Supplementary for Persistent-Transient Duality: A Multi-mechanism Modeling for Human-Object Interaction

Hung Tran[1], Vuong Le[2], Svetha Venkatesh[1], Truyen Tran[1]
[1]Applied AI Institute, Deakin University, [2]Amazon
{tduy, svetha.venkatesh, truyen.tran}@deakin.edu.au, levuong@amazon.com

In this supplementary material, we provide extra details and analysis of the proposed Persistent-Transient Duality concept and models. They include:

1. Dataset and experiment settings.
2. Model size comparison.
3. Detailed numeric performance results.
4. Extra visualizations.
5. Analysis of the impact of the post-dating window sizes.
6. Additional analysis on the egocentric property in WBHM
7. Ablation studies on Bimanual Action Dataset
8. Generalization analysis on Bimanual Action Dataset
9. The effectiveness of Persistent-Transient Duality in Trajectory Prediction

1. Dataset and experiment settings

1.1. Whole-Body Human Motion Dataset

Dataset details. We retrieved the WBHM dataset using the provided API[1] [7]. The retrieving process involves selecting videos that contain at least one human entity and a table. Our retrieved dataset includes 233 videos with 20 different object types, which is a bigger version of the dataset used in the previous work [2] (190 videos, 15 object types). Among 233 videos in our WBHM dataset, 215 contain 1 human and 18 with 2 humans. The moving space of a human in a video ranges from 0.1m to 3.6m, average at 1.02m and with the median of 0.97m. These statistics indicate that this dataset covers a wide natural motion variety of HOI activities, hence caompatible with the PTD model.

The raw features are stored in C3D files, each of which contain the 3D motion of humans and objects in a video. From the raw features, we extracted skeletal vectors of 18 joints ($x \in \mathbb{R}^{54}$) to represent human entities, and 3D bounding box vectors of 8 vertices ($x \in \mathbb{R}^{24}$) to represent object entities, sampled at 10Hz, consistent with the compared method [2]. The selected joint positions are shown in Fig. 1

[1]https://motion-database.humanoids.kit.edu/faq/#access_methods

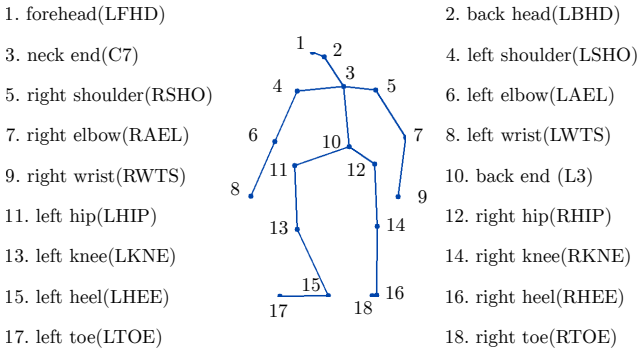


Figure 1: **Human Visual Feature in WBHM** is a vector of 18 3D joints.

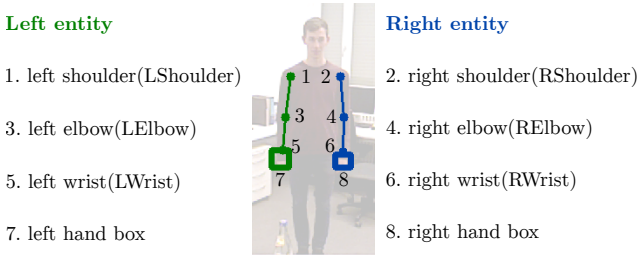
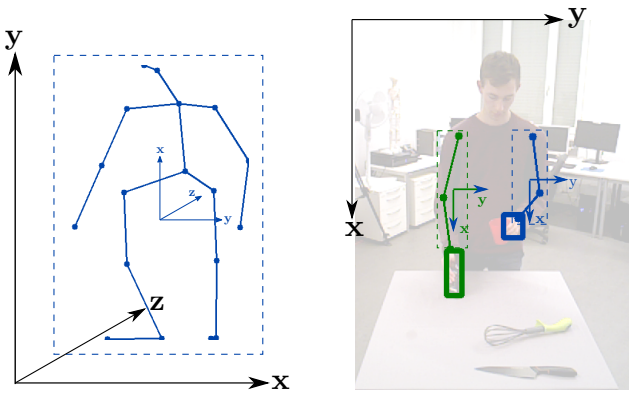


Figure 2: **Human Visual Feature in Bimanual Action Dataset**. In this dataset, each arm is a separate human entity. The visual feature of entity is a concatenation of the arm keypoints and the hand bounding box.

Network parameters and settings. We used GRU in both Persistent Channel, Transient Channel. In Persistent Channel, the human and object GRUs has the hidden states of 512 and 128 dimensions, respectively. Similarly, 512 and 128 are also the dimension sizes of the center and leaf GRUs in the Transient Channel. The Transient-Persistent messages and the entities' raw features are embedded into 64-dimensional vectors. In the Transient Channel, the centroid of the human entity is chosen to be the center of the bounding box around the skeleton (See Fig. 3).

All experiments were conducted on a single GPU



(a) WBHM Dataset (b) Bimanual Action Dataset

Figure 3: Egocentric representations of Transient Channels in WBHM and Bimanual Action Dataset.

NVIDIA Tesla V100 32GB installed on a server of 256 processors and 1024 GB of memory running on Ubuntu 20.04.4 LTS. The network is implemented in Python 3.8.5 with PyTorch 1.7.1 and Deep Graph Library (DGL) 0.5.3 [11]. Our models were trained with the batch size of 128 and optimized by the Adam optimizer [4] with the initial learning rate of 0.002, which decreases by 3% every 5 epochs. We first trained our model with teacher-forcing for 100 epochs, then fine-tuned it with unrolling mechanism for 200 epochs. The experiments were managed and recorded using Hydra [12]. The full implementation will be made publicly available upon acceptance.

