Inter-X: Towards Versatile Human-Human Interaction Analysis **Appendix**

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

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

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

A. Extra experiments

In this section, we report the results for the remaining four settings of 1) Human interaction captioning; 2) Causal order inference; 3) Stylized human interaction generation, and 4) Personality assessment.

A.1. Human interaction captioning

Human interaction captioning aims to generate precise and diverse textual descriptions given the human interaction sequences. We follow [6] and evaluate for motion captioning models, *i.e.*, RAEs [20], Seq2Seq [14], SeqGAN [4] and TM2T [6]. Similar to the text-conditioned interaction generation task, we simply modify the input and output dimensions to extend these models to two-person settings and change the motion representations to SMPL-X [12].

We follow the same protocol as text-conditioned interaction generation to split our dataset into training, testing and validation sets. Following [6], we also adopt the R Precision and multimodal distance, together with the Bleu [11], Rouge [9], Cider [18] and BertScore [22] to extensively evaluate the performance of the motion captioning models.

The quantitative results are demonstrated in Tab. A.1. We can conclude that TM2T [6] achieves state-of-the-art performance for all the metrics. RAEs [20] fails to model long-term dependencies between human-human interaction sequences and texts, thus leading to low R Precision and linguistic evaluation metrics. Seq2seq [14] and SeqGAN [4] perform better than RAEs [20] by introducing the attention operation and the adversarial learning paradigm.

A.2. Causal order inference

Causal order inference aims to determine the order of the actor and the reactor in the interaction sequences. Similar to the human interaction recognition task, we adopt the models of ST-GCN [21], 2s-AGCN [16], HD-GCN [8], CTR-GCN [2] and MS-G3D [10] as the backbone and model this problem as a binary classification task. From the quantitative results in Tab. A.2, we can derive that MS-G3D [10] yields state-of-the-art performance over all the other methods. However, we found that this task is not that simple, and the performance is far from satisfactory, *i.e.*, only **76.8%**.

A.3. Stylized human interaction generation

We implement the stylized human interaction generation based on the vanilla human interaction generations models, *i.e.*, Action2Motion [5], ACTOR [13], MDM [17], MDM-GRU [3, 17] and Actformer [19]. We add the familiarity level as a style code injected into the model as in [1]. We also report the Frechet Inception Distance (FID) [7], action recognition accuracy, diversity, and multimodality in Tab. A.3. From Tab. A.3, we can derive that Actformer achieves the best FID score and Accuracy, and MDM achieves the best Diversity and Multimodality score.

A.4. Personality assessment

Personality assessment is to automatically obtain the personalities through human interactions. Different from the previous dataset splitting methods, we split the train/test/val sets by person IDs with the ratio of 0.8, 0.15 and 0.05. We also adopt the models of ST-GCN [21], 2s-AGCN [16], HD-GCN [8], CTR-GCN [2] and MS-G3D [10] as the backbone and model this problem as a regression task. We report the R^2 values for each personality element. From the results in Tab. A.4, we can derive that MS-G3D [10] achieves the best performance over all the other methods, except for the element of “Agreeableness”, and CTR-GCN [2] achieves the best R^2 score for the “Agreeableness”.

B. SMPL-X optimization details

Formally, our SMPL-X parameters consist of the body pose parameters $\theta \in \mathbb{R}^{N\times 55\times 3}$, translation $t \in \mathbb{R}^{N\times 3}$ and the shape parameters $\beta \in \mathbb{R}^{N\times 10}$, where N is the number of

[†] Corresponding authors

Methods	R Precision↑			MM Dist↓	Bleu@1↑	Bleu@4 ↑	Rouge ↑	Cider ↑	BertScore ↑
	Top 1	Top 2	Top 3						
Real Desc	0.442	0.645	0.778	3.126	-	-	-	-	-
RAEs [20]	0.094	0.127	0.245	7.554	28.6	9.7	34.1	25.9	10.2
Seq2Seq [14]	0.273	0.436	0.619	4.285	53.8	18.5	45.2	61.9	27.1
SeqGAN [4]	0.206	0.398	0.563	5.447	45.4	14.1	36.8	52.3	21.4
TM2T [6]	**0.375**	**0.583**	**0.674**	**3.493**	**56.8**	**21.6**	**48.2**	**75.5**	**32.7**

Table A.1. Experimental results of human interaction captioning on the Inter-X dataset. **Bold** indicates best results.

