Proxy-Bridged Game Transformer for Interactive Extreme Motion Prediction

Supplementary Material

This appendix contains supplementary explanations and experiments to support our proposed proxy-bridged game Transformer (PGformer). Section A supplements the settings for the three datasets, including the descriptions of the ExPI, illustrations for the three splits, and explanations for CMU-Mocap and MuPoTS-3D settings used in our experiments. Section B provides more experiment details, results, ablation studies and visualizations.

A. More Information about the Dataset

A.1. ExPI Settings

As described in Section 4.1, 16 actions are recorded in the ExPI dataset, which are split into three data splits: common action split, single action split and unseen action split. Seven of them are common actions (A1–A7), performed by both of the 2 couples. In our experiment, we mainly forcus on the common action split and unseen action split.

We use superscript and subscript to denote the couple number and action split respectively, for example, the common action performed by couple 1 is denoted as $\mathcal{A}_c^1$. The other nine actions are couple-specific and performed by only one of the couples. The actions A8–A13 in unseen action split are performed by couple 1, denoted as $\mathcal{A}_u^1$; while the actions A14–A16 performed by couple 2 are represented as $\mathcal{A}_u^2$.

Common action split. The common actions performed by different couples of actors are considered as training and testing data. Then, training and testing sets contain the same actions but are performed by different persons. In our experiment, following the setting in [13], $\mathcal{A}_c^2$ is the training set and $\mathcal{A}_c^1$ is the testing set.

Single action split. In this split, 7 action-wise models are trained independently for each common action by treating the action from couple 2 as the training set and the same action from couple 1 as the corresponding testing set.

Unseen action split. The entire set of common actions including $\mathcal{A}_c^1$ and $\mathcal{A}_c^2$ are used as the training set for unseen action split, while the unseen actions $\{\mathcal{A}_u^1, \mathcal{A}_u^2\}$ are used as the testing set. Since the testing actions do not appear in the training process, this unseen action split aims at measuring the generalization ability of models.

A.2. CMU-Mocap and MuPoTS-3D Settings

CMU-Mocap contains a large number of scenes with a single person moving and a small number of scenes with two persons interacting and moving. Wang *et al.* [30] sampled from these two parts and mix them together as their training data. All the CMU-Mocap data were made to consist of 3 persons in each scene, and the testing set was sampled from CMU-Mocap in a similar way. The generalization ability of the model is evaluated by testing on the MuPoTS-3D (2 – 3 persons) and Mix1 (6 persons) datasets with the model trained on the entire CMU-Mocap dataset.

B. Experiments

B.1. More Implementation Details

ExPI For training our PGformer on ExPI, we follow the same implementation settings as in [13]. Specifically, we predict future motion for 1 second in a recursive manner based on the observed motion of 50 frames. The network is trained by the Adam optimizer with an initial learning rate of 0.005, which is decayed by a rate of $0.1^{1/E}$ (E is the total number of epochs) every epoch. Our model is trained for 40 epochs with a batch size of 32, and the average MPJPE loss is calculated for 10 predicted frames. And we find that XIA-GCN [13] also has to be trained by 40 epochs to achieve the reported results.

CMU-Mocap and MuPoTS-3D The model predicts the future 45 frames (3 s) given 15 frames (1 s) of history as input. All the persons' pose sequences are forwarded in parallel to the PGformer layers to capture fine relations across themselves and other persons. The gravity loss is not applied to control the center of gravity since the motions in the two datasets are moderate.

Since these two datasets consist of 2–3 persons in each scene, our XQA module should be made adaptive to them. Specifically, each person is denoted as $\boldsymbol{E}^l$, and other persons are concatenated by time as $\boldsymbol{E}^f$ (e.g., 3 persons mean 3 pairs of $\boldsymbol{E}^l$ and $\boldsymbol{E}^f$). This implementation can be adaptive to any number of persons regardless of parameters. The attention score map $\boldsymbol{A} \in \mathbb{R}^{T\times((n-1)\times T)}$, where n is the number of persons, could still be shared by $\boldsymbol{Q}^l$ and $\boldsymbol{Q}^f$, but it only be used to obtain $\boldsymbol{O}^l$ for simplicity ($\boldsymbol{O}^f$ is omitted). The entire process is conducted in an iterative manner over n with the shared parameters. Here we just provide a straightforward solution for $\geq$ 3 extension, and this approach can be easily applied to the scenarios with more than 3 individuals. Instead of squeezing M frames $\boldsymbol{X}_{T-M+1:T}$ into one vector q, we use the last frame x_t as q directly.

