Appendix:
Role-aware Interaction Generation from Textual Description

1. Evaluation Metrics

Accuracy: Accuracy evaluates whether the generated interactions correspond to the input language. Fig. 1 shows the model used for evaluation, which predicts the category of the input interaction. The output features after Global Average Pooling in Fig. 1 is used for evaluating the following FID, Diversity, and Multimodality.

FID: FID is a commonly used metric in the evaluation of generative models, which measures the distance between the generated and true data distributions. When mean and covariance of features of the generated interactions and the ground truth interactions are expressed as (μ, Σ) and (μ', Σ'), respectively, FID is calculated as

$$\mathrm{FID} = \|\mu - \mu'\|_2 + \mathrm{Tr}(\Sigma + \Sigma' - 2\sqrt{\Sigma\Sigma'})\,. \tag{1}$$

Diversity: Diversity measures the variance of the interactions. When two sets of S_d features of interactions generated from random text input are expressed as $[v_1, ..., v_{S_d}]$ and $[v'_1, ..., v'_{S_d}]$, respectively, Diversity is calculated as

$$\mathrm{Diversity} = \frac{1}{S_d}\sum_{i=1}^{S_d} \|v_i - v'_i\|_2\,. \tag{2}$$

Multimodality: Multimodality measures the variance of the interactions in each category. When two sets of S_m features of interactions generated from the same category c are expressed as $[v_{c,1}, ..., v_{c,S_m}]$ and $[v'_{c,1}, ..., v'_{c,S_m}]$, respectively, Multimodality is calculated as

$$\mathrm{Multimodality} = \frac{1}{C \times S_m}\sum_{c=1}^{C}\sum_{i=1}^{S_d} \|v_{c,i} - v'_{c,i}\|_2\,. \tag{3}$$

Mutual Consistency: Mutual Consistency is our proposed metric, which evaluates whether the generated interaction of two humans is mutually consistent. The model is shown in Fig. 1. We prepare a special token called [CLS] and we train the output from it to predict whether the two input actions are correct. This is similar to the training of the next sentence prediction task in BERT [1]. In this study, 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, 0 otherwise. Mutual Consistency is calculated as

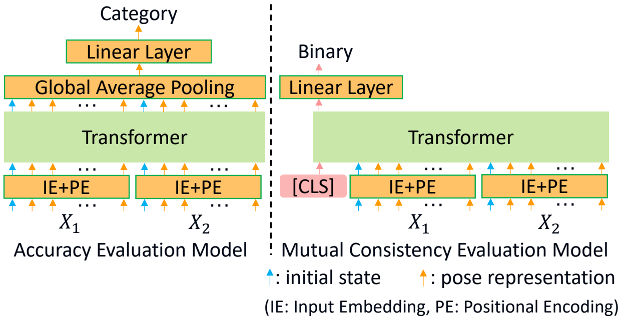


Figure 1. The models for evaluating Accuracy and Mutual Consistency are shown. Input embedding (IE) and positional encoding (PE) are performed as in the proposed model.

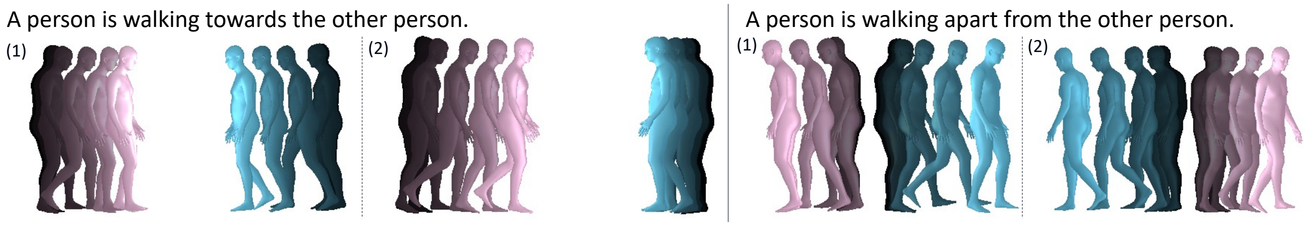
$$\mathrm{MutualConsistency} = \frac{1}{M}\sum_{i}^{M} f(X'_{i1}, X'_{i2})\,. \tag{4}$$

2. Dataset

As mentioned in the paper, 26 interaction categories in NTU-RGB+D 120 [2] dataset are used in the experiment. As shown in Table 1, we translate the labels into corresponding descriptions. For asymmetric interactions, in which there is an actor and a receiver, we translate the description into active and passive voice descriptions.