Method	Accuracy (%)
ST-GCN [21]	62.3
2s-AGCN [16]	68.2
HD-GCN [8]	70.6
CTR-GCN [2]	74.5
MS-G3D [10]	**76.8**

Table A.2. Experimental results of causal order inference on the Inter-X dataset. **Bold** for best results.

frames. We initialize the subjects' shape β based on their height and weight as [15]. Then a two-stage SMPL-X optimization algorithm is adopted to our Mocap data to obtain the SMPL-X parameters.

In the first stage, we only optimize the pose parameters except that of fingers. The joint energy term

$$\mathbb{E}_j = \frac{1}{N}\sum_{i=0}^{N}\sum_{j\in\mathcal{J}} \|\boldsymbol{J}_j^i(\mathbb{M}(\theta_b, t) - \boldsymbol{g}_j^i\|_2^2 \tag{1}$$

aims to fit the SMPL-X joints to our captured skeleton data, where $\mathcal{J}$ denotes the joint set, $\mathbb{M}$ is the SMPL-X parametric model, $\boldsymbol{J}_j^i$ is the joint regressor function for joint j at i-th frame, θ_b is the pose parameters excluding fingers, $\boldsymbol{g}$ is the Mocap skeleton data. A smoothing term

$$\mathbb{E}_{smooth} = \frac{1}{N-1}\sum_{i=0}^{N-1}\sum_{j\in\mathcal{J}} \|\boldsymbol{J}_j^{i+1} - \boldsymbol{J}_j^i\|_2^2 \tag{2}$$

alleviates the pose jittering between frames. A regularization term

$$\mathbb{E}_r = \|\theta_b\|_2^2 \tag{3}$$

constrains the pose parameters from deviating too much. In total, our optimization objective at the first stage is:

$$\mathbb{E}_1 = \lambda_j\mathbb{E}_j + \lambda_{smooth}\mathbb{E}_{smooth} + \lambda_r\mathbb{E}_r, \tag{4}$$

and we set $\lambda_j, \lambda_{smooth}, \lambda_r = 1, 0.1, 0.01$.

For the second stage, we append the finger pose parameters and jointly optimize the whole-body pose parameters. We especially emphasize fingers' optimization, thus we separate fingers' pose parameters from the body part. Our optimization objective in the second stage is summarized as:

$$\mathbb{E}_b = \lambda_j\mathbb{E}_j + \lambda_{smooth}\mathbb{E}_{smooth} + \lambda_r\mathbb{E}_r, \tag{5}$$

$$\mathbb{E}_h = \lambda_{j_h}\mathbb{E}_{j_h} + \lambda_{smooth_h}\mathbb{E}_{smooth_h} + \lambda_{r_h}\mathbb{E}_{r_h}, \tag{6}$$

$$\mathbb{E}_2 = \mathbb{E}_b + \mathbb{E}_h, \tag{7}$$

we set $\lambda_j, \lambda_{smooth}, \lambda_r = 1, 0.1, 0.01$ for the body part and $\lambda_{j_h}, \lambda_{smooth_h}, \lambda r_h = 10, 0.01, 0.001$ for fingers.

C. The action categories

We provide the names of the 40 human-human interaction categories in Tab. A.5.

D. Samples of textual annotations

We provide some samples of the textual annotations of our Inter-X dataset in Fig. F.1.

E. More visualization results

We provide the rendered RGB frames based on the Unreal Engine in Fig. F.2. We also provide more visualization samples of Inter-X in the supplementary video.

F. Boarder impacts

With our proposed Inter-X dataset, one can facilitate the generative models for synthesizing human-human interaction sequences given detailed textual descriptions with plenty of applications in AR/VR and gaming. For perceptual tasks of human action recognition, one can also build intelligent models for intelligent surveillance.

References

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Unpaired motion style transfer from

Method	FID↓	Acc.↑	Div.→	Multimod.→
Real	$0.281^{\pm 0.002}$	$0.990^{\pm 0.0000}$	$12.890^{\pm 0.028}$	$22.391^{\pm 0.195}$
Action2Motion [5]	$21.182^{\pm 13.319}$	$0.737^{\pm 0.0005}$	$11.492^{\pm 0.032}$	$14.934^{\pm 0.258}$
ACTOR [13]	$9.796^{\pm 0.862}$	$0.867^{\pm 0.0003}$	$11.862^{\pm 0.039}$	$15.174^{\pm 0.245}$
MDM [17]	$11.762^{\pm 1.854}$	$0.912^{\pm 0.0002}$	$\mathbf{13.025}^{\pm \mathbf{0.028}}$	$\mathbf{21.742}^{\pm \mathbf{0.106}}$
MDM(GRU) [17]	$31.688^{\pm 4.492}$	$0.753^{\pm 0.0006}$	$12.259^{\pm 0.039}$	$16.271^{\pm 0.206}$
Actformer [19]	$\mathbf{8.544}^{\pm \mathbf{0.684}}$	$\mathbf{0.932}^{\pm \mathbf{0.0006}}$	$12.116^{\pm 0.062}$	$16.122^{\pm 0.183}$