Evaluation metrics are the average errors of humans and objects across all prediction time steps. The error of the i^{th} entity is formulated as:

$$\frac{1}{L \cdot M_i} \sum_{t=T+1}^{T+L} \sum_{j=1}^{M_i} \left\| \hat{y}_{i,j}^t - y_{i,j}^t \right\|_2 , \qquad (1)$$

where L is the number of prediction steps and M_i is the number of 3D points in the visual feature of the i^{th} entity. We have $M_i = 18$ for human entities and $M_i = 8$ for object entities. The average error metric of the human entity is equivalent to the standard Mean Per Joint Position Error (MPJPE) metric commonly used in the motion forecasting literature.

1.2. Bimanual Action Dataset

Dataset details. We represent a human entity as the concatenation of a 3-point arm joints (shoulder, elbow, wrist) and a hand bounding box, resulting in a 10-dimension vector (See Fig. 2). An object entity is represented as concatenation of the one-hot object type vector and the 2D bounding box location ($x \in \mathbb{R}^4$). These features are sampled at 10Hz, similar with the those in WBHM.

	WBHM	Bimanual Action
GRU	3.64M	3.42M
CRNN-OPM	3.69M	3.47M
CRNN-OPM-LI	3.73M	3.50M
STS-GCN	2.16M	2.13M
MotionMixer	4.35M	4.27M
PTD (Ours)	3.68M	3.42M

Table 1: Model size comparison.

Network parameters and settings. In this experiment, we re-use the model hyper-parameters in WBHM and choose the centroid of the human entity to be the center of the bounding box around the arm (See Fig. 3).

The system and Python environment are also similar to those in WBHM. We use the batch size of 64 and the Adam optimizer with the initial learning rate of 0.0015, decreasing by 3% every 5 epochs. Our models were first trained using teacher forcing for 10 epochs and then were fine-tuned for 50 epochs.

Evaluation metrics are defined in Eq. (1) with $M_i = 3$ for human key points and $M_i = 2$ for hand and object bounding boxes.

2. Model size comparison

To accurately estimate the contribution of the new modeling scheme, we measure the size of the models using the number of parameters. These measurements are reported in Tab. 1.

The comparable size of PTD with other models confirms that the superior performance of PTD is caused by the proposed modeling scheme, not by the increase in computation power.

3. Detailed numeric performance results

The quantitative performance in WBHM Dataset. We extend the average results in Tab. 2 of the main paper and report the errors at each time step in Tab. 5 (for humans) and Tab. 6 (for objects).

The quantitative performance in Bimanual Action Dataset. The errors at each time of humans and objects in Bimanual Action Dataset are also reported in Tab. 7 and Tab. 8.

4. Extra visualizations

WBHM Dataset We extend the visualizations in Fig. 5, Fig. 6 of the main paper, plotting additional qualitative examples to Fig. 7, Fig. 8 below. The illustrations in Fig. 10 further demonstrate the superior performance of PTD in WBHM dataset compared to the single-mechanism CRNN-OPM-LI [2], whereas those in Fig. 8 exhibit the operation of the Switch Module in different HOI scenarios.

Additionally, in Fig. 9, we demonstrate the controllability of the Transient switch, forcing it to not turn on where it naturally does. The visualization shows that the model seems to accept this intervention, and reasonably forecast the motion as if the agent changes their mind. This idea opens a wide road for exploration into counterfactual and intervention modeling.

Bimanual Action Dataset We also provide qualitative examples of the models in Bimanual Action Dataset in Fig. 10, Fig. 11, demonstrating the superior performance of PTD compared to CRNN-OPM-LI (Fig. 10), and showing the operations of the Switch Module (Fig. 11) in this dataset.

5. The effects of the post-dating window size

We vary the value of the post-dating window size ω (Eq. 14, main paper) and measure its impact on the model performance. As plotted in Fig. 4, the model achieves the best result when we train the switch module to anticipate the next 0.1 seconds (1 time steps) in both WBHM and Bimanual Action Datasets.

6. Additional ablation analysis on the egocentric property in WBHM

We provide the additional ablation analysis on the effectiveness of different aspects of the egocentric property in WBHM in Tab. 2. They include:

1. Without the Egocentric representation. We take off the transformation to egocentric system and have Transient channels running on the global features. This ablation results in weaker performance, especially in human prediction.

2. Without the Egocentric computational structure. We then study the importance of the egocentric graph structure by replacing it with a densely connected graph. This results in even worse performance in both human and object prediction.

3. Without both egocentric, the network cannot adapt to the perspective change when the human starts to interact with objects, resulting in a significant drop in performance.

	Ablation	Human	Obj
1	w/o egocentric rep.	86.10	70.70
2	w/o egocentric struct.	86.20	71.90
3	w/o egocentric property	87.90	72.30
	Full PTD model	**85.53**	**70.69**

Table 2: Ablation studies in Bimanual Action Dataset.

7. Ablation studies on Bimanual Action Dataset

We examine the roles of PTD's core components by making ablations from the model. The performances on Bimanual Action Dataset are reported in Tab. 3. They include:

1. **Without Transient channel**. We turned off Transient Channel. Being alone, Persistent Channel performs significantly worse than when in pair with its Transient partner, showing the significance of the duality.

2. **Without Persistent channel.** Here we instead look at the performance of the Transient Channel alone. It also has much worse performance than the full duality.

3. **Without Egocentric representation.** We take off the transformation to egocentric system (Eq. 5, main paper) and have the model running on the global features. This ablation hurts the performance proving the appropriateness of egocentric representation for the Transient process.