3. Qualitative Results

Fig. 2 shows the generated interactions that involves walking motions. As can be seen, the method is able to generate two walking motions that are moving towards and apart from each other very accurately. Fig. 3 shows the generated asymmetric interactions. Fig. 4 shows the generated symmetric interactions. In both asymmetric and symmetric interactions, our model generate appropriate interactions



A person is walking towards the other person.

A person is walking apart from the other person.

Figure 2. Generated symmetric interactions that involves walking motions.

according to the roles of the input descriptions. We include some animated figures of examples as additional supplementary material to aid visualization.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 1

[2] 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. 1

Table 1. All interaction categories and the descriptions translated from the categories. When the interaction is asymmetric, we prepare active (A) and passive (P) voice descriptions.

Class	Asymmetric	Descriptions
punching or slapping other person	✓	(A) A person is punching or slapping the other person. (P) A person is punched or slapped by the other person.
kicking other person	✓	(A) A person is kicking the other person. (P) A person is kicked by the other person.
pushing other person	✓	(A) A person is pushing the other person. (P) A person is pushed by the other person.
pat on back of other person	✓	(A) A person is patting on the back of the other person. (P) A person is patted on the back by the other person.
point finger at the other person	✓	(A) A person is pointing a finger at the other person. (P) A person has a finger pointed at by the other person.
hugging other person		A person is hugging the other person.
giving something to other person	✓	(A) A person is giving something to the other person. (P) A person is given something by the other person.
touch other person's pocket	✓	(A) A person is touching the other person's pocket. (P) A person has a pocket touched by the other person.
handshaking		A person is shaking hands with the other person.
walking towards each other		A person is walking towards the other person.
walking apart from each other		A person is walking apart from the other person.
hit other person with something	✓	(A) A person is hitting the other person with something. (P) A person is hit by the other person with something.
wield knife towards other person	✓	(A) A person is wielding a knife at the other person. (P) A person has a knife pointed at by the other person.
knock over other person	✓	(A) A person is knocking over the other person. (P) A person is knocked over by the other person.
grab other person's stuff	✓	(A) A person is grabbing the other person's stuff. (P) A person has a stuff grabbed by the other person.
shoot at other person with a gun	✓	(A) A person is shooting at the other person with a gun. (P) A person is shot at with a gun by the other person.
step on foot	✓	(A) A person is stepping on the other person's foot. (P) A person has a foot stepped on foot by the other person.
high-five		A person is doing a high-five with the other person.
cheers and drink		A person is cheering and drinking with the other person.
carry something with other person		A person is carrying something with the other person.
take a photo of other person	✓	(A) A person is taking a photo of the other person. (P) A person has a photo taken by the other person.
follow other person	✓	(A) A person is following the other person. (P) A person is followed by the other person.
whisper in other person's ear	✓	(A) A person is whispering in the other person's ear. (P) A person is being whispered to by the other person.
exchange things with other person	✓	A person is exchanging things with the other person.
support somebody with hand	✓	(A) A person is supporting the other person with a hand. (P) A person is supported with a hand by the other person.
finger-guessing game		A person is doing finger-guessing game with the other person.