B.2. More Discussions on Quantitative Results

ExPI. We further compare the performance gains of our PGformer with XIA-GCN [13] and HRI [20] for each joint in Figure 5. As can be seen, our proposed method gets better results almost on all the joints, and larger performance gains are achieved for the joints of limbs. Since joints on the limbs usually have higher motion frequencies, the figure indicates that our PGformer can better handle high-frequency motions. Comparing ours and XIA-GCN on the follower, larger improvements are achieved for joints on the head and shoulder. We reasonably conjecture that the follower has more extreme motions in Lindy-hop dancing actions (see qualitative results for verification), and our approach can better handle extreme motions.

For SPGSN, we apply it adaptively to the ExPI dataset, and decompose the body joints into upper body and lower body following the same spirit as in its experiments on Human3.6M, CMU Mocap and 3DPW datasets.

Though BP [25] is a **contemporaneous work**, we still compare ours with BP on ExPI in Table B5. Since BP used different data and training settings from them used in other models (e.g., XIA, MSR and HRI), we train BP by the training setup provided by ExPI benchmark [13] for fair comparisons. We also train our PGformer by the training settings provided by BP (see the results of BP trained by XIA and PGformer trained by BP). Besides, BP concatenates the joints of the two persons as the nodes of GCN and apply the spatial-temporal GCN, which means the number of persons should be fixed, while our PGformer can be adaptive to different numbers of persons.

Table B5. Results of MPJPE for the compared models. The mean and standard deviation, denoted as avg and std, are computed by 5 runs. We run the code provided by **BP official GitHub** directly and report the results in brackets since it has experiments on ExPI.

Time (sec)	0.2	0.4	0.6	1.0
PGformer avg ± std	53 ± 0.0	108 ± 0.4	156 ± 1.2	231 ± 1.4
PGformer (trained by BP)	48	100	149	229
BP-paper (our run)	39 (46)	86 (97)	129 (145)	202 (225)
BP (trained by XIA)	74	134	181	256

B.3. More Comparisons on Quantitative Results

Due to the limited space in main paper, we remove some results of AME in Appendix Table B6. And we provide a more complete percentages of improvement of our PGformer compared with other methods at different forecasting time in Figure B7.

B.4. More Qualitative Results

More qualitative results are provided at the end of this Appendix. We show the examples from each action in Figures B8 to B11. From these examples, with the increase of the forecasting time, the result of our PGformer becomes better than those of other compared methods that independently predict the motions of each person (HRI [20] and MSR-GCN [8]) or only study the interactions between the historical motions (XIA-GCN [13]). For some extreme actions, taking A4 as an example, the poses predicted by MSR-GCN and XIA-GCN at 1 sec forecasting time are weird or look far apart from the ground truths. Nonetheless, our proposed PGformer successfully predicts the poses which are closer to the ground truths.

B.5. More Ablation Study

Ablation studies are performed by using different components and hyperparameters of our network on common actions to identify their roles. In the main paper, we compared different designs of *proxy* impacting the attention map, and here $\boldsymbol{P}' \in \mathbb{R}^{T\times T}$ is given by: $\boldsymbol{P}' = \boldsymbol{W}_t\boldsymbol{T}\boldsymbol{T}^\top(\boldsymbol{W}_t)^\top$. $\otimes$ and $\oplus$ denote broadcast element-wise multiplication and addition, respectively. The results show that the way in Eq. (6) influencing the bidirectional information performs the best.

