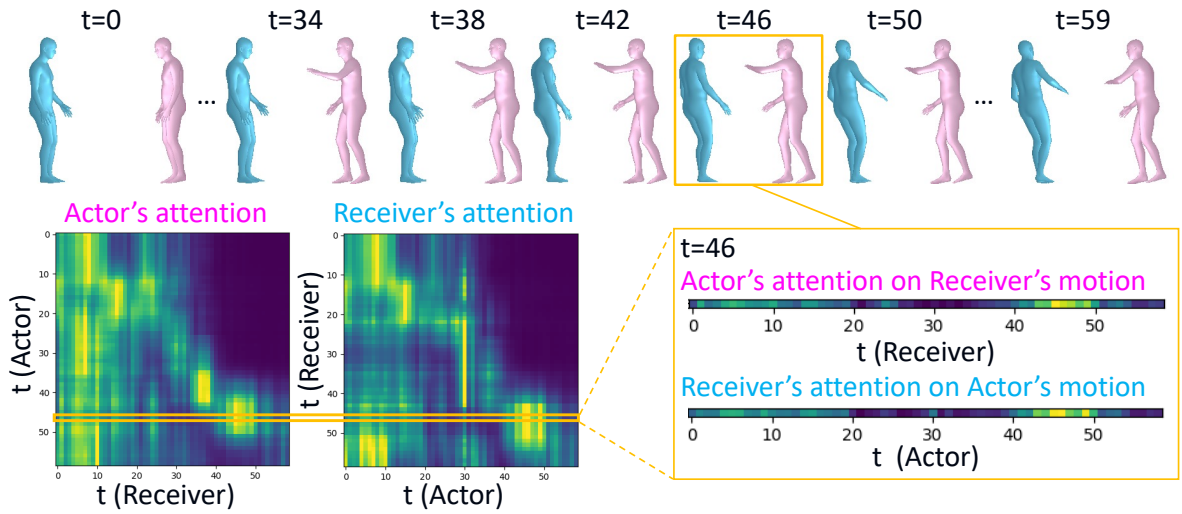

Figure 6. Visualization of one node in the cross attention when generating an interaction in "punch or slap" category. This shows that our model learned to take into account the other human motion features in the temporal proximity.

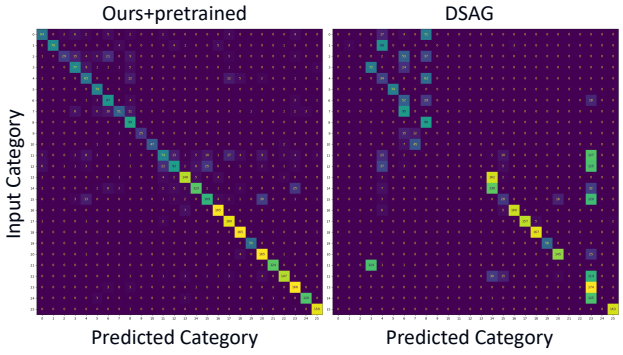

Figure 7. Visualization of confusion matrix when classifying interactions generated by Ours+pretrained and DSAG, respectively.

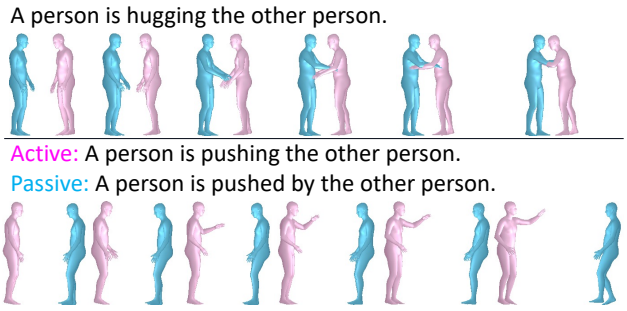

Figure 8. Examples of failure cases. The distance between two humans is either too close or too far.

6. Conclusion

In this study, we focused on a novel role-aware interaction generation task. To avoid introducing additional annotation costs, we proposed a method that can learn to generate actions while separating the motions of an actor and a receiver. We demonstrated that our model can generate consistent interactions between two humans while being faithful to individual instructions. The proposed method outperformed prior methods not only in existing evaluation metrics, but also in the proposed metric evaluating consistency.

