RELI11D: A Comprehensive Multimodal Human Motion Dataset and Method —Supplementary Material

Ming Yan[1,2,3*] Yan Zhang[1,3*] Shuqiang Cai[1,3] Shuqi Fan[1,3] Xincheng Lin[1,3] Yudi Dai[1,3]
Siqi Shen[1,3†] Chenglu Wen[1,3] Lan Xu[4] Yuexin Ma[4] Cheng Wang[1,3]
[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University
[2]National Institute for Data Science in Health and Medicine, Xiamen University
[3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, School of Informatics, Xiamen University
[4]Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University

We would like to thank the reviewer for reading the supplementary material.

In Appendix A, we describe the performance of state-of-the-art methods after training from scratch based on the RELI11D dataset. Then, We evaluate the generalization ability of LEIR in Appendix B, add additional trajectory legend in Appendix C, and perform an ablation study of the baseline in Appendix D. Appendix E presents the cross-dataset evaluation result of RELI11D.

For evaluation, we report Procrustes-Aligned mean per-joint position error (PMPJPE), mean per-joint position error (MPJPE), percent correct keypoints (PCK) , Per Vertex Error (PVE) and Acceleration Error (mm/s^2) (ACCEL). Except for PCK, which is a percentage indicator, error indicators are all in millimeters.

We describe the details of the RELI11D dataset in Appendix F. And in Appendix G, we present the detail of the multi-modal benchmark. We further describe the details of the multi-modal baseline LEIR in Appendix H, and the details of different modality fusion strategies in Appendix I.

A. HPE results in RELI11D

To study the performance of multiple state-of-the-art HPE methods on RELI11D. We train these methods from scratch on RELI11D and compare their performance with their variations without retraining. The studied methods are: RGB-based method (HybrIK [13], NIKI [12]), global-RGB-based method (GLAMR [22]), Event-based method (EventHPE [23]), LiDAR-based method (LiDARCap [14]), and RGB+LiDAR-based method (ImmFusion [2] and FusionPose [4]). Please refer to Appendix G for their detailed descriptions.

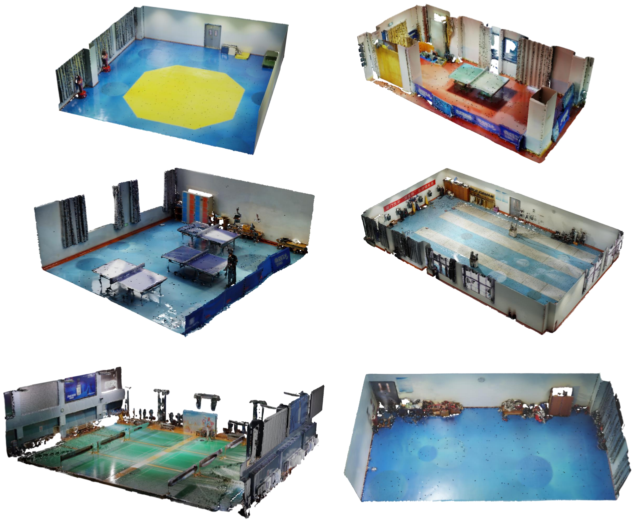


Figure 1. Scenes in RELI11D dataset.

Their results are shown in Tab. 1. RELI11D dataset consists of many rapid and complex movements. All these methods perform poorly on RELI11D without retraining (results without the * mark). After retraining, the performance of these methods improves a lot. For example, after retraining, HybrIK [13]'s MPJPE indicator results improve by about 40%. The LiDAR-based method LiDARCap improves the index by 55% after retraining. These demonstrate that RELI11D brings a significant challenge to existing methods. After incorporating the knowledge of challenging poses from RELI11D, the performance of these methods improves. Moreover, we find that the proposed baseline, LEIR, performs the best among all the methods thanks to its ability to effectively utilize the RGB, LiDAR, and Event modalities.