We further ablate the pose encoder/decoder of our PGformer, the inner elements of our XQA module with *proxy* and different hyperparameters in Table B7. The variants with different pose encoding and decoding networks are first compared, and here 'w/ GCN (enc)' indicates only using a GCN layer as the pose encoder while using FC layers as the pose decoder. Following the same spirit, our proposed model, which can be denoted as 'w/ GCN (dec)', uses an FC layer as the pose encoder and GCNs as the pose decoder. And 'w/ GCN (both)' uses GCNs both in the pose encoder and decoder. 'w/o GCN' uses FC layers instead of GCNs in the pose encoder and decoder. For all the variants, they use a one-layer encoding network and a four-layer decoding network, which means the numbers of layers in the pose encoding and decoding network are kept the same whether FC layers or GCNs are used. From the ablation results, we can observe that using either FC layers or GCNs as the pose encoding and decoding network has a negligible impact on the performances, but using GCNs as the pose decoder is more suitable. In our experiment, we also find that the pose decoding network with four layers performs better since modeling the relationships of the joints is important for the task of extreme motion prediction.

We then ablate the different hyperparameters including the number of templates (M), number of layers and dimensions for model and FFN. From the results, we can find that setting M a small number ($M = 3$ is suggested in our proposed architecture) is sufficient to build *proxy*. From Tables 4 and B7, our suggested architecture has 4 PGformer layers in the encoder/decoder with D = 128 and d_{ffn}=1024 for model dimension and FFN, and 4 heads in MHA with d_h=64 for dimension of each head, which is simpler but more suitable.

Table B6. Results of **AME** on the common action split with the two evaluation metrics (in *mm*). Lower values mean better performances. The best and second best performances are respectively marked in **bold** and underlined.

Action	A1 A-frame				A2 Around the back				A3 Coochie				A4 Frog classic				A5 Noser				A6 Toss Out				A7 Cartwheel				AVG			
Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
Res-RNN [21]	59	102	132	167	62	112	152	229	57	102	139	215	48	85	113	157	51	90	120	167	53	94	126	183	74	131	178	265	58	102	137	197
LTD [19]	51	92	116	132	51	91	116	**148**	43	80	103	130	38	70	89	111	39	70	90	116	42	75	94	123	52	101	139	198	45	83	107	137
HRI [20]	34	69	97	130	44	84	115	150	32	65	91	121	27	56	82	112	28	58	85	121	34	66	88	115	42	83	120	171	34	69	97	131
MSR [8]	41	75	99	126	54	96	129	180	41	74	98	135	34	61	82	106	33	59	79	109	42	71	93	124	57	103	146	210	43	77	104	141
XIA [13]	32	68	99	128	41	82	116	163	29	58	84	116	24	**50**	**73**	**96**	**24**	51	75	109	**31**	62	86	114	41	81	115	160	32	65	93	127
Ours	**31**	**66**	**93**	**120**	**40**	**78**	**109**	150	**27**	**54**	**77**	**109**	**23**	**50**	74	98	**24**	**49**	**71**	**104**	**31**	**61**	**84**	**112**	**37**	**77**	**111**	**155**	**30**	**62**	**88**	**121**

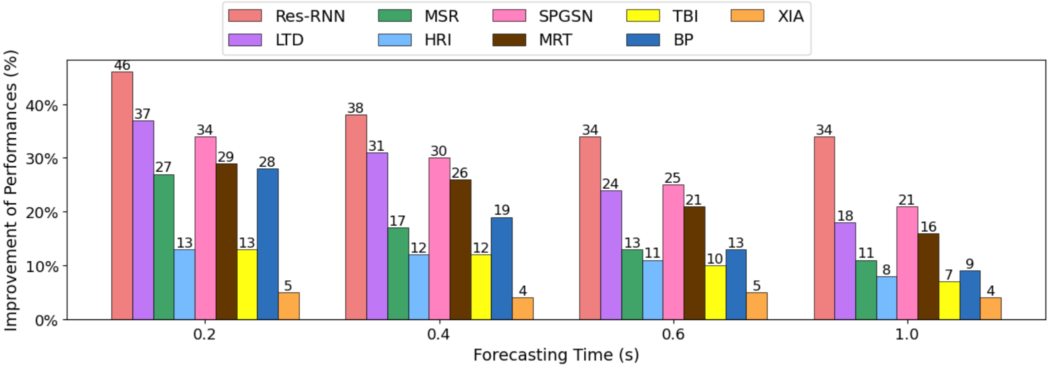


Figure B7. Percentages of improvement of our PGformer compared with other methods at different forecasting time, on the common action split, which are measured by taking the average of the percentages of improvement of average JME and AME error.