4. **Without Egocentric computational structure.** We then replace the egocentric computational structure (Eq. 4, main paper) with a fully-connected graph. This also has a detrimental effect on the model's performance.

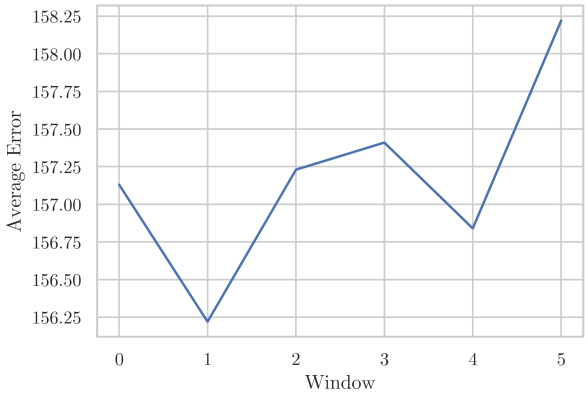
5. **Without both Egocentric egocentric aspects,** the model has worse performance than when each individual aspect is removed.

6. **Heuristic switch**. This experiment probes the need for the *Transient Switch* by replacing it with a heuristic rule. This hard-coded switch can still make use of the duality and have better performance than the single Persistent channel (row 1). However, being too stiff, it cannot represent the switching patterns and failed to reach the full potential of the duality (row 0).

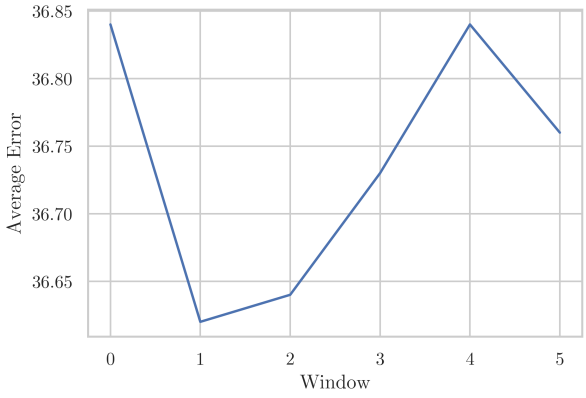
7. **Switch without spatial discount factor.** Without the discount factor γ_i^t (Eq. 10, main paper), the *Transient Switch* module could not respond fast enough to immediate social interaction development, resulting in slightly weaker performance.

8. **Switch with only discount factor.** However, this quick change factor could not do the job by itself because it is susceptible to noisy patterns in crowded scenes. This results in an even worse performance than in case 4.

9. **Without switch loss.** We study the role of the switch's direct supervision (Sec 3.7, main paper) by setting switch loss weight $\lambda = 0$. This unsupervised switch only relies on



(a) WBHM



(b) Bimanual Action Dataset

Figure 4: The effect of the post-dating window size ω on the model performance. The model achieves the best results when anticipating future interaction in the next 0.1 seconds ($\omega = 1$) in both WBHM and Bimanual Action datasets.

	Ablation	Arm Keypoints	Hand	BoxObj
1	w/o Transient channel	12.1	19.1	7.1
2	w/o Persistent channel	13.3	19.5	8.0
3	w/o egocentric rep.	11.0	19.0	6.9
4	w/o egocentric struct.	11.1	19.1	7.0
5	w/o both egocentric	11.3	19.4	7.0
6	w/ heuristic switch	11.2	19.3	7.2
7	w/o γ	11.0	**18.8**	**6.8**
8	w/ only γ	11.0	19.0	**6.8**
9	w/o switch loss	12.0	19.1	7.1
10	w/o multistage training	11.1	18.9	7.0
	Full PTD model	**10.9**	**18.8**	**6.8**

Table 3: Ablation studies in Bimanual Action Dataset.

weak gradient flowing back from prediction loss and delivers significantly weakened performance.

10. **Without multistage training.** Finally, we study the impact of the multistage training Sec 3.7, main paper) on the model performance. Without such a training procedure, the model suffers from accumulating losses during early epochs and capture less accurate motion pattern, resulting in a decrease in the model performance.

8. Generalization analysis on Bimanual Action Dataset

We compare the performance of PTD and other models on test sequences of lengths different from the ones used in training: **(1) observation length T varies, (2) prediction length L varies**.

The results are measured as the total average error (pixel)

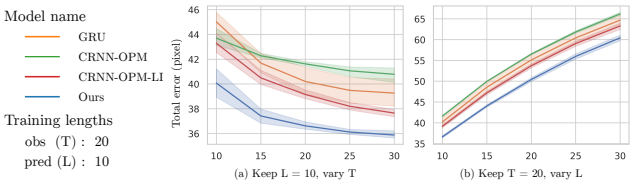


Figure 5: Generalization analysis on Bimanual Action Dataset

of human and objects. As shown in Fig. 5, PTD consistently outperforms other baselines in both generalization scenarios.

9. Additional application: Trajectory prediction

The PTD modeling is generic and applicable to a wide range of human motion modeling tasks. We demonstrate this universality by trialing it on Pedestrian Trajectory (PedTraj) prediction. For this purpose we will only use the single-prediction setting. The extension to multi-prediction setting can be done and is outside the scope of this supp.

Adaptation to Trajectory Prediction. In PedTraj setting, all entities are of class human. Their features $\{x_i^t\}_{t=1}^{T}$ and predicted output $\{y_i^t\}_{t=T+1}^{T+L}$ are sequences of 2D pedestrian coordinates. The implementation modification includes:

1. For Persistent channel, we use a Goal-driven Trajectory Prediction Network (GTP) [10] that estimates the long-term intention of the pedestrian on selecting the destinations in the scene.