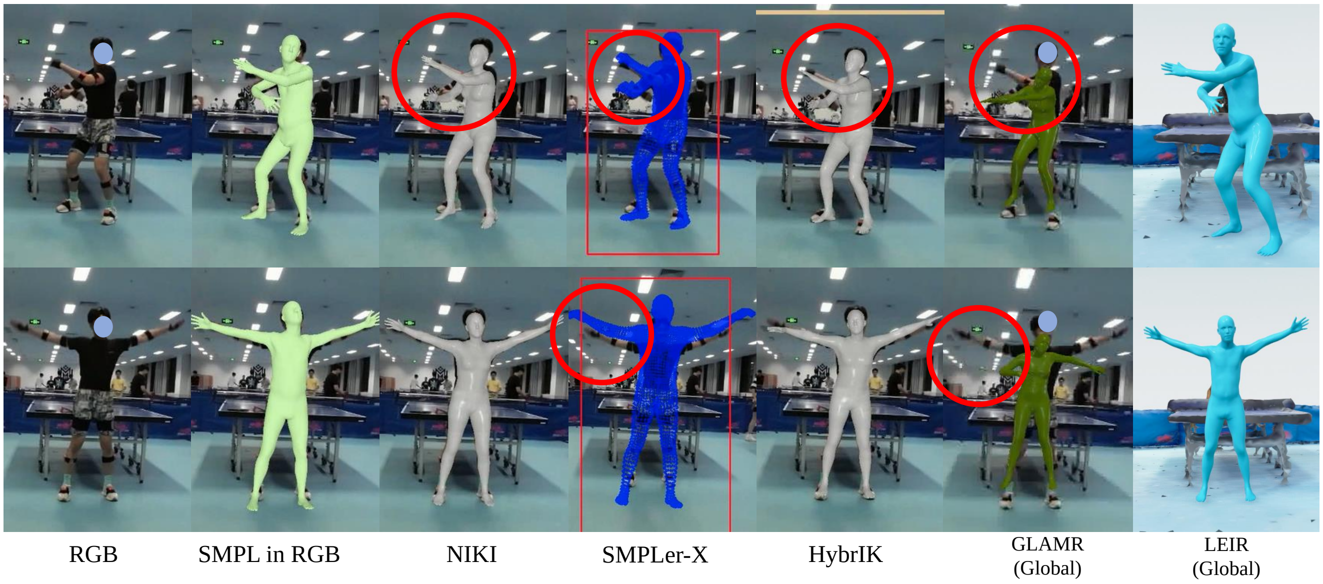


Figure 2. Generalization experiment on unseen rapid motion with complicated hand depth ambiguity. Red circles mark areas with obvious errors. The proposed baseline, LEIR, can accurately model the motions.

Input Modality	Method	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
RGB	HybrIK [13]	58.39	249.34	163.91	255.98	0.53
	HybrIK* [13]	38.37	141.13	132.18	216.43	0.71
	NIKI [12]	55.62	196.68	142.48	198.10	0.61
	NIKI* [12]	43.28	120.36	112.63	165.71	0.73
Global RGB	GLAMR [22]	47.83	202.66	179.59	346.35	0.65
	GLAMR* [22]	44.15	163.90	155.84	277.13	0.70
Event	EventHPE [23]	-	193.70	115.72	224.59	0.52
	EventHPE* [23]	-	167.65	109.57	202.46	0.56
LiDAR	LiDARCap [14]	54.42	144.51	106.20	176.98	0.67
	LiDARCap* [14]	36.85	64.43	52.19	75.92	0.84
RGB+ LiDAR	ImmFusion [2]	49.19	175.00	159.62	187.31	0.67
	ImmFusion* [2]	48.74	123.08	103.16	154.81	0.72
	FusionPose [4]	44.89	136.15	110.19	166.94	0.75
	FusionPose* [4]	42.29	97.58	74.51	106.31	0.81
LiDAR+ RGB+Event	LEIR	**23.90**	**49.19**	**40.87**	**61.86**	**0.92**

Table 1. Retrain SOTA HPE methods in RELI11D. *Represents this method is retrained from scratch based on RELI11D.

B. Generalization to Unseen Motions

In this work, we have collected a small set of fast motions with complicated hand-depth ambiguity. Actors perform various rapid and unpredictable hand movements, and such motions are not seen in the RELI11D dataset. We use these motions to test the generalization ability of the proposed baseline and other state-of-the-art methods. The visualization results are shown in Fig. 2. Most methods cannot accurately reconstruct the details of these movements. It can be seen from the results that there exist some in-precise predicted poses for NIKI [12], SMPler-X [1], HybriK [13], and GLAMR [22]. LEIR can capture these motions accurately thanks to its ability to use RGB, LiDAR point clouds, and events streams together.

Strategy	ACCEL↓	MPJPE↓	PA-MPJPE↓	PVE↓	PCK0.3↑
(a)ImmFusion-Based	34.94	95.27	70.24	141.67	0.79
(b)FusionPose-Based	35.27	59.50	51.08	72.10	0.88
(c)MMCA w/o Multi-TE	36.11	58.37	49.42	70.75	0.89
(d)MMCA w/o Multi-CA&TE	31.59	61.79	51.32	75.50	0.87
(e)MMCA $with$ J & V	**26.69**	56.15	47.09	69.03	0.89
MMCA(Ours)	27.07	**55.36**	**45.72**	**67.74**	**0.90**

Table 2. Ablation study of LEIR using different fusion strategies.

