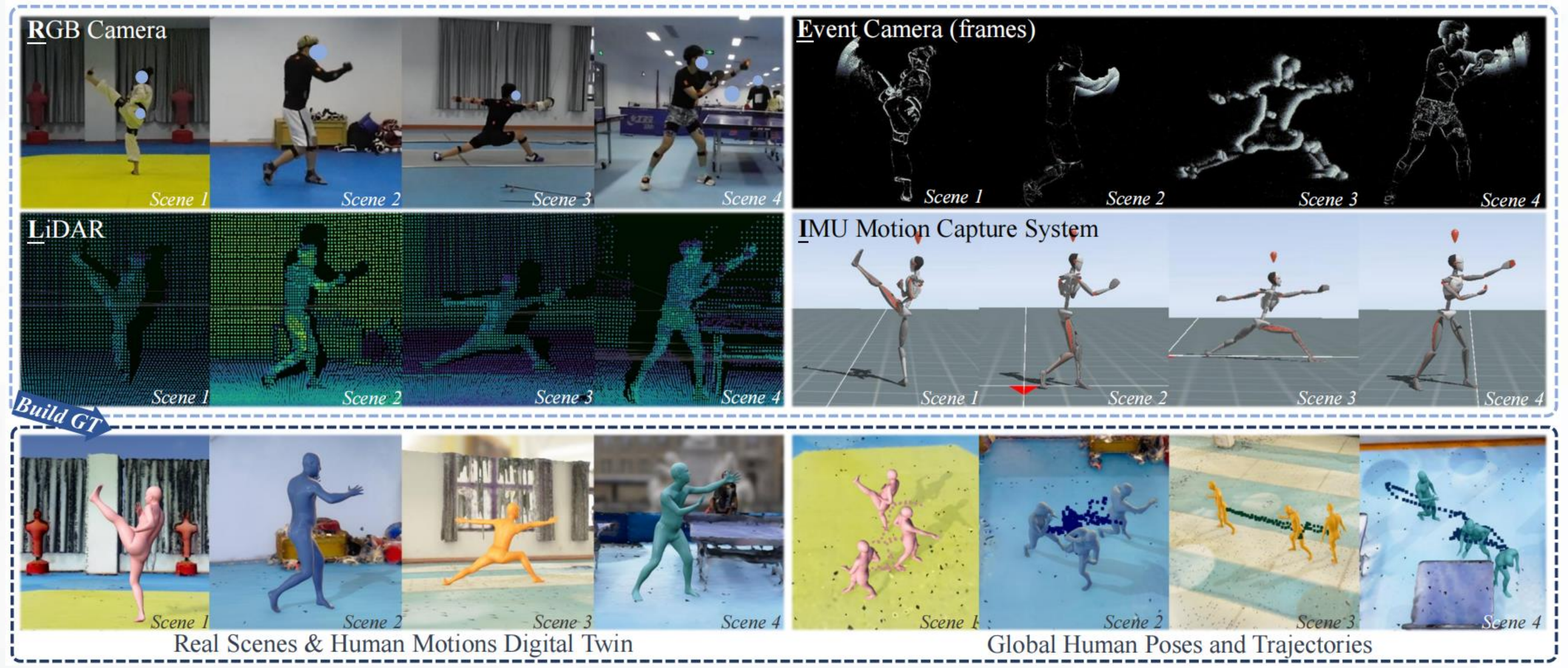
SEATTLE, WA CVPR JUNE 17-21, 2024

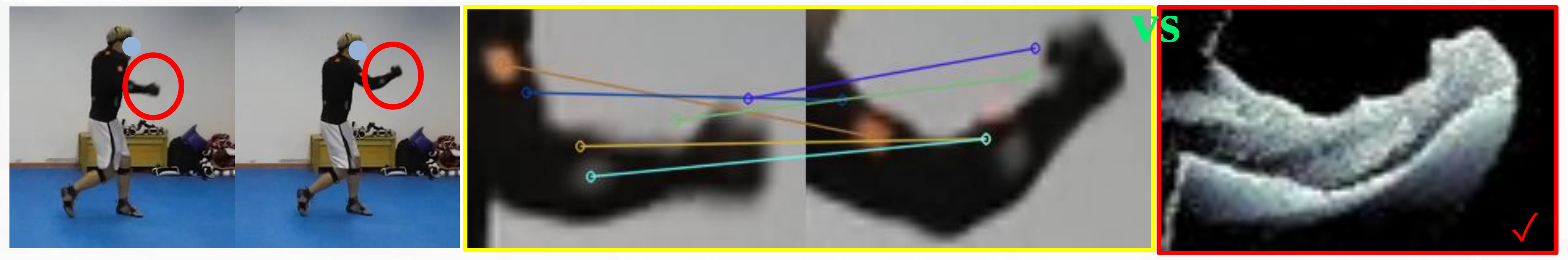
Session Order: 201, Poster Session 1, Wednesday, June 19, 10:30 - 12:00
SEATTLE CONVENTION CENTER, SEATTLE, WASHINGTON USA