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

Method	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
ST-GCN [21]	21.16	25.38	34.91	23.67	13.02
2s-AGCN [16]	23.46	31.27	38.72	24.88	13.57
HD-GCN [8]	25.92	33.19	41.33	26.83	14.29
CTR-GCN [2]	27.78	35.41	43.52	**29.43**	15.63
MS-G3D [10]	**28.36**	**37.88**	**46.23**	29.07	**16.35**

Table A.4. The R^2 values results (%) of the personality assessment on the Inter-X dataset. **Bold** for best results.

video to animation. *ACM Transactions on Graphics (TOG)*, 39(4):64–1, 2020. 1

[2] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, pages 13359–13368, 2021. 1, 2, 3

[3] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 1

[4] Yusuke Goutsu and Tetsunari Inamura. Linguistic descriptions of human motion with generative adversarial seq2seq learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4281–4287. IEEE, 2021. 1, 2

[5] 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. 1, 3

[6] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*, pages 580–597. Springer, 2022. 1, 2

[7] 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. 1

[8] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*, pages 10444–10453, 2023. 1, 2, 3

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1

[10] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, pages 143–152, 2020. 1, 2, 3

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[12] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 1

[13] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *CVPR*, pages 10985–10995, 2021. 1, 3

[14] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 1, 2

[15] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3d measurements. *IEEE transactions on visualization and computer graphics*, 25(5): 1887–1897, 2019. 2

[16] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 1, 2, 3

A01: Hug	A02: Handshake	A03: Wave	A04: Grab
A05: Hit	A06: Kick	A07: Posing	A08: Push
A09: Pull	A10: Sit on leg	A11: Slap	A12: Pat on back
A13: Point finger at	A14: Walk towards	A15: Knock over	A16: Step on foot
A17: High-five	A18: Chase	A19: Whisper in ear	A20: Support with hand
A21: Finger-guessing	A22: Dance	A23: Link arms	A24: Shoulder to shoulder
A25: Bend	A26: Carry on back	A27: Massage shoulder	A28: Massage leg
A29: Hand wrestling	A30: Chat	A31: Pat on cheek	A32: Thumb up
A33: Touch head	A34: Imitate	A35: Kiss on cheek	A36: Help up
A37: Cover mouth	A38: Look back	A39: Block	A40: Fly kiss

Table A.5. The action categories of Inter-X.