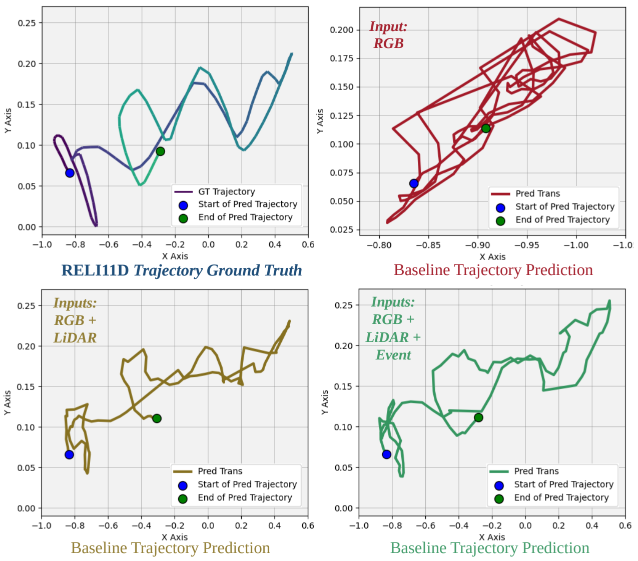
C. Global Trajectory Prediction

As mentioned in Sec 5.5 of the main paper, the T-Error measures the translation error is depicted in Tab.6. And the additional predicted trajectory is plotted in Fig. 3. It shows that the method based on two-dimensional representation performs very poorly on global trajectory indicators, which is also the bottleneck of the current global HPE method based on monocular cameras. Incorporating the LiDAR point clouds with global trajectory information improves the global motion indicators (low T-Error, similarity between the curve and ground truth). This observation indicates a promising trend of multimodal methods that fuse global information.

D. LEIR Ablation Study

In this section, we perform the ablation study for the multi-modal baseline, LEIR. As it is a multi-modal method, we consider different ways to fuse different modalities. Fig. 4 depicts different modal fusion strategies, and Appendix I describes them in details.

The experimental results of these methods are summarized in Tab. 2. Our baseline uses MMCA to fuse different modalities. It is observed that, in terms of ACCEL, MMCA



RELI11D *Trajectory Ground Truth*

Baseline Trajectory Prediction

Baseline Trajectory Prediction

Baseline Trajectory Prediction

Figure 3. **Qualitative experiments on trajectory comparison in different modalities.** The lower right corner is the trajectory prediction result of the three-modal baseline network, which is better than other modalities. Unit: m.

Metrix	Train / Test	LH26M	RELI11D	LH26M+RELI11D
ACCEL	LH26M	45.88	130.92	51.30
	RELI11D	54.42	36.85	29.54
MPJPE	LH26M	80.08	237.45	94.73
	RELI11D	144.51	64.43	55.20
PMPJPE	LH26M	67.51	149.89	76.80
	RELI11D	106.20	52.19	46.22
PVE	LH26M	102.24	289.11	109.25
	RELI11D	176.98	75.92	65.38
PCK0.3	LH26M	0.85	0.57	0.83
	RELI11D	0.65	0.84	0.90

Table 3. Cross-dataset evaluation using LiDARCap [14].

performs slightly worse than MMCA + J&V, which requires additional inputs. However, in terms of the MPJPE, PA-MPJPE, PVE, and PCK0.3, MMCA performs better than all the other strategies.

E. Cross Dataset Evaluation

In this section, we evaluate the quality of RELI11D through cross-dataset evaluation. We have shown in the submitted paper that for single-modality input, the LiDAR-based method (i.e., LiDARCap) performs better than other-modality-based methods. Therefore, we use LiDARCap as the studied method to evaluate the quality of RELI11D.

LiDARCap is trained based on the training sets of LH26M (LiDARHuman26M [14]), RELI11D, and their combination (LH26M+RELI11D). Then, it is evaluated based on the test sets of LH26M and RELI11D, respectively. The results are shown in Tab. 3.

As shown in Tab. 3, we can draw the following conclusions. When LiDARCap is trained on one of the datasets, it performs poorly on the other. In particular, when it is trained on LH26M and tested on RELI11D, its performance is worse than the other way around, which shows that the daily activities included in LH26M are not enough for LiDARCap to estimate rapid and complex human motion postures. Furthermore, if LiDARCap is trained based on the combination of the two datasets, its performance on both datasets is significantly improved, and the error will be further reduced by about 60%. This indicates a domain gap between the LiDAR modality of the two datasets and that the two datasets *complement* each other.