2. The Transient channel is similar to the one for HOI-M in Sec. 3.4, main paper, except that the geometrical distances in Eq. 4, main paper include both relative distances and directions between pedestrians. The egocentric transformation f_{ego} (Eq. 5, main paper) now includes both translation, rotation, and scaling.

Experiment setting. We use the popular ETH and UCY datasets[8, 5] preprocessed into world coordinates[3]. We follow the common settings of observing 8 time steps (2.4s) and predicting 12 steps (3.6s); one scene is left-out for testing and the remaining are used for training. We also follow scene-based methods [6, 10, 9] to filter out the data with unavailable scene images. The scene images are used to extract the representation of all possible destinations, similar to [10]. The compared models are retrained with these settings if they were trained differently.

As multi-prediction is out of scope of this paper, and to concentrate only on motion modeling, we take the simple deterministic settings where only one prediction is generated. We compare PTD with the SoTA deterministic predictors, namely GTP [10] and SR-LSTM[13]. Extending PTD to multi-prediction and comparing it to variational methods is considered a future work.

Network parameters and Settings In PedTraj, we first embed the pedestrian locations into 8-dimensional vectors. Following HOI-M, we also use GRUs as RNN units and MLPs as readout functions in this application. In Persistent Channel, the GRUs of the pedestrians have the hidden states of 16 dimensions. The dimensions of the center GRU, the leaf GRU in Transient Channel, and the GRU in Switch Module are 32, 16, and 8, respectively. We embed the cross-channel messages between Persistent Channel and Transient Channel into 16-dimensional vectors. All models are trained with the batch size of 64 and optimized by the Adam optimizer [4] with a learning rate of 0.001.

The system and environment for all experiments are similar to those in HOI-M. We train the models for 300 epochs. The experiments are managed and recorded using Hydra [12] and Weight-and-Biases frameworks [1]. The full implementation will be made publicly available upon acceptance.

Quantitative Evaluation. The means and standard deviations of prediction errors on 5 independent runs are reported in Tab. 4. The duality of persistent and transient processes once again shows its power in modeling the interleaving mechanisms corresponding to the overall plan and the interruptive social interactions.

Visual analysis. We visualize the operations of the full PTD model and GTP in Fig. 6. The visualization shows that the pair of Persistent/ Transient channels and the Transient switch operate in synergy to successfully handle the interaction, avoid the potential collision and maintain the overall navigation plan.

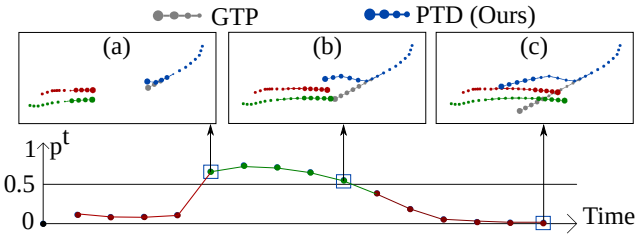


Figure 6: PTD on PedTraj Prediction. Along the course of the blue pedestrian, the Transient switch anticipates a collision (a), the switching score p^t increases, activating the Transient channel to handle the social interaction which steers to avoid the collision (b). When the interaction is over, p^t goes down, terminates the transient channel, and returns the control to the Persistent channel which guides the trajectory back to the original plan (c). In contrast, the baseline GTP model [10] (grey trajectory) does not have this mode-switching and failed to avoid the collision.

References

[1] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com. 5

[2] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 1, 3

[3] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 5

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2, 5

[5] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 5

[6] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2019. 5

[7] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015. 1

[8] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social be-

Prediction time(s)	0.6	1.2	1.8	2.4	3.0	3.6 (FDE)	Average (ADE)
GTP	$\mathbf{0.11 \pm 0.00}$	0.27 ± 0.00	0.44 ± 0.01	0.62 ± 0.01	0.84 ± 0.01	1.07 ± 0.01	0.51 ± 0.01
SR-LSTM	0.12 ± 0.01	0.26 ± 0.02	0.43 ± 0.04	0.64 ± 0.05	0.88 ± 0.08	1.17 ± 0.11	0.53 ± 0.04
PTD (Ours)	$\mathbf{0.11 \pm 0.00}$	$\mathbf{0.26 \pm 0.01}$	$\mathbf{0.42 \pm 0.01}$	$\mathbf{0.59 \pm 0.01}$	$\mathbf{0.78 \pm 0.01}$	$\mathbf{0.98 \pm 0.02}$	$\mathbf{0.48 \pm 0.01}$