Some interactions require additional information to achieve higher precision. For example, amount of force an actor applied to push a receiver would help generate a more accurate reaction. Incorporating physics simulation to further enhance the model is one direction for future work. Although we considered one form of interaction, between two humans, there remains a wider variety, such as interactions in a group. We would like to enhance the proposed method to handle cases where more than two actors are in a motion sequence, and to produce a dataset that contains a wider range of interactions, as well as their detailed descriptions.

Acknowledgements

We would like to thank Masayoshi Kondo from LINE Corporation and Dong-Hyun Hwang from NAVER Cloud Corporation for helpful discussions and assistance.

References

[1] Okan Arikan and David A Forsyth. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)*, 2002.

[2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, 2022.

[3] Michael Büttner and Simon Clavet. Motion matching-the road to next gen animation. *Proc. of Nucl. ai*, 2015.

[4] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023.

[5] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *ECCV*, 2022.

[6] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

[8] Aman Goel, Qianhui Men, and Edmond S. L. Ho. Interaction Mix and Match: Synthesizing Close Interaction using Conditional Hierarchical GAN with Multi-Hot Class Embedding. *Computer Graphics Forum*, 2022.

[9] 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*, 2022.

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

[11] Debtanu Gupta, Shubh Maheshwari, Sai Shashank Kalakonda, and Manasvi Vaidyula. Dsag: A scalable deep framework for action-conditioned multi-actor full body motion synthesis. In *WACV*, 2023.

[12] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics*, 2019.

[13] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020.

[14] Nils Hasler, Bodo Rosenhahn, Thorsten Thormahlen, Michael Wand, Jürgen Gall, and Hans-Peter Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, 2009.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, and Bernhard Nessler. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[17] Yoshiki Ito, Quan Kong, Kenichi Morita, and Tomoaki Yoshinaga. Efficient and accurate skeleton-based two-person interaction recognition using inter- and intra-body graphs. In *ICIP*, 2022.

[18] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *ICME Workshop*, 2014.

[19] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017.

[20] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021.

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

[22] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *TASLP*, 2017.

[23] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In *ECCV*, 2012.

[24] 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*, 2019.

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

[26] 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. In *TPAMI*, 2020.

[27] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. In *ACM Transactions on Graphics (TOG)*, 2015.

[29] Shubh Maheshwari, Debtanu Gupta, and Ravi Kiran Sarvadevabhatla. Mugl: Large scale multi person conditional action generation with locomotion. In *WACV*, 2022.

[30] Qianhui Men, Howard Leung, Edmond S. L. Ho, and Hubert P. H. Shum. A two-stream recurrent network for skeleton-based human interaction recognition. In *ICPR*, 2020.

[31] Qianhui Men, Hubert P.H. Shum, Edmond S.L. Ho, and Howard Leung. Gan-based reactive motion synthesis with class-aware discriminators for human–human interaction. *Computers & Graphics*, 102:634–645, 2022.

[32] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *CVPR*, 2022.

[33] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *ECCV*, 2022.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[35] Mauricio Perez, Jun Liu, and Alex C. Kot. Interaction relational network for mutual action recognition. In *IEEE Trans. Multimedia*, 2021.

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

[37] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, 2022.

[38] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 2016.

[39] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, 2021.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[41] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *CVPR*, 2013.

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

[43] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[44] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *WACV*, 2021.

[45] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[46] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[47] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, 2020.

[48] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 2022.

[49] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022.

[50] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. *ECCV*, 2022.

[51] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *ICLR*, 2023.

[52] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *ICCV Workshop*, 2011.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[54] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014.

[55] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[56] Qing Yu and Kent Fujiwara. Frame-level label refinement for skeleton-based weakly-supervised action recognition. In *AAAI*, 2023.

[57] 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 Workshop*, 2012.

[58] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023.

[59] 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.

[60] Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge, and Tsuhan Chen. Spatio-temporal phrases for activity recognition. In *ECCV*, 2012.

[61] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.