F. RELI11D Dataset

F.1. Scene Surface Reconstruction

To study the interaction between the human body and the scene, we use Trimble X7, a high-precision scene scanning device, to record the scene. Each scene contains more than 40 million colored point clouds. For ease of geometry operations in 3D space, we convert all point clouds in dense models into mesh using Poisson reconstruction [9, 10]. In Fig. 1, we show 6 reconstructed scenes, where human motions were recorded.

F.2. Consolidated Optimization Stage

Coordinates and Notations. In this work, there are three coordinates systems: 1) IMU coordinate $\{I\}$: its origin is at the pelvis joint of a human, and $X/Y/Z$ axis is pointing to the right/upward/forward of the human. 2) LiDAR Coordinate $\{L\}$: its origin is the center of the LiDAR, and $X/Y/Z$ axis is pointing to the right/forward/upward of the LiDAR. 3) World coordinate $\{W\}$: it is the scene's coordinate. We use the subscript $k, k \in Z^+$ to represent frame index, and the superscript, I or L or W, to specify the frame's coordinate system. For example, a LiDAR point clouds frame is represented as $P^L = \{P^L_k, k \in Z^+\}$

A 3D scene is represented as $\boldsymbol{S}$. $M^W_k = (T^W_k, \theta^W_k, \beta)$ represents the kth frame of human motion in the world coordinate system, where T^W_k is the 3 dimensions translation parameter, θ^W_k is 24×3 dimensions pose parameters, β is 10 dimensions shape parameter. The global translation $T^W_k \in \mathbb{R}^3$ represents the position of the SMPL model in the three-dimensional space. The pose parameters θ^W_k are determined by the orientation of the pelvic joint $R^W_k \in \mathbb{R}^{1\times3}$ and the other 23 joints $\in \mathbb{R}^{23\times3}$ relative to its parent Rotation composition of level nodes. The constant parameter β in $\mathbb{R}^{10}$ represents the shape of the human body. We use the Skinned Multiplayer Linear (SMPL) [15] body model $\varPhi$ to obtain $V_k, F_k = \varPhi(M^W_k)$, where the body vertices

$V_k \in \mathbb{R}^{6890\times 3}$ and the faces $F_k \in \mathbb{R}^{13690\times 3}$.

F.2.1 Optimization Loss Detail

We utilize contact aware loss $\mathcal{L}_{contact}$, smoothness loss $\mathcal{L}_{smoo}$ and geometry loss $\mathcal{L}_{geo}$ to perform consolidated optimization of global poses and trajectories to obtain accurate and scene-natural human motion annotation. We minimize the overall loss which is defined as follows.

$$\mathcal{L} = \lambda_c \mathcal{L}_{contact} + \lambda_s \mathcal{L}_{smoo} + \lambda_g \mathcal{L}_{geo} \tag{1}$$

where $\lambda_c,\ \lambda_s,\ \lambda_g$ are loss coefficients. $\mathcal{L}$ is minimized with a gradient descent algorithm [11].

Contact Aware Loss. The $\mathcal{L}_{contact}$ term combines scene constraints $\mathcal{L}_{sceneC}$ and self-penetration constraints $\mathcal{L}_{selfC}$ to improve the quality of local human poses. The $\mathcal{L}_{contact}$ is expressed as:

$$\mathcal{L}_{contact} = \lambda_{sceneC}\mathcal{L}_{sceneC} + \lambda_{selfC}\mathcal{L}_{selfC} \tag{2}$$

where λ_{sceneC} and λ_{selfC} are coefficients of these loss terms.

$\mathcal{L}_{sceneC}$ penalizes the vertices in the human SMPL mesh that penetrates the scene mesh to ensure that the human mesh remains collision-free with the 3D scene during optimization. For each vertex v, we find the closest vertex S_v on the scene S. If the dot product between the distance vector from v to S_v and the normal vector at S_v is positive, it indicates penetration. $\mathcal{L}_{sceneC}$ is expressed as:

$$\mathcal{L}_{sceneC} = \frac{1}{k}\sum_{i=1}^{k}\sum_{\mathbf{v}\in M_i} max\left(0, (S_v - \mathbf{v})\cdot \mathbf{n}_{S_v}\right) \tag{3}$$

where $\mathbf{n}_{S_v}$ is the normalized normal vector of S_v.

$\mathcal{L}_{selfC}$ penalizes the self-penetration in human motion M_k^W. Following [19], we divide the body mesh M_k^W into 12 separate regions $\mathcal{R}$ including the torso, arms, hands, legs, and head. The $\mathcal{L}_{selfC}$ is expressed as:

$$\mathcal{L}_{selfC} = \frac{1}{k}\sum_{i=1}^{k}\sum_{A\in\mathcal{R}}\sum_{B\in\mathcal{R}\ B\neq A}\sum_{\mathbf{a}\in A}\max\left(0, (\mathbf{a}-\mathbf{b})\cdot\mathbf{n}_b\right) \tag{4}$$

where A and B are different regions of the body, and the vertices $\mathbf{a}$ and $\mathbf{b}$ belong to regions A and B, respectively. $\mathbf{n}_b$ represents the normalized normal vector at $\mathbf{b}$. Self-penetration is determined by the positive dot product between vectors $(\mathbf{a}-\mathbf{b})$ and $\mathbf{n}_b$.