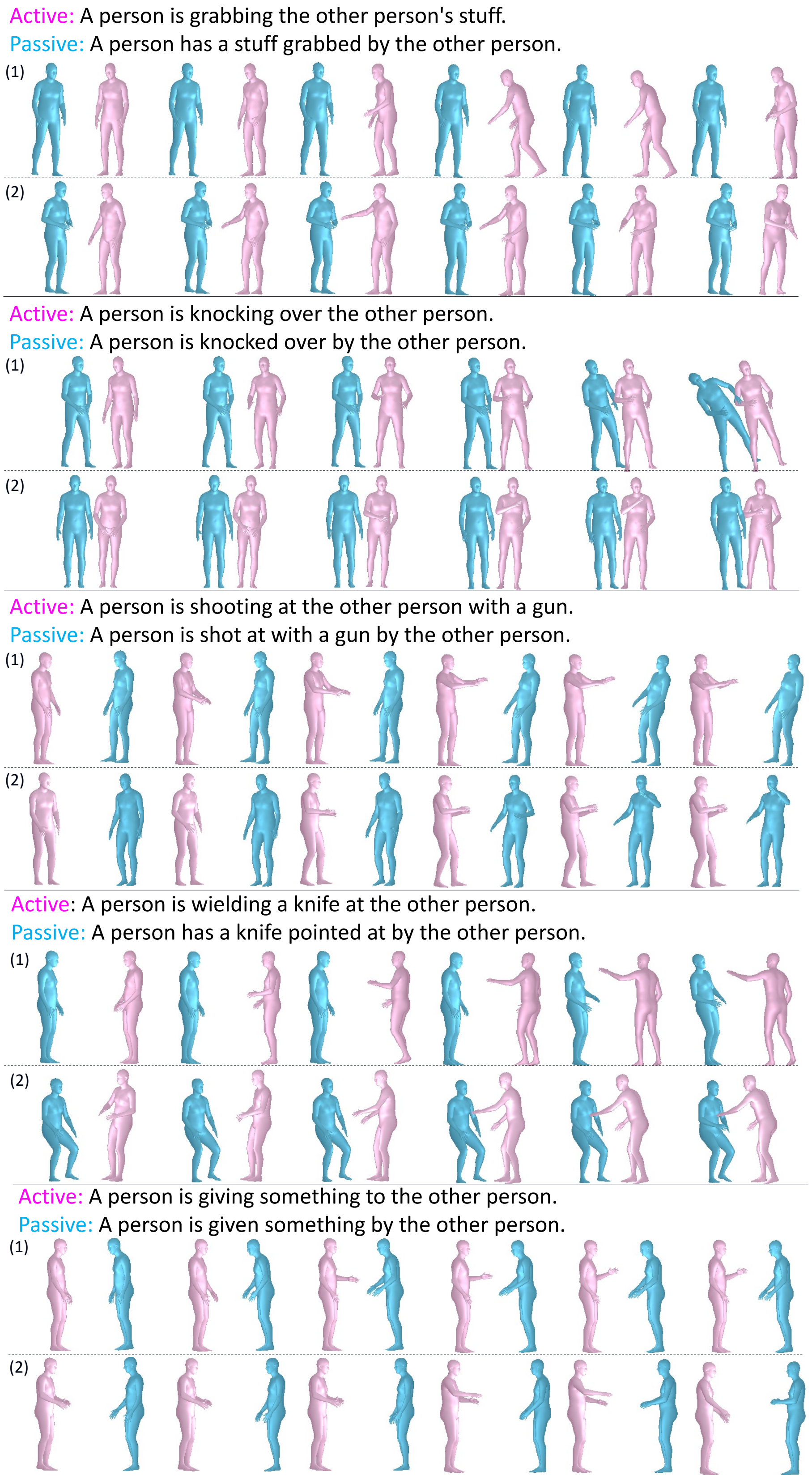

Active: A person is grabbing the other person's stuff.
Passive: A person has a stuff grabbed by the other person.
(1)
(2)
Active: A person is knocking over the other person.
Passive: A person is knocked over by the other person.
(1)
(2)
Active: A person is shooting at the other person with a gun.
Passive: A person is shot at with a gun by the other person.
(1)
(2)
Active: A person is wielding a knife at the other person.
Passive: A person has a knife pointed at by the other person.
(1)
(2)
Active: A person is giving something to the other person.
Passive: A person is given something by the other person.
(1)
(2)

Figure 3. Generated asymmetric interactions. Each human act their own role according to the active or passive voice description.