Table 4: Prediction errors in PedTraj

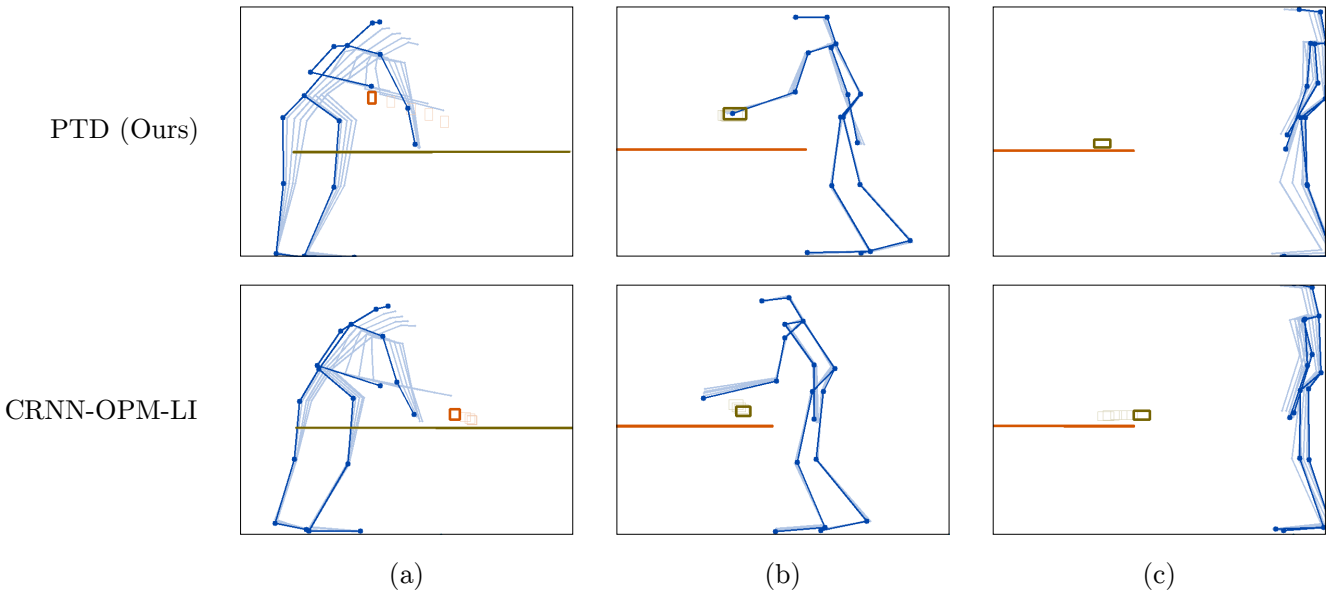


Figure 7: Visual Comparison in WBHM Dataset

havior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 5

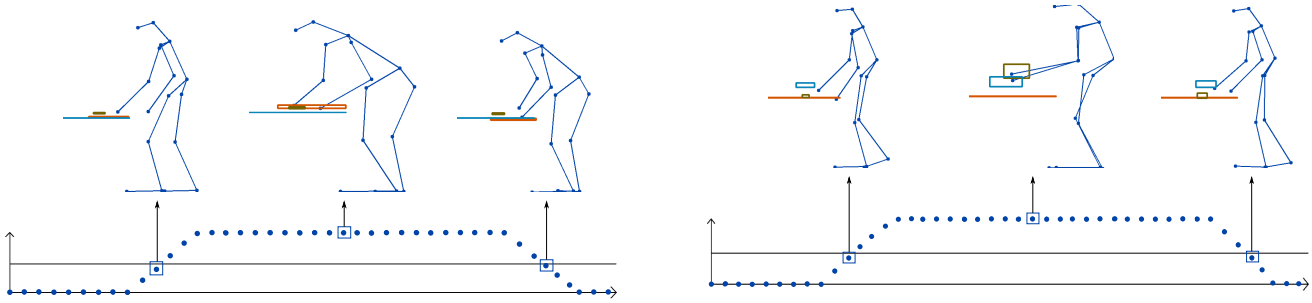
[9] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019. 5

[10] Hung Tran, Vuong Le, and Truyen Tran. Goal-driven long-term trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 796–805, 2021. 4, 5

[11] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019. 2

[12] Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. 2, 5

[13] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019. 5



(a) (b)