RELI11D: A Comprehensive Multimodal Human Motion Dataset and Method

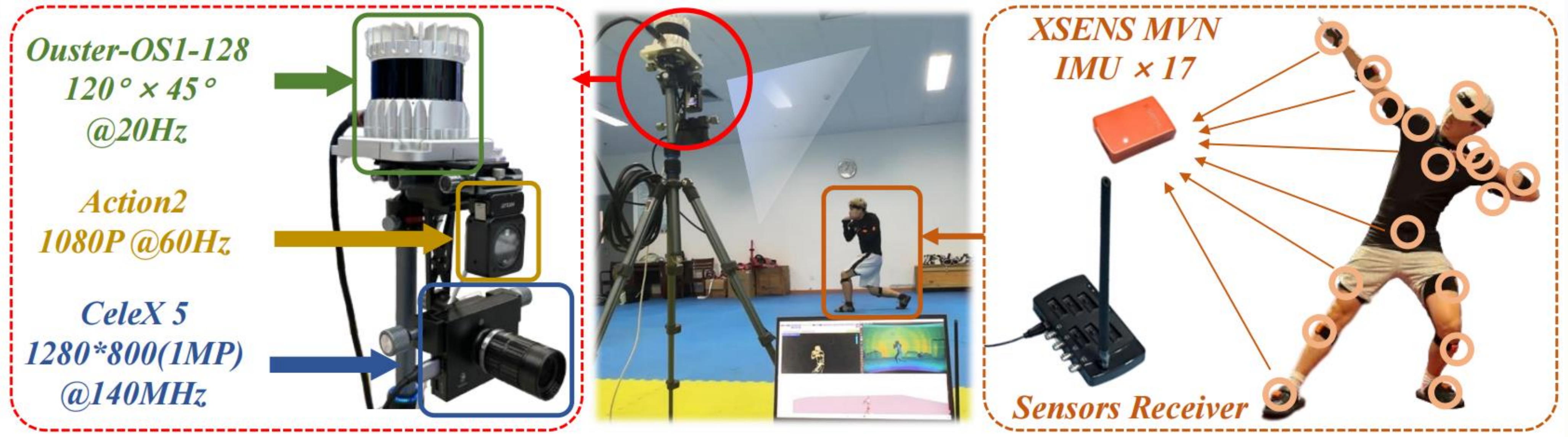
Ming Yan[1]*, Yan Zhang[1]*, Shuqiang Cai[1], Shuqi Fan[1], Xincheng Lin[1], Yudi Dai[1], Siqi Shen[1]†, Chenglu Wen[1], Lan Xu[2], Yuexin Ma[2], Cheng Wang[1]