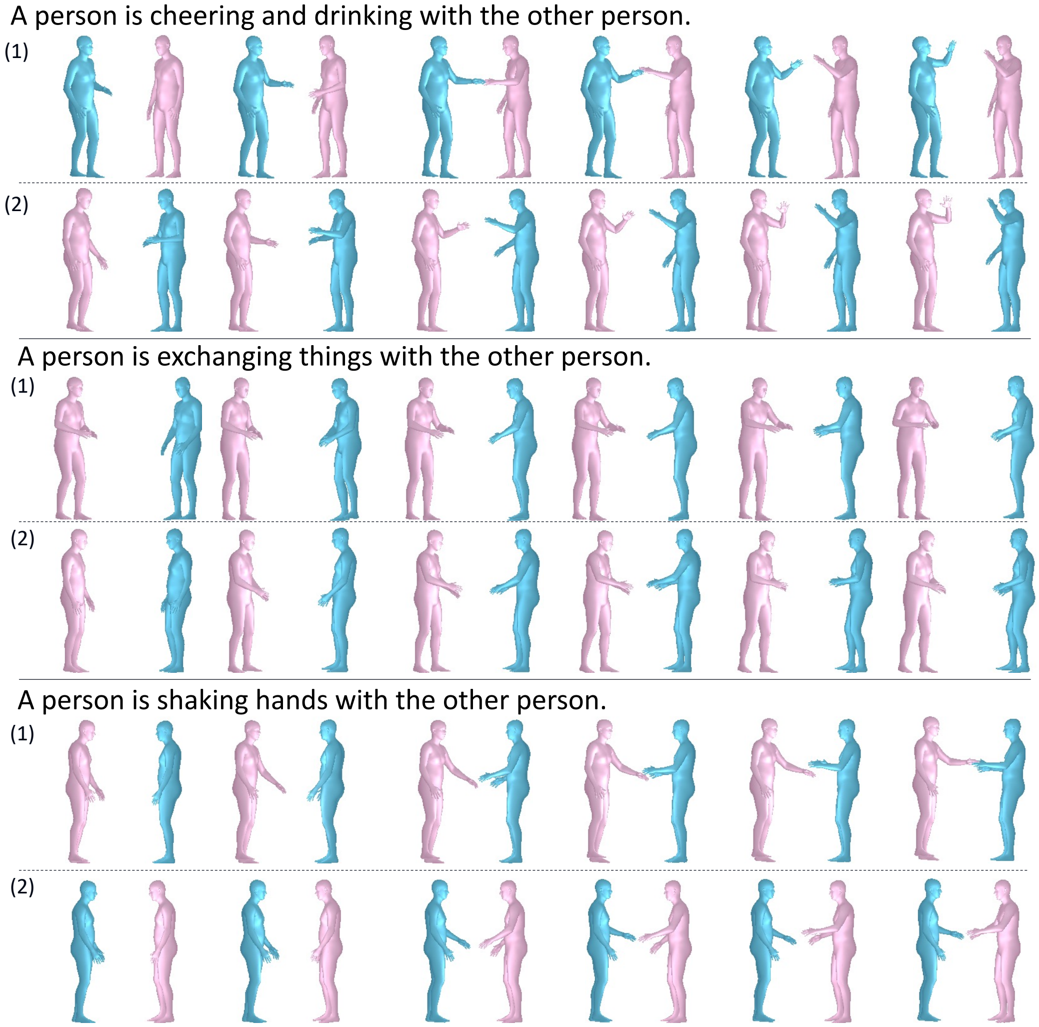

A person is cheering and drinking with the other person.
(1)
(2)
A person is exchanging things with the other person.
(1)
(2)
A person is shaking hands with the other person.
(1)
(2)

Figure 4. Generated symmetric interactions. Both of two humans perform the same motion cooperatively according to the same description.