B.6. Results of More Metrics

In the main body, we reported Aligned Mean Error (AME) results mainly on ExPI. Here, we expand our evaluation to all datasets (CMU-Mocap and MuPoTS-3D) and report AME and Final Displacement Error (FDE) [24] metric for further comparison. Results of AME and FDE on ExPI and CMU-Mocap is shown in B8. Although T2P surpass PGformer in CMU-Mocap, PGformer keep the highest performance on ExPI.

B.7. Hyperparameter Justification

In our two-person setup, we observed that typically one person often plays a more stable role (e.g., the base in an acrobatic pair), which we term the "leader," while the other is more dynamic (the "follower"). We found that applying a stronger gravity constraint on the leader helps the overall prediction remain physically plausible. The leader's center-of-mass (CoM) should not fluctuate unrealistically (e.g., the base shouldn't hop wildly if they are lifting someone). Thus, we set λ_l = 0.01 to moderately penalize large CoM height deviations for the leader. The follower, conversely, might perform jumps or be lifted, so their CoM can vary more; a heavy penalty could dampen these legitimate motions. Therefore, we chose a much smaller weight λ_f = 0.0001 for the follower's gravity loss, just enough to curb egregious failures (like the follower "floating" or sinking) without hindering necessary movement. We tried several orders of magnitude for these weights. If λ_l is too high (e.g., 0.1), the leader's movements became overly constrained – the model would sometimes predict an unnatural crouch to

Table B7. Ablation study on the pose encoder/decoder, the inner elements of XQA module with *proxy* and different hyperparameters. d_{ffn} is the hidden dimension of the FFN.

	JME				AME			
Time (sec)	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
Proposed	**53**	**108**	**156**	**231**	**30**	**62**	**88**	**121**
w/ GCN (enc)	53	108	157	233	31	63	90	125
w/ GCN (both)	53	108	157	233	31	62	88	122
w/o GCN	53	109	158	234	30	62	88	123
$M = 8$	53	109	159	236	31	63	90	125
$M = 16$	53	109	158	235	31	62	88	123
3-layer	53	110	159	236	31	63	90	124
6-layer	53	110	161	238	31	63	90	125
d_{ffn}=512	54	111	162	240	31	64	92	127
d_{ffn}=2048	54	110	161	237	31	63	91	126

Table B8. Results of AME and FDE on ExPI and CMU-Mocap.

		ExPI			CMU-Mocap		
	Method	0.2	0.6	1.0	0.2	0.6	1.0
AME	TBIFormer	34	98	133	27	84	118
	T2P [14]	34	96	128	**24**	**78**	**110**
	XIA	32	93	127	-	-	-
	Ours	**30**	**88**	**121**	27	82	116
FDE	TBIFormer	36	114	180	18	72	133
	T2P [14]	35	111	176	**17**	**66**	**127**
	XIA	32	106	174	-	-	-
	Ours	**31**	**103**	**168**	20	74	133

Table B9. Comparison of computational cost by using 40G-A100 with batch size of 32/64 in ExPI/CMU-Mocap's training.