Smoothness Loss. We use $\mathcal{L}_{smoo}$ to ensure the smoothness of the global human motion spatially and temporally. The $\mathcal{L}_{smoo}$ is expressed as:

$$\mathcal{L}_{smoo} = \lambda_{trans}\mathcal{L}_{trans} + \lambda_{poses}\mathcal{L}_{poses} + \lambda_{joints}\mathcal{L}_{joints} \tag{5}$$

where λ_{trans}, λ_{joints} and λ_{poses} are coefficients of these loss terms. The trajectory smoothing term $\mathcal{L}_{trans}$ smooths human trajectories by minimizing the acceleration of the pelvis. It is defined as:

$$\mathcal{L}_{trans} = \frac{1}{k-2}\sum_{i=1}^{k-2}\|T_{i+2} - 2T_{i+1} + T_i\|_2^2 \tag{6}$$

The body posture smoothing term $\mathcal{L}_{poses}$ maintains the stability of the entire human body motion by minimizing the axial angular velocity of each pelvis-related joint. It is expressed as:

$$\mathcal{L}_{pose} = \frac{1}{k-2}\sum_{i=1}^{k-2}\|\theta_{i+2} - \theta_{i+1} + \theta_i\|_2^2 \tag{7}$$

The human joints smoothing term $\mathcal{L}_{joints}$ promotes the smoothness of the overall motion by minimizing the acceleration of all the SMPL joints except the root joint. It is expressed as:

$$\mathcal{L}_{jts} = \frac{1}{k-2}\sum_{i=1}^{k-2}\|J(M^*_{i+2}) - 2J(M^*_{i+1}) + J(M^*_i)\|_2^2 \tag{8}$$

where the 23 pelvis-relative joints are regressed from the motions by $J(M^*_i) \in \mathbb{R}^{23\times 3}$.

Geometry Loss. The point cloud contains the geometry information of human movement, they can be used to guide the reconstructed SMPL model.

Registering the global human body SMPL and point clouds in three-dimensional space may improve the quality of reconstructed SMPL model. However, traditional registration methods such as Iterative Closest Point (ICP) are not be suitable for aligning sparse and partial body points with dense and complete SMPL grids. These methods usually rely on large overlapping areas and similar point set densities to obtain accurate results.

For each estimated human mesh, following [5], we use Hidden Point Removal (HPR) [8] to remove invisible mesh vertices from the LiDAR perspective. We then use iterative closest point (ICP) [18] to register the visible vertices to $\mathcal{P}$, which is the segmented human body point cloud. We project the body mesh in LiDAR coordinates to select the visible body vertices $\mathcal{P}'$. For each frame, we use $\mathcal{L}_{geo}$ to minimize the distance between body point $\mathcal{P}$ and vertex $\mathcal{P}'$ in Chamfer Distance. For each frame, the $\mathcal{L}_{geo}$ is defined as follows.

$$\mathcal{L}_{geo} = \frac{1}{k}\sum_{i=1}^{k}\sum_{\hat{p}\in P_i}\frac{1}{|P_i|}\min_{\hat{p'}\in P'_i}\|\hat{p'} - \hat{p}\|_2^2. \tag{9}$$

G. Details of Benchmarks

LiDAR-Based. We use three LiDAR-Based methods to estimate human posture, and the results of all methods are finally uniformly converted into SMPL models to calculate indicators. First, we use P4Transformer [6], a Transformer-based encoder specifically designed to process raw point cloud videos in both spatial and temporal dimensions. It reduces the amount of data that the Transformer needs to process by proposing 4D convolution of points as a feature extractor for spatiotemporal point clouds, thereby encoding the local structure of spatiotemporality. We use the point cloud features obtained using P4Transformer [6] and input them into SMPL to obtain the predicted human pose.

Second, we use PCT [7], which exploits the inherent order invariance of Transformer to avoid the need to explicitly define the order of point cloud data. PCT [7] learns feature representation through attention mechanism. The input point cloud is processed through the input embedding module, and then the attention output of each attention layer is concatenated along the feature dimension. A linear transformation is then applied to obtain the final PCT result. We use the same strategy as P4Transformer to input the results into SMPL to obtain the predicted human pose.