Figure 8: Additional visualizations of the Switching Behavior in WBHM Dataset

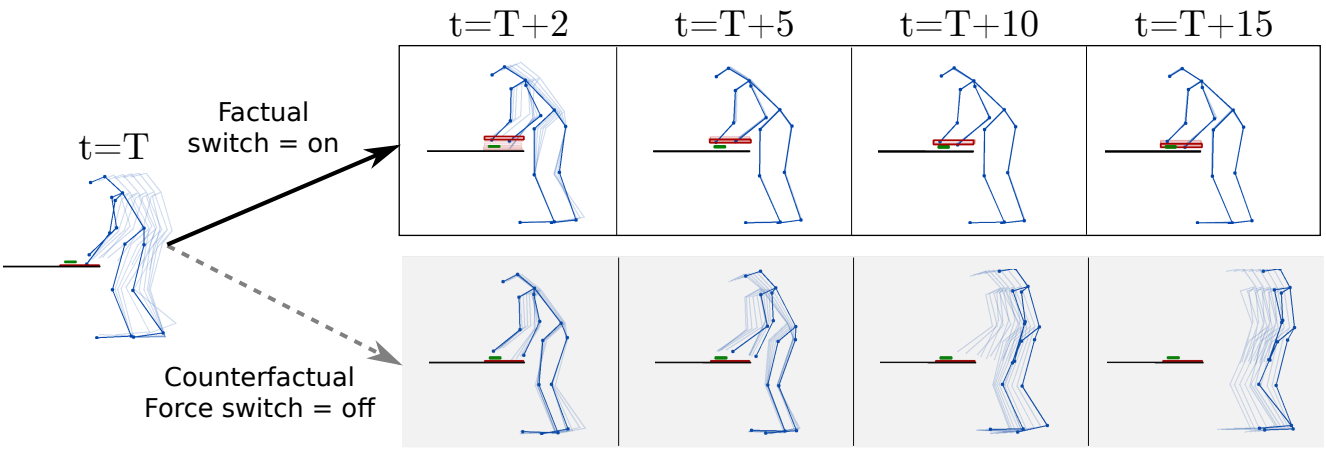
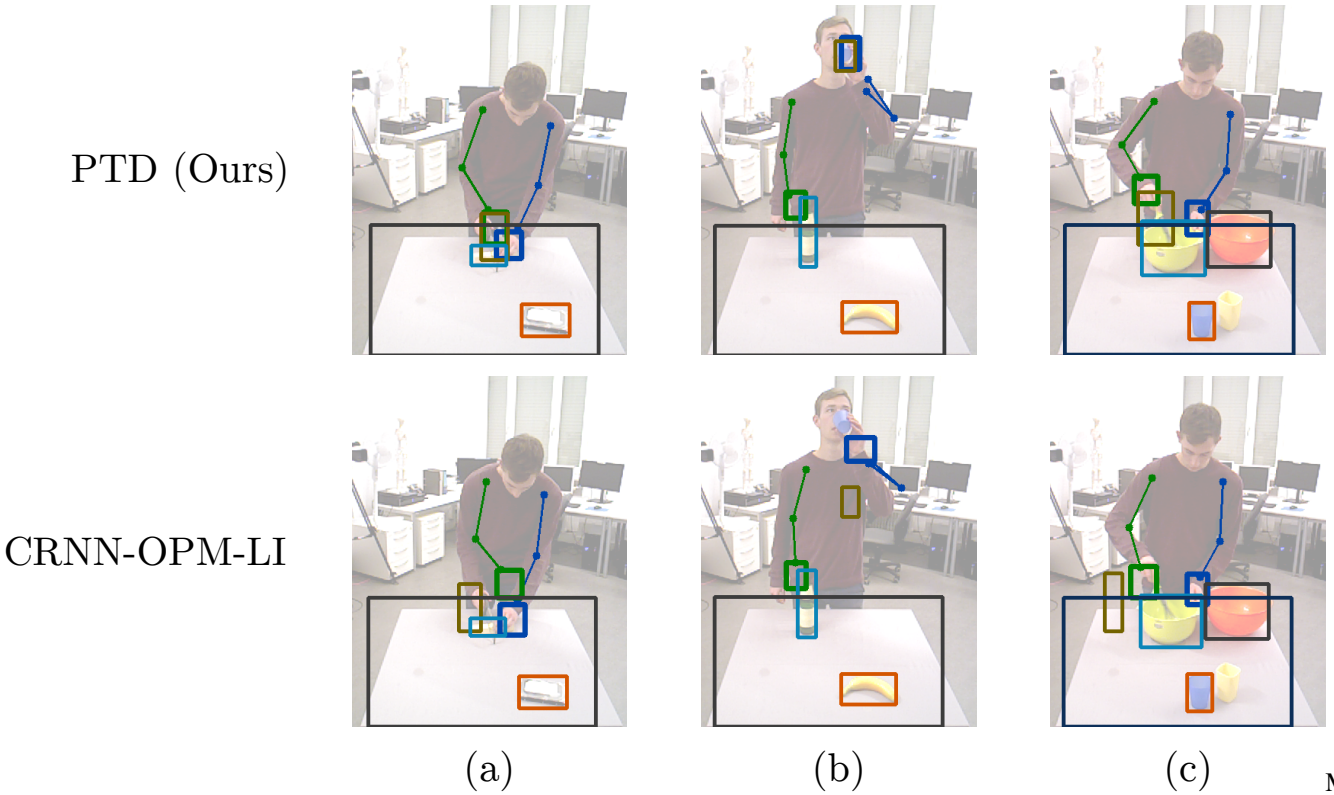


Figure 9: Controllability of the forecasting. We force the Transient switch to not turn on where it naturally does. The visualization shows that the model seems to accept this intervention, and reasonably forecast the motion as if the agent changes their mind



March 15, 2023 (23:59 GMT)

Figure 10: Visual Comparison in Bimaual Action Dataset

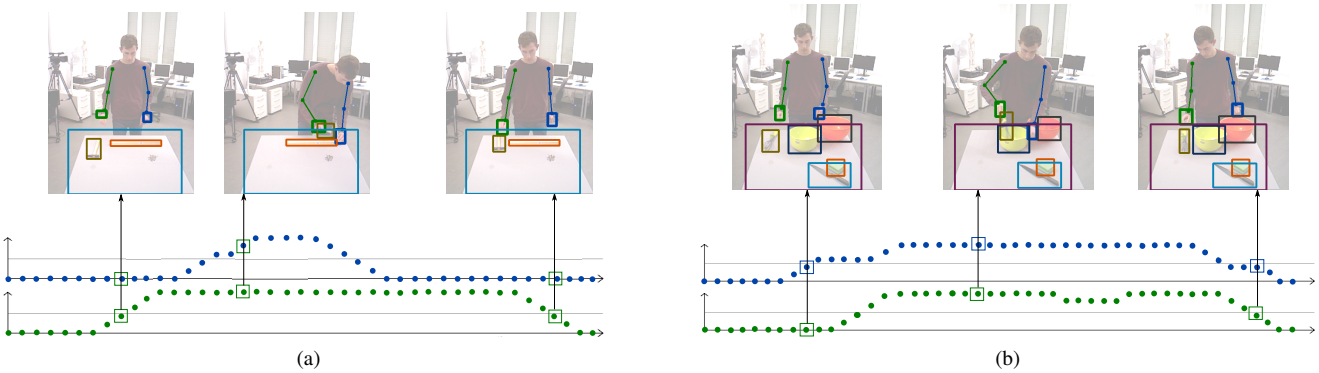


Figure 11: Additional visualizations of the Switching Behavior in Bimanual Action Dataset