	ExPI			CMU-Mocap	
Model	Params	GPU Memory	Train Time	Params	Train Time
MSRGCN	12.73M	7%	45s/iter	15.10M	52s/iter
MRT	5.52M	12%	15s/iter	6.61M	20s/iter
TBIFormer	6.65M	18%	30s/iter	7.26M	39s/iter
XIA	8.50M	6%	25s/iter	9.8M	31/iter
Ours	7.89M	5%	40s/iter	7.17M	49s/iter

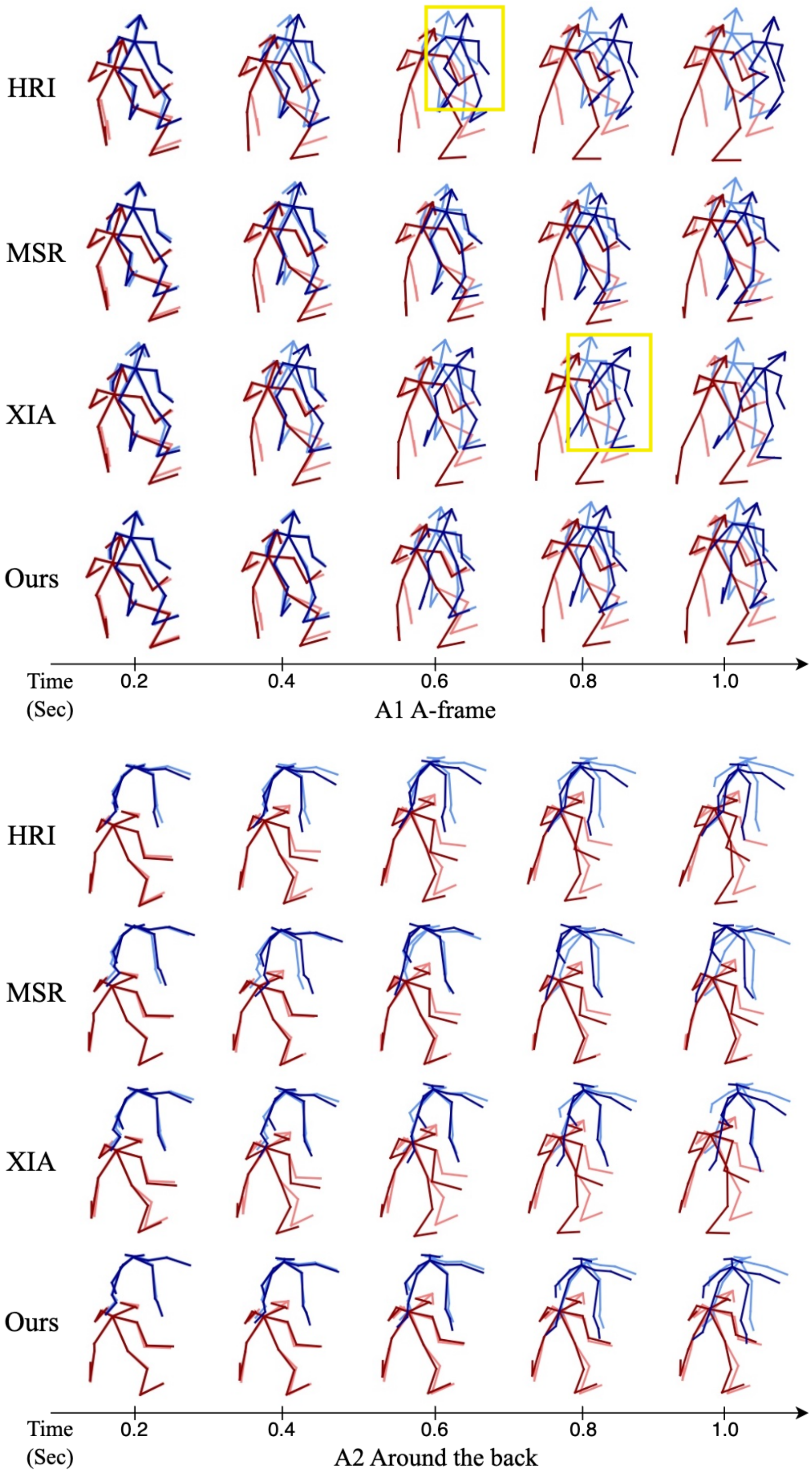
minimize CoM change, hurting accuracy. If λ_l is too low (e.g., 0.001), we observed occasional instabilities in long-term predictions (the leader's pose would drift upward or downward over 1s in unrealistic ways). The value 0.01 provided a good balance, significantly improving long-term stability with minimal impact on short-term accuracy. Similarly for λ_f, values higher than 0.0001 started to noticeably impede the follower's extreme motions (e.g., under-predicting jump height), whereas lower provided no benefit. Thus 0.0001 was the sweet spot for follower: it subtly guides the CoM without sacrificing dynamics.