Finally, we also conduct experiments using LiDARCap [14], which is designed for human pose prediction. This method includes the extraction of point cloud features and the solution of human posture, so we can directly input the point cloud data provided in the data set and obtain the predicted human posture. LiDARCap [14] consists of three main components: a temporal encoder, an inverse motion solver, and an SMPL optimizer. The first step is to process the point cloud through PointNet++[17] and extract the 1024-dimensional global descriptor. Then, a temporal encoder implemented using GRU is used to fuse the temporal information in consecutive frames. Next, an MLP decoder is used to predict the positions of human joints based on the fused features. The predicted joint positions are combined with the 1024-dimensional features and fed into ST-GCN, which computes the predicted pose parameters. Finally, joint positions are calculated in the SMPL optimizer, resulting in predicted human body information, including pose and other relevant details.

RGB-Based. We use three RGB-based methods to reconstruct human pose. First, we test HybrIK, It employs a pre-trained HRNet network to extract 2048-dimensional image features. This network learns the twist angle and shape using a fully connected layer and generates heatmaps through deconvolution layers for the regressive calculation of 3D joint points. Subsequently, HybrIK utilizes Twist-and-Swing decomposition along with a series of IK methods to predict final human poses. Second, we conduct experiment with NIKI, which is a neural inverse kinematics solution for estimating 3D human pose and shape. It utilizes a bidirectional error decoupling IK algorithm, which is based on INN. We employ a simplified version of HybrIK to predict the parameters necessary for the bidirectional decoupling process. Subsequently, a single-layer NIKI is used to determine the final outcome. It is important to note that, for the HybrIK and NIKI methods, we mitigate errors from inaccurate target detection by employing ground truth bounding boxes as inputs for detection results. This strategy enables us to concentrate exclusively on the precision in reconstructing the human body. The third method we use is SMPLer-X. SMPLer-X is an extensive human model trained with over 4.5 million data points gathered from various sources, utilizing ViT-Huge as its core architecture. It has demonstrated strong performance across multiple benchmarks, prompting us to also conduct tests and experiments using this model.

Event-Based. We use two Event-Based methods. First, we test EventHPE [24]. After we aggregate the events into image-like event frames, we select inter-frame events as one packet and convert the event packets into 4-channel event frames. We pad and resize the event frame to 256 × 256. The event frames are then fed into FlowNet to infer optical flow. And use the ShapeNet that comes with the original article to receive event frames and optical flow to estimate attitude changes and global translation changes. We use the ground-truth pose and shape of the first frame as the starting pose and shape to subsequently estimate the body pose and shape at each time point. Second, in EventPointPose [3], events are scaled to 346×260. Then we select time slice K=4 to rasterize the event point set and sample 2048 points. The sampled rasterized event point cloud is then processed by the point converter backbone. Features output from the backbone network are fed into a linear layer to predict the 2D locations of key points on the human body. We select the "Last Label" setting to generate labels.

LiDAR+RGB-Based. We use two methods in experiments to predict human posture based on point clouds and RGB images. The first is ImmFusion [2], which can be divided into three main parts. In the first part, images and point clouds are passed through the modal masking module. This module randomly selects two modes, ensuring that the model is robust and not biased towards a specific mode. In the second part, PointNet++[17] is used to extract features from the point cloud. This extraction process produces local cluster features. Furthermore, MLP is adopted to obtain 1024-dimensional global features L^{pc}. For images, HrNet is used to extract local mesh features, which are then converted into global features using CNN. At the same time, the local grid features are adjusted by MLP to match the size of the local cluster features of the point cloud, generating features represented as L^{im}. The global features of the two modalities are fused together using a small Transformer module, and the template's vertices and joints are also incorporated into this fusion process. Finally, the fused global feature G is obtained. The last part combines the features L^{im}, G and L^{pc} to obtain the fusion features through the Fusion Transformer Module. Finally, the fused features are input into SMPL to obtain human pose.

The second is FusionPose [4], which is a method that combines 3D point clouds and 2D perspective RGB images for human pose prediction. In their proposed IPAFusion approach, a fusion technique is introduced to effectively combine the two modes. To extract features from point clouds, they adopt PointNet, which provides global feature representation. This global feature is then combined with the original feature to obtain the final feature, denoted as f_p. This fusion operation can be expressed as $f_p = LN(p + SelfAttention(p))$, where LN represents layer normalization. Similarly, for image modality, HrNet is selected to extract features. This process produces the final feature f_i, which can be expressed as $f_i = LN(i + SelfAttention(i))$, where i represents the initial image feature. In the image-to-point attention fusion part, they utilize cross-attention to fuse two features. Point cloud features f_p act as queries, while image features f_i act as values and keys. Finally, the fused features are input into GRU+MLP to obtain the predicted SMPL human pose.