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
zv	22	43	64	83	102	120	137	153	168	182
running avg. 2	32	53	73	92	111	128	145	161	176	190
GRU	21.4 ± 0.5	31.2 ± 0.4	40.6 ± 0.5	48.6 ± 0.5	56.8 ± 0.4	64.4 ± 0.8	71.8 ± 0.4	79.6 ± 0.8	87.0 ± 0.6	94.2 ± 0.7
CRNN-OPM	33.4 ± 1.0	37.0 ± 0.9	47.0 ± 0.9	54.6 ± 1.2	61.8 ± 1.2	69.4 ± 1.0	76.2 ± 1.2	83.4 ± 1.4	90.4 ± 1.4	97.4 ± 1.4
CRNN-OPM-LI	23.2 ± 0.4	32.6 ± 0.5	41.4 ± 0.5	49.4 ± 0.5	57.4 ± 0.5	64.6 ± 0.8	72.0 ± 1.1	79.0 ± 1.1	86.2 ± 1.5	93.6 ± 1.9
STS-GCN	47.9 ± 4.9	48.2 ± 4.5	52.5 ± 3.6	59.9 ± 2.7	66.6 ± 1.9	74.5 ± 1.9	81.9 ± 1.7	88.9 ± 1.6	95.0 ± 1.6	100.8 ± 1.7
MotionMixer	12.8 ± 0.5	23.2 ± 0.6	31.8 ± 0.7	39.8 ± 0.9	47.2 ± 1.0	55.0 ± 1.1	63.0 ± 1.2	70.8 ± 1.1	78.4 ± 0.9	85.4 ± 1.1
Ours	15.0 ± 0.0	26.0 ± 0.0	35.8 ± 0.4	44.6 ± 0.5	52.4 ± 0.5	59.8 ± 0.7	66.8 ± 0.7	73.4 ± 1.0	79.6 ± 1.0	85.8 ± 0.7

(a) From 0.1s to 1.0s

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
zv	196	208	220	232	242	253	262	272	281	289
running avg. 2	203	215	227	238	249	259	268	278	286	295
GRU	101.2 ± 0.7	108.8 ± 0.7	116.0 ± 1.1	123.8 ± 1.3	131.0 ± 1.1	138.4 ± 1.4	145.4 ± 1.9	152.0 ± 2.1	158.0 ± 2.6	163.8 ± 2.8
CRNN-OPM	104.0 ± 1.3	110.6 ± 1.4	117.4 ± 1.4	124.0 ± 1.1	130.4 ± 1.4	136.8 ± 1.2	142.8 ± 1.9	148.8 ± 1.9	154.6 ± 2.4	160.2 ± 2.7
CRNN-OPM-LI	100.4 ± 2.0	107.8 ± 2.4	115.2 ± 2.9	122.0 ± 3.0	129.2 ± 3.2	136.2 ± 3.2	143.2 ± 3.2	149.2 ± 3.1	155.6 ± 2.8	161.0 ± 2.9
STS-GCN	106.4 ± 1.6	111.7 ± 2.1	117.0 ± 2.1	122.3 ± 2.2	127.7 ± 2.8	133.5 ± 3.1	139.5 ± 3.4	145.2 ± 3.5	150.9 ± 3.8	157.0 ± 3.9
MotionMixer	92.3 ± 1.0	99.4 ± 1.0	106.5 ± 1.1	113.6 ± 1.3	120.5 ± 1.6	127.4 ± 1.7	134.3 ± 2.0	141.5 ± 2.4	148.5 ± 2.6	155.8 ± 2.8
Ours	91.6 ± 1.0	97.6 ± 1.0	103.2 ± 1.5	108.8 ± 1.3	114.8 ± 1.3	120.2 ± 1.7	125.8 ± 1.7	131.2 ± 1.5	136.6 ± 1.9	141.6 ± 2.0

(b) From 1.0s to 2.0s

Table 5: The mean and standard deviation of the human error at each time step in WBHM Dataset

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
zv	15	28	40	52	64	76	88	101	113	125
running avg. 2	21	33	45	57	68	80	93	105	118	130
GRU	14.0 ± 0.0	26.0 ± 0.0	37.4 ± 0.5	48.6 ± 0.5	59.6 ± 0.5	70.8 ± 1.0	82.4 ± 0.8	94.0 ± 1.1	105.6 ± 1.5	116.8 ± 1.5
CRNN-OPM	13.4 ± 0.5	24.6 ± 0.5	33.8 ± 0.7	42.8 ± 1.0	50.6 ± 1.2	58.6 ± 1.2	66.8 ± 1.5	74.8 ± 2.0	82.6 ± 2.2	89.6 ± 2.1
CRNN-OPM-LI	12.0 ± 0.6	21.6 ± 1.0	29.6 ± 1.4	37.2 ± 2.5	44.0 ± 3.2	50.8 ± 3.4	57.8 ± 4.0	64.8 ± 4.0	71.0 ± 4.0	77.0 ± 4.0
STS-GCN	-	-	-	-	-	-	-	-	-	-
MotionMixer	-	-	-	-	-	-	-	-	-	-
Ours	10.0 ± 0.0	19.4 ± 0.5	27.6 ± 0.5	35.4 ± 0.5	42.4 ± 0.5	49.2 ± 0.7	55.4 ± 0.5	61.4 ± 1.0	67.2 ± 0.7	72.8 ± 0.7

(a) From 0.1s to 1.0s

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
zv	137	148	160	171	183	193	204	214	225	235
running avg. 2	141	153	164	176	187	198	209	219	229	240
GRU	127.8 ± 1.7	138.6 ± 2.1	149.6 ± 2.1	160.2 ± 2.5	170.6 ± 2.7	181.0 ± 3.0	191.0 ± 3.3	200.8 ± 3.3	210.6 ± 3.6	220.6 ± 3.6
CRNN-OPM	96.2 ± 2.1	102.4 ± 2.1	108.4 ± 2.1	114.0 ± 1.9	119.6 ± 1.9	124.8 ± 1.9	129.6 ± 2.2	134.4 ± 2.4	139.0 ± 3.2	144.4 ± 3.5
CRNN-OPM-LI	82.4 ± 3.9	87.4 ± 4.3	92.4 ± 4.3	97.2 ± 4.3	101.4 ± 4.4	105.2 ± 4.1	108.6 ± 4.0	112.0 ± 4.4	115.0 ± 4.4	118.0 ± 4.4
STS-GCN	-	-	-	-	-	-	-	-	-	-
MotionMixer	-	-	-	-	-	-	-	-	-	-
Ours	77.8 ± 0.7	82.8 ± 0.7	87.0 ± 1.1	91.2 ± 1.0	95.6 ± 0.8	99.6 ± 0.8	103.8 ± 0.7	107.8 ± 0.7	111.6 ± 1.0	115.8 ± 1.2