B.8. Computational Cost

The comparison of computational cost is shown in B9. The model size of our PGformer is insensitive to the input sequence's shape, and the change in model parameters is caused by the variation in the number of body joints in the first embedding layer.

Ethics Statement. Our original intention for this research is to protect people's safety in autonomous vehicles, collision avoidance for robotics and surveillance systems. The potential negative societal impacts include: (1) our approach can be used to synthesize highly realistic human motions, which might lead to the spread of false information; (2) there are still concerns about the invasion of people's privacy since our approach requires real behavioral information as input, and we are concerned that this may expose the identity information. Nonetheless, on the positive side, our model operates on the processed human skeleton representations instead of the raw data, which contains much less identification information.

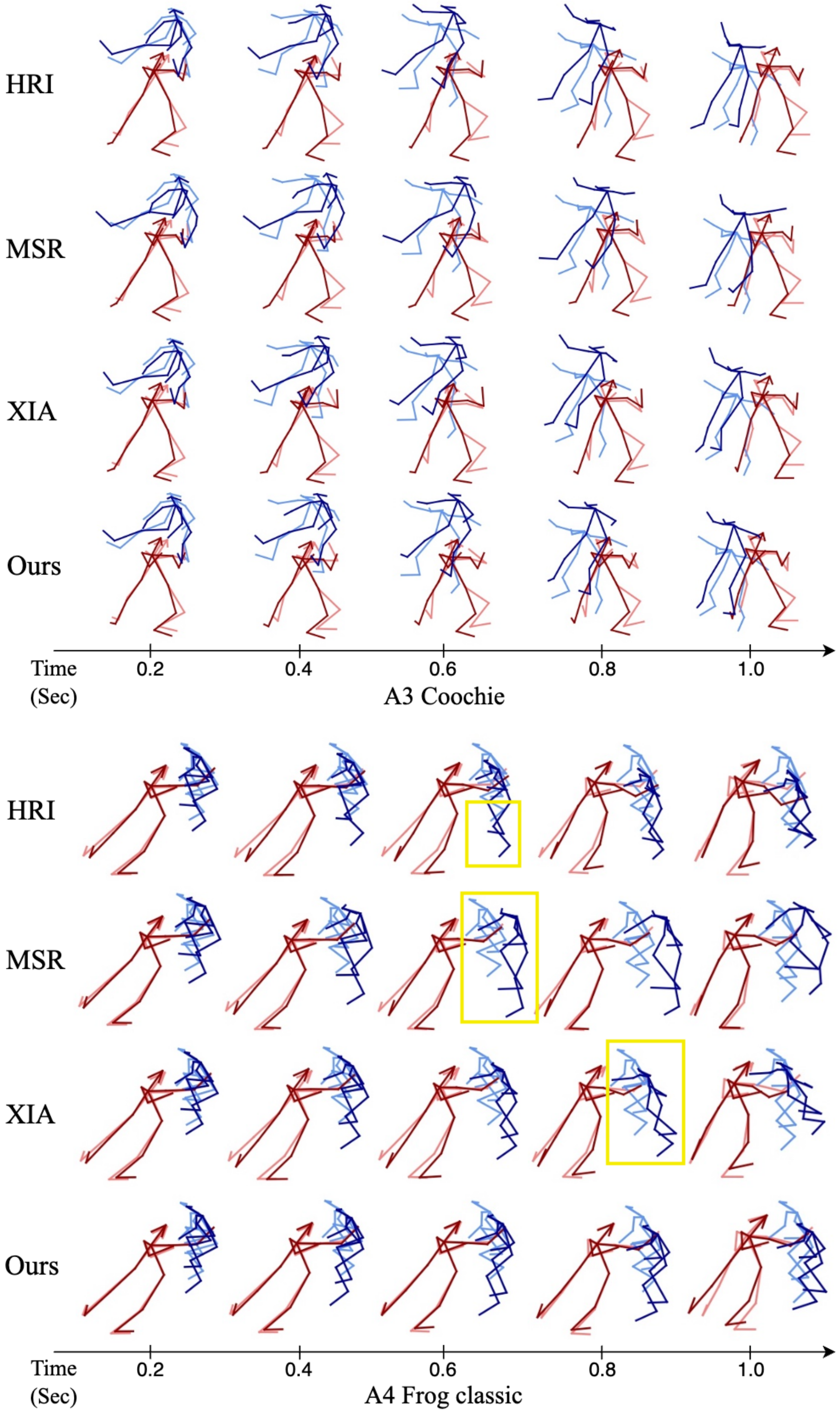
Discussion of Limitations. This paper mainly focuses on modeling multi-person extreme actions, while the motions from different actions vary greatly. Hence, it is hard to verify the effectiveness of our PGformer on other extreme actions due to the lack of such datasets. Besides, we only conduct the ablation study on ExPI to decide the architecture of our model. The performances on CMU-Mocap and MuPoTS-3D datasets would be further improved if tuning some hyperparameters.