[17] 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. 1, 3

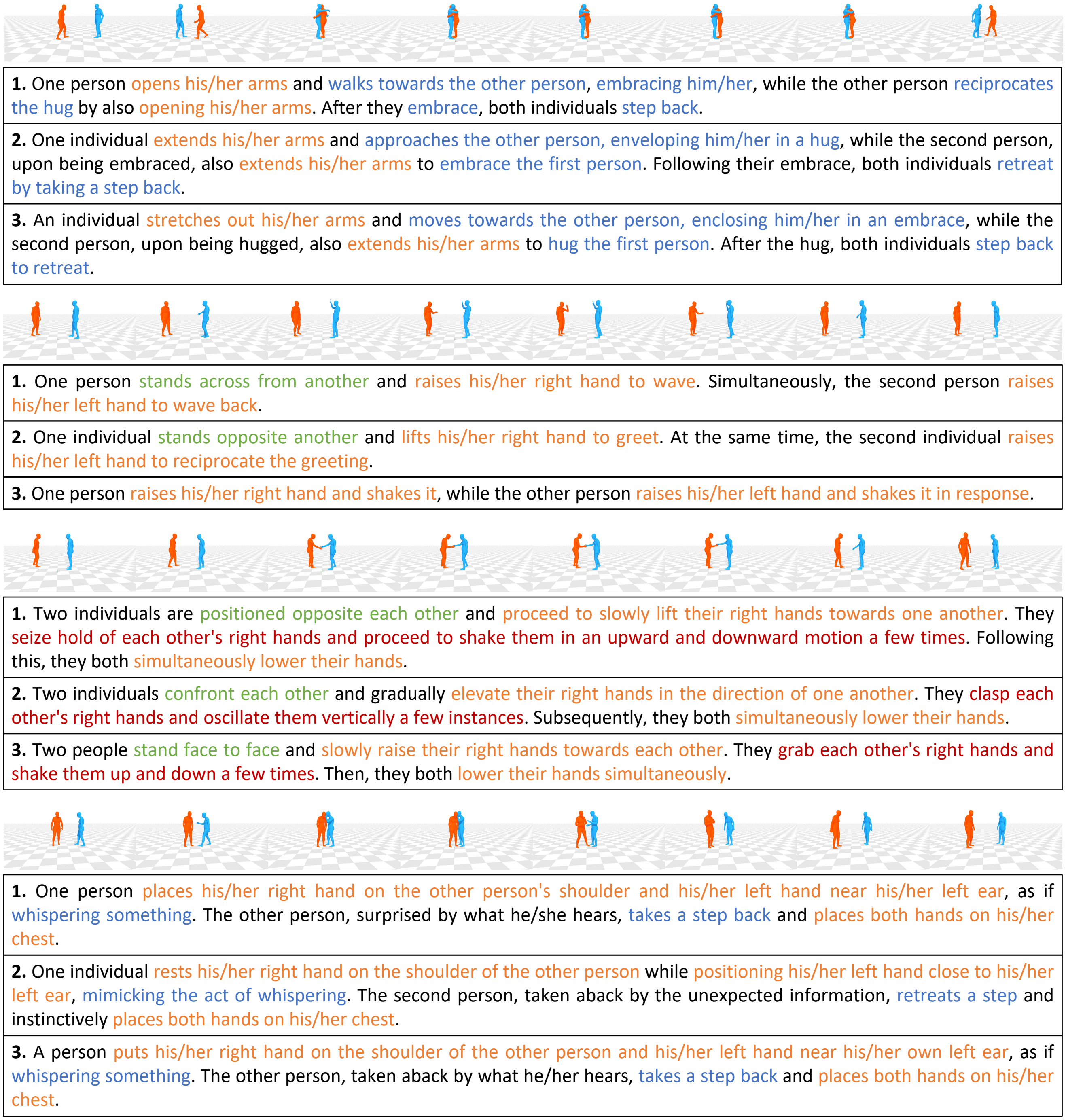
[18] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[19] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *ICCV*, pages 2228–2238, 2023. 1, 3

[20] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*, 3(4):3441–3448, 2018. 1, 2

[21] 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. 1, 2, 3

[22] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 1



1. One person opens his/her arms and walks towards the other person, embracing him/her, while the other person reciprocates the hug by also opening his/her arms. After they embrace, both individuals step back.

2. One individual extends his/her arms and approaches the other person, enveloping him/her in a hug, while the second person, upon being embraced, also extends his/her arms to embrace the first person. Following their embrace, both individuals retreat by taking a step back.

3. An individual stretches out his/her arms and moves towards the other person, enclosing him/her in an embrace, while the second person, upon being hugged, also extends his/her arms to hug the first person. After the hug, both individuals step back to retreat.

1. One person stands across from another and raises his/her right hand to wave. Simultaneously, the second person raises his/her left hand to wave back.

2. One individual stands opposite another and lifts his/her right hand to greet. At the same time, the second individual raises his/her left hand to reciprocate the greeting.

3. One person raises his/her right hand and shakes it, while the other person raises his/her left hand and shakes it in response.

1. Two individuals are positioned opposite each other and proceed to slowly lift their right hands towards one another. They seize hold of each other's right hands and proceed to shake them in an upward and downward motion a few times. Following this, they both simultaneously lower their hands.

2. Two individuals confront each other and gradually elevate their right hands in the direction of one another. They clasp each other's right hands and oscillate them vertically a few instances. Subsequently, they both simultaneously lower their hands.