[1]*Xiamen University* [2]*ShanghaiTech University*

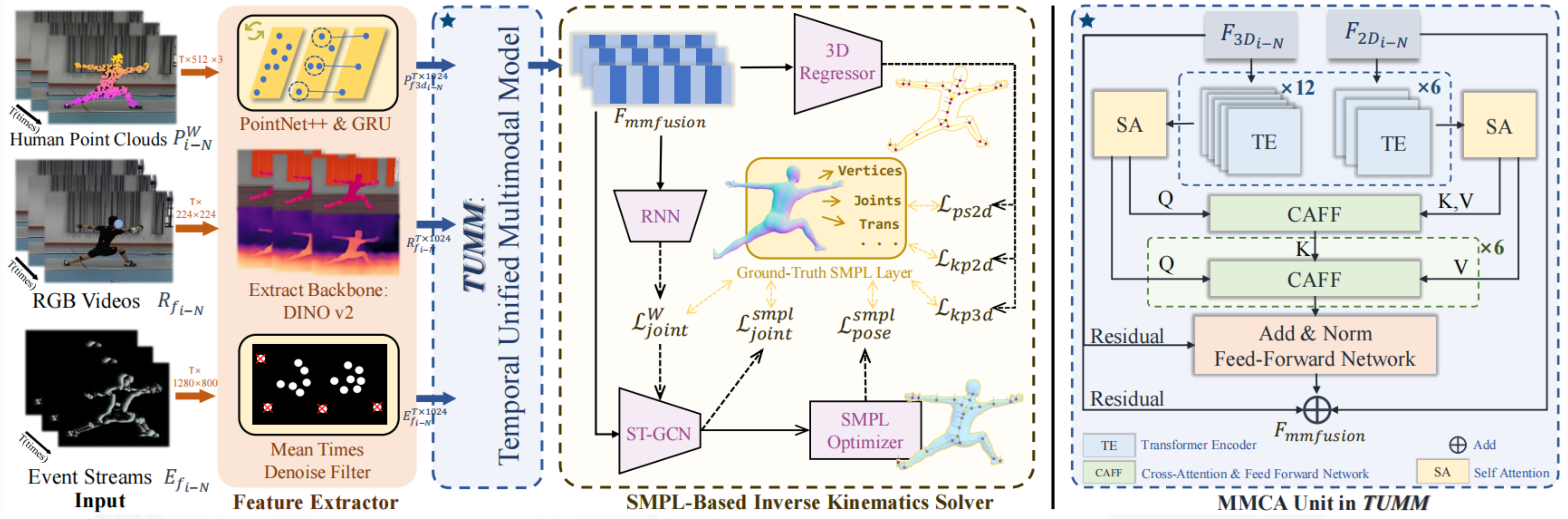

RGB Camera
Event Camera (frames)
LiDAR
IMU Motion Capture System
Scene 1
Scene 2
Scene 3
Scene 4
Build GT
Real Scenes & Human Motions Digital Twin
Global Human Poses and Trajectories

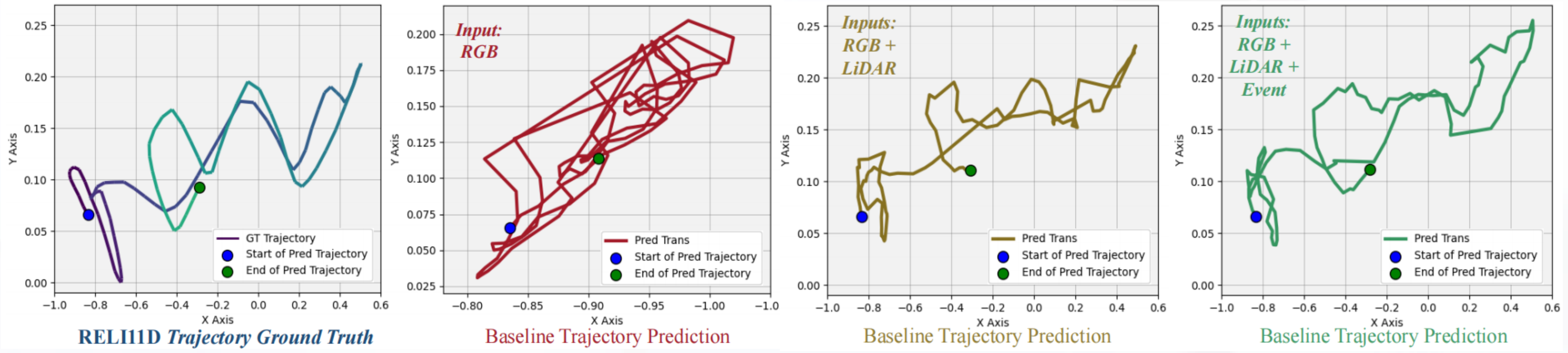


Fast and complex movements can be seen everywhere in reality. However, it is difficult to fully capture human body movements. It requires both accurately capturing complex postures and accurately positioning the human body in the scene. The community needs a high-quality dataset and method to fill the gap.



For holistic human motion understanding, we present RELI11D. It records the motions of 10 actors performing 5 sports in 7 scenes, including 3.32 hours of synchronized four different modality and provides precise annotations. We built a portable collection system that integrates different modal hardware devices to collect in different real-world scenarios. And we propose a Consolidated Optimization, which includes global Pose Loss, human joint Loss, scene awareness Loss, et al.