A1 A-frame

A2 Around the back

Figure B8. Qualitative results of actions A1 – A2 on the common action split. Dark red/blue represents the prediction results, while light red/blue indicates the ground truths.



A3 Coochie

A4 Frog classic

Figure B9. Qualitative results of actions A3 – A4 on the common action split. Dark red/blue represents the prediction results, while light red/blue indicates the ground truths.

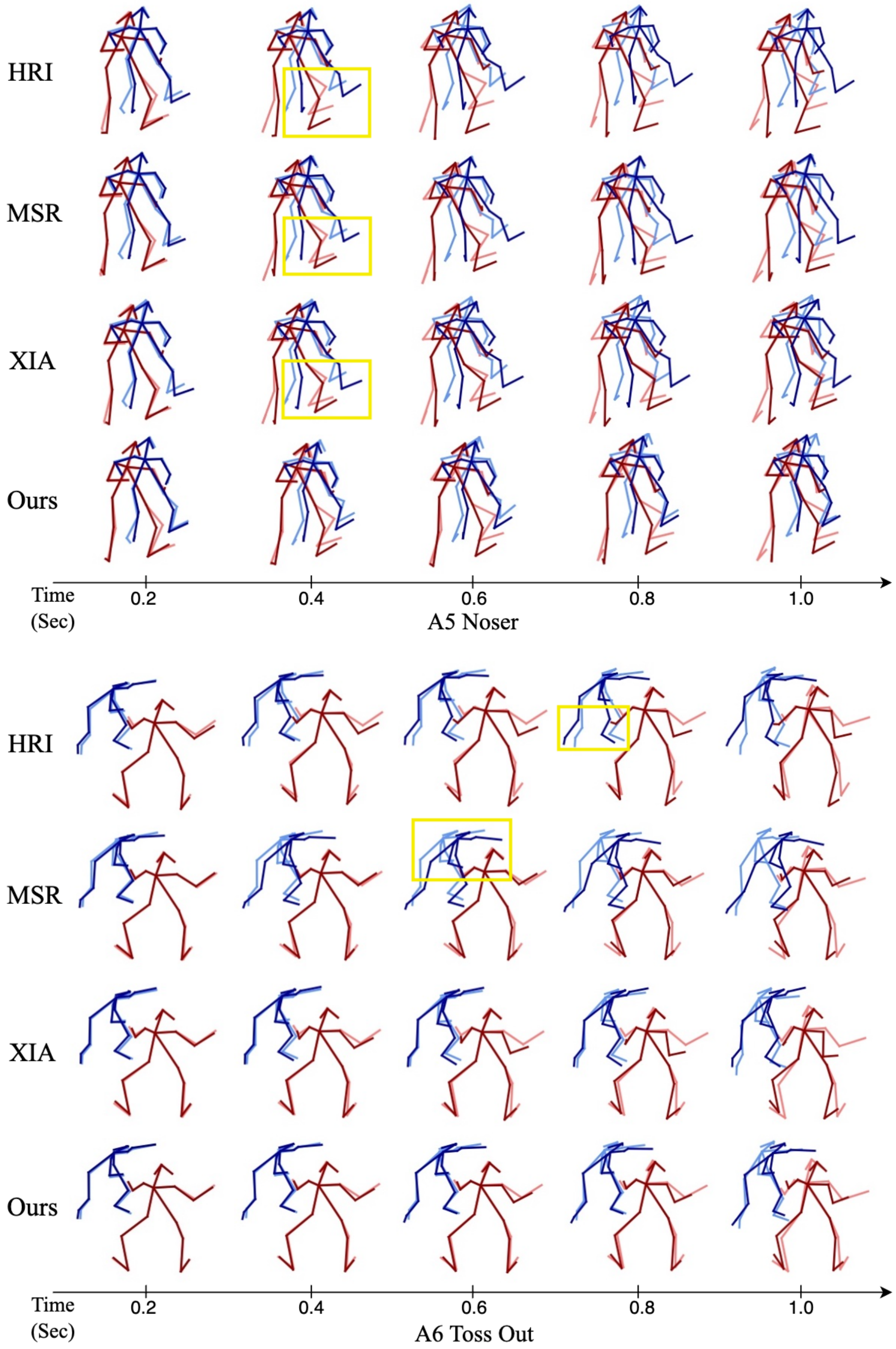


Figure B10. Qualitative results of actions A5 – A6 on the common action split. Dark red/blue represents the prediction results, while light red/blue indicates the ground truths.

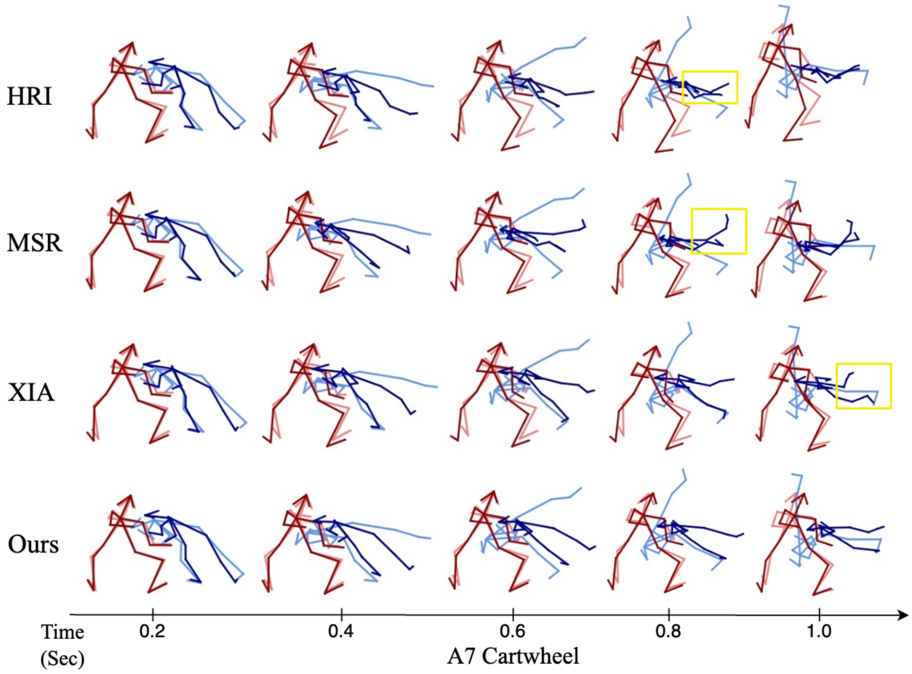


Figure B11. Qualitative results of action A7 on the common action split. Dark red/blue represents the prediction results, while light red/blue indicates the ground truths.