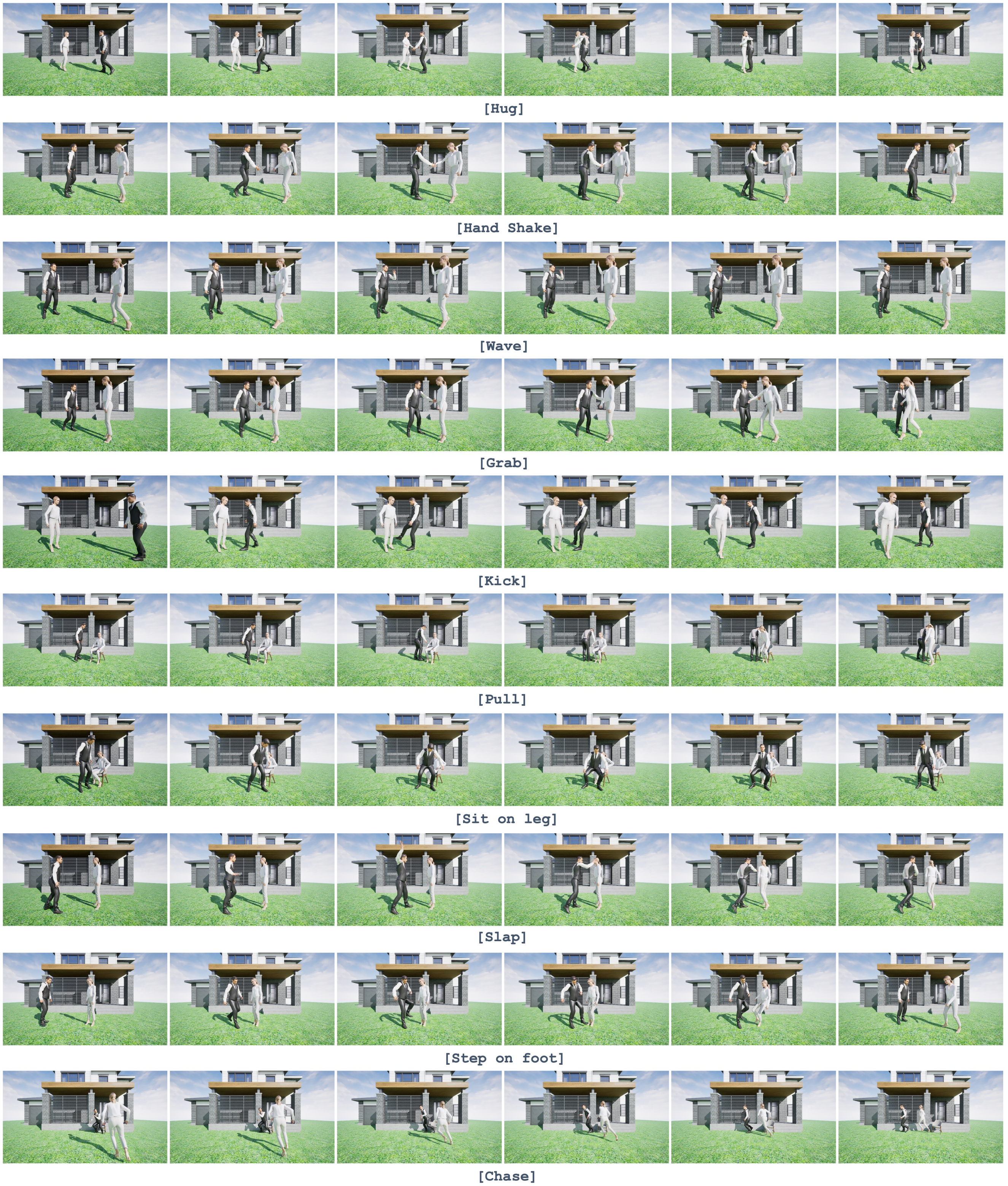
3. Two people stand face to face and slowly raise their right hands towards each other. They grab each other's right hands and shake them up and down a few times. Then, they both lower their hands simultaneously.

1. One person places his/her right hand on the other person's shoulder and his/her left hand near his/her left ear, as if whispering something. The other person, surprised by what he/she hears, takes a step back and places both hands on his/her chest.

2. One individual rests his/her right hand on the shoulder of the other person while positioning his/her left hand close to his/her left ear, mimicking the act of whispering. The second person, taken aback by the unexpected information, retreats a step and instinctively places both hands on his/her chest.

3. A person puts his/her right hand on the shoulder of the other person and his/her left hand near his/her own left ear, as if whispering something. The other person, taken aback by what he/her hears, takes a step back and places both hands on his/her chest.

Figure F.1. Some samples of the textual annotations of the Inter-X dataset.



[Hug]

[Hand Shake]

[Wave]

[Grab]

[Kick]

[Pull]

[Sit on leg]

[Slap]

[Step on foot]

[Chase]

Figure F.2. The visualization results of the rendered RGB frames based on the Unreal Engine.