(b) From 1.0s to 2.0s

Table 6: The mean and standard deviation of the object error at each time step in WBHM Dataset

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
zv	3.7	6.3	8.3	10	11.9	13.5	14.9	16.2	17.6	18.8
running avg. 2	4.9	7.2	9	10.8	12.6	14.1	15.4	16.8	18.1	19.2
GRU	3.9 ± 0.2	6.7 ± 0.2	8.9 ± 0.3	10.7 ± 0.4	12.4 ± 0.4	13.9 ± 0.4	15.1 ± 0.5	16.3 ± 0.5	17.4 ± 0.5	18.4 ± 0.5
CRNN-OPM	4.3 ± 0.2	7.1 ± 0.2	9.3 ± 0.2	11.1 ± 0.1	12.8 ± 0.1	14.3 ± 0.1	15.6 ± 0.2	16.8 ± 0.2	17.9 ± 0.3	19.0 ± 0.3
CRNN-OPM-LI	3.9 ± 0.1	6.6 ± 0.2	8.7 ± 0.2	10.4 ± 0.3	12.0 ± 0.3	13.5 ± 0.3	14.6 ± 0.4	15.7 ± 0.4	16.7 ± 0.4	17.7 ± 0.4
STS-GCN	5.0 ± 0.8	6.8 ± 0.5	8.7 ± 0.5	10.2 ± 0.5	11.8 ± 0.3	13.1 ± 0.4	14.3 ± 0.4	15.2 ± 0.5	16.2 ± 0.4	17.2 ± 0.4
MotionMixer	3.0 ± 0.1	5.5 ± 0.1	7.7 ± 0.1	9.9 ± 0.2	11.7 ± 0.2	13.3 ± 0.2	14.7 ± 0.2	15.9 ± 0.2	17.1 ± 0.2	18.1 ± 0.2
Ours	3.7 ± 0.1	6.2 ± 0.1	8.0 ± 0.1	9.5 ± 0.2	11.0 ± 0.2	12.2 ± 0.2	13.3 ± 0.2	14.2 ± 0.2	15.2 ± 0.2	16.1 ± 0.2

(a) Keypoints

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
zv	8.6	12.9	15.9	18.4	21.1	23.7	25.7	27.6	29.7	31.5
running avg. 2	10.2	13.9	16.6	19.2	21.8	24.2	26.2	28.2	30.1	31.7
GRU	8.6 ± 0.3	12.9 ± 0.5	16.0 ± 0.7	18.5 ± 0.9	21.0 ± 0.9	23.1 ± 1.0	24.7 ± 1.1	26.3 ± 1.1	27.8 ± 1.1	29.1 ± 1.1
CRNN-OPM	9.7 ± 0.8	13.8 ± 0.6	16.8 ± 0.5	19.4 ± 0.4	21.7 ± 0.2	23.8 ± 0.1	25.5 ± 0.2	27.1 ± 0.2	28.7 ± 0.2	30.1 ± 0.3
CRNN-OPM-LI	8.6 ± 0.1	12.8 ± 0.2	15.7 ± 0.2	18.1 ± 0.3	20.4 ± 0.2	22.3 ± 0.2	23.8 ± 0.3	25.2 ± 0.3	26.6 ± 0.3	27.8 ± 0.3
STS-GCN	-	-	-	-	-	-	-	-	-	-
MotionMixer	-	-	-	-	-	-	-	-	-	-
Ours	8.4 ± 0.1	12.4 ± 0.1	14.9 ± 0.2	17.0 ± 0.2	19.0 ± 0.2	20.8 ± 0.2	22.1 ± 0.2	23.3 ± 0.2	24.6 ± 0.2	25.8 ± 0.3

(b) Hand boxes

Table 7: The mean and standard deviation of the human errors at each time step in Bimanual Action Dataset

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
zv	2.7	4.3	5.1	5.8	6.8	7.7	8.4	9.1	9.8	10.5
running avg. 2	3.3	4.6	5.3	6.1	7.1	7.9	8.6	9.4	10.1	10.7
GRU	2.7 ± 0.0	4.3 ± 0.0	5.1 ± 0.0	5.9 ± 0.0	6.8 ± 0.0	7.7 ± 0.0	8.4 ± 0.0	9.1 ± 0.0	9.9 ± 0.0	10.5 ± 0.0
CRNN-OPM	2.7 ± 0.0	4.3 ± 0.0	5.2 ± 0.1	6.0 ± 0.1	6.9 ± 0.1	7.8 ± 0.1	8.5 ± 0.1	9.3 ± 0.1	10.0 ± 0.2	10.7 ± 0.2
CRNN-OPM-LI	2.7 ± 0.0	4.3 ± 0.0	5.1 ± 0.0	5.9 ± 0.0	6.8 ± 0.1	7.7 ± 0.1	8.4 ± 0.1	9.2 ± 0.1	9.9 ± 0.1	10.6 ± 0.1
STS-GCN	-	-	-	-	-	-	-	-	-	-
MotionMixer	-	-	-	-	-	-	-	-	-	-
Ours	2.7 ± 0.0	4.2 ± 0.0	5.1 ± 0.1	5.9 ± 0.1	6.7 ± 0.1	7.5 ± 0.1	8.1 ± 0.1	8.8 ± 0.1	9.4 ± 0.1	10.0 ± 0.1

Table 8: The mean and standard deviation of the object error at each time step in Bimanual Action Dataset