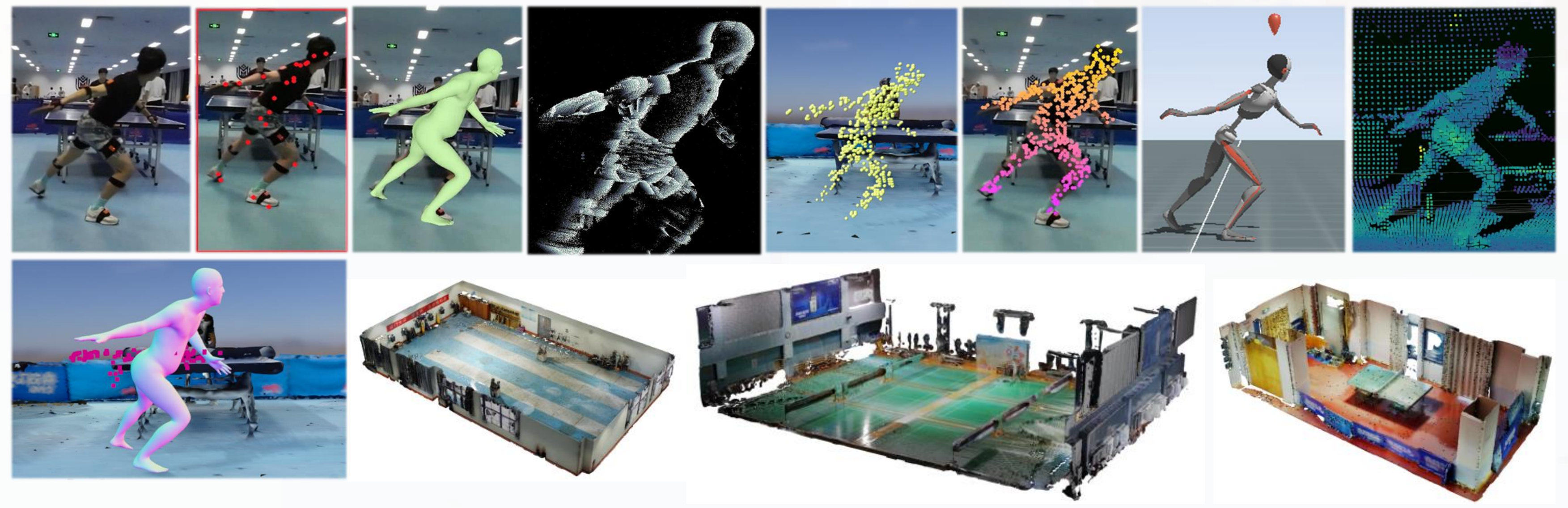



RELI11D *Trajectory Ground Truth*
Baseline Trajectory Prediction
Baseline Trajectory Prediction
Baseline Trajectory Prediction

In order to effectively integrate the data of each modality and use it for global human pose estimation, we propose a Baseline, LEIR. First, input data from different modalities. Next, we use the feature extractor corresponding to different modalities to obtain features. The modal features enter the Temporal Unified Multimodal Model, and we propose the MMCA Unit to fuse the features. Finally, in the SMPL-Based Inverse Kinematics Solver, we design a variety of Loss to constrain data of different dimensions and finally predict the global human poses. In method experimental part, First, we present with different inputs of LEIR, and the model index of three-modal input was the best. This result confirms the importance of multi-modal methods in human pose estimation. In addition, we conduct cross-dataset validation, and our method also performs better on other datasets. Furthermore, we perform visualization experiments on the prediction results of global trajectories. Combine with the analysis of the previous quantitative experimental results, it is show that the multi-modal method can be of considerable help in global human pose estimation.

Affiliation: Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China. National Institute for Data Science in Health and Medicine, Xiamen University, China. Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China. Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University, China

Project Page: ***http://www.lidarhumanmotion.net/reli11d/***



Input Modality	ACCEL↓	MPJPE↓	PA-MPJPE↓	G-MPJPE↓	T-Error↓	PCK0.3↑
LiDAR	31.26	59.93	48.23	125.77	195.72	0.89
RGB	28.43	62.71	54.11	557.81	710.84	0.88
Event	34.45	107.78	83.64	605.45	743.71	0.59
LiDAR+ RGB	27.07	55.36	45.72	122.32	168.61	0.90
LiDAR+ Event	25.41	57.79	46.70	123.75	178.97	0.89
LiDAR+ RGB+Event	**23.90**	**49.19**	**40.87**	**115.36**	**146.13**	**0.92**

Input Modality	Method	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
LiDAR	P4Tranformer [24]	66.33	172.04	150.65	206.75	0.51
	PCT [27]	59.19	144.40	116.99	174.33	0.67
	LiDARCap [48]	54.42	144.51	106.20	176.98	0.67
RGB	HybrIK [47]	58.39	249.34	163.91	255.98	0.53
	NIKI [45]	55.62	196.68	142.48	198.10	0.61
	SMPLer-X [9]	50.15	171.97	128.02	185.83	0.66
Event	EventHPE [98]	-	193.7	115.72	224.59	0.52
	EventPointPose [14]	-	16.24(2D)	10.91(2D)	-	0.69(2D)
RGB+ LiDAR	ImmFusion [12]	49.19	175.00	159.62	187.31	0.67
	FusionPose [19]	44.89	136.15	110.19	166.94	0.75

In the experimental part, we first evaluate on the dataset RELI11D. In the qualitative experiment, we show the different stages of data set annotation, and it can be seen that our annotation has the results closest to real actions. In the Benchmarks experiment, we compare the human pose estimation method based on 2D video and global human pose estimation. It can be seen from the visualization experiments that existing methods cannot estimate fast-moving limb movements well, and almost all methods cannot estimate high-leg movements. The dual-modality based method FusionPose shows the best local human pose estimation results. Current methods all perform poorly in global human pose estimation.