Global RGB. We conduct experiments using two RGB-based methods incorporating global translation. The first method we explore is GLAMR. It begins by utilizing a Generative Motion Infiller to fill in human motion, effectively addressing issues of occlusion. Next, GLAMR employs a newly proposed global trajectory predictor to estimate future global trajectories. Finally, the method optimizes both the global trajectory of the human body and camera parameters simultaneously, resulting in the generation of global motion in world coordinates. Similar to NIKI, we utilize HybrIK as the motion prediction backbone for GLAMR, achieving the final human action through several stages.The second method, TRACE, is an end-to-end approach designed for inferring human actions within scenes. This method utilizes temporal features extracted from images and optical flow maps predicted to learn multiple feature maps. These feature maps are then processed using various head structures, enabling the completion of different human pose estimation tasks. For our experiments, we use inputs with a resolution of 1920x1080 and employ TRACE to deduce the final human pose and global trajectory for experimental evaluation.

H. Details of LEIR

Our proposed multi-modality baseline, LEIR, focuses on predicting the 3D pose of the human in world coordinate system based on synchronized LiDAR point clouds, RGB images and event streams. It contains feature extractors, the temporal unified multimodal model (TUMM), and SMPL-based inverse kinematics solver. The details of these three modules are explained as follows.

H.1. Feature Extractor

To extract the corresponding feature for each RGB frame, we first project the point cloud of the body onto the image, which could determine the boundary of the point cloud in this frame. Then, the bounding box that corresponds to the human body can be obtained. We crop the image from the bounding box and feed the image into a RGB encoder (DINOv2 [16]). The feature for RGB modality, $R_{f_{i-N}}$ is obtained.

The point clouds features are obtained through feeding the point clouds through the PointNet++ [17] network and a GRU network.

To appropriately handle noise in event frames, we utilize an average time sampling filter combined with adjacent point denoising [20], which helps to enhance the visibility of changes in human body movement across frames. After noise reduction, we aggregate all the events within a frame based on their pixel location and polarity to generate a new event frame that resemble an image. The features of this event frame, denoted as $E_{f_{i-N}}$, are then extracted using DINOv2 [16].

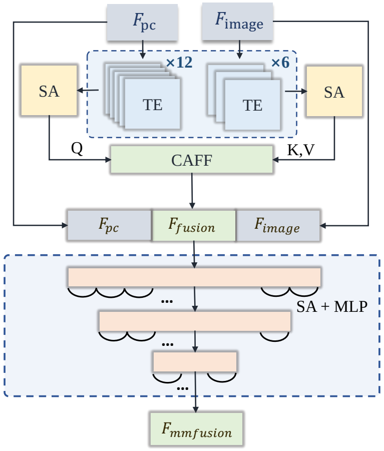
H.2. Temporal Unified Multimodal Model

The temporal unified multimodal model (TUMM) module is proposed to fully utilize the 3D geometric information of point clouds, the appearance details from RGB images and temporal dynamics in event streams.

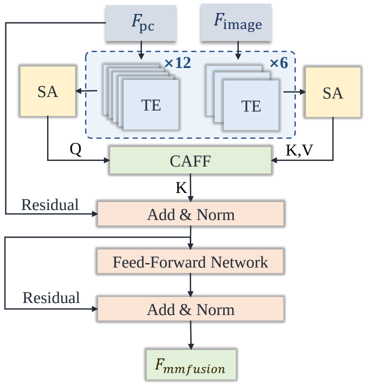
To fuse the features of point clouds and images, we employ the multi-modal cross attention unit (MMCA), which enables effective fusion of information from different modalities by leveraging a cross-attention module.

LEIR is a flexible method which can use different combinations of modalities as input. For single-modality input, the TUMM module consists only one step. In this step, the extracted input features are fed into the MMCA unit whose cross-attention module is replaced with a self-attention module. For two-modality input, the TUMM module consists one step. In this step, features from two modalities are fed into the MMCA unit.

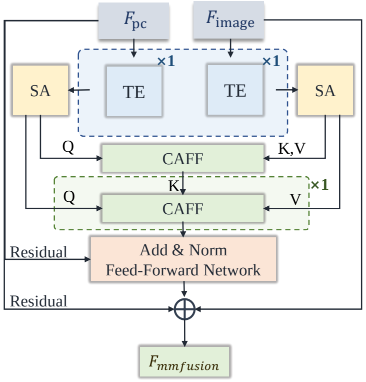
For three-modality input, the TUMM modules contains two steps. In the first step, the LiDAR point clouds and the RGB images are fused using the MMCA, LiDAR point clouds and the event frames are fused using MMCA as well. In the second step, the features obtained from the first step are further fused using MMCA, which allows a comprehensive integration of the features from different modalities. For the second step of TUMM, the 2D (right) branch of MMCA is replaced by a 3D branch.



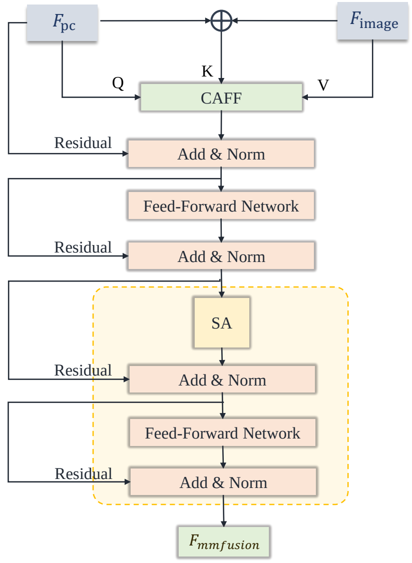
(a) ImmFusion-Based Strategy



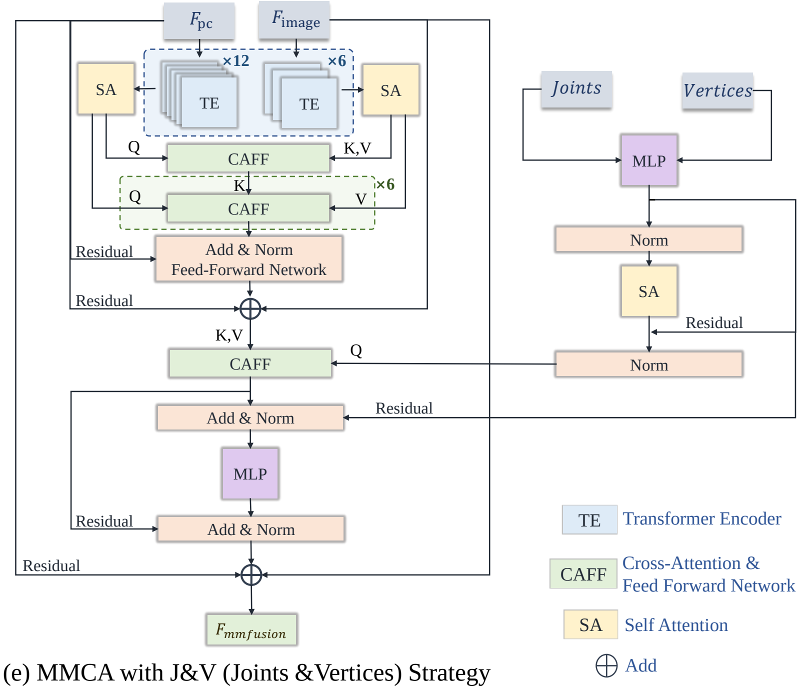
(b) FusionPose-Based Strategy



(c) MMCA *w/o* Multi-TE Strategy



(d) MMCA *w/o* Multi-CA&TE Strategy



(e) MMCA with J&V (Joints &Vertices) Strategy

Figure 4. **Different strategies of fusing multiple modalities for LEIR.** (a) Fusion Strategy of ImmFusion [2]. (b) Fusion Strategy of FusionPose [4] (c) MMCA without Multi-Transformer-Encoder (d) MMCA without Multi-Cross-Attention. (e) MMCA + Joints and Vertices.

H.3. SMPL-based inverse kinematics solver

The fused features $F_{mmfusion}$ obtained from TUMM, are used in this solver, which consists of three branches. In the first branch, the extracted features are inputted into a 3D regressor, which is responsible for estimating the 3D joints and camera intrinsic parameters. To guide the training and ensure accurate estimation, three loss functions are employed in this branch. The first loss function, $\mathcal{L}_{ps2d}$, serves as a projection loss, which ensures that the 2D appearance of the SMPL model aligns with the human body in pixel coordinates. By minimizing the discrepancy between the projected 2D model and the observed human body in the image, this loss function aids in achieving accurate alignment and pixel-level correspondence. This loss is defined as:

$$\mathcal{L}_{ps2d} = \mathcal{L}_{shape2d} + \mathcal{L}_{pose2d} \tag{10}$$

$\mathcal{L}_{ps2d}$ consists two terms that specifically target the pose and shape parameters of the SMPL model. The shape term β is the 10-dimensional shape parameter of the SMPL model. The pose term, θ, is a $N \times 3 \times 3$ rotation matrix, where N is 24 and represents the number of joint points. $L_{shape2d}$ and L_{pose2d} are defined as follows.

$$\mathcal{L}_{shape2d} = \frac{1}{10}\sum_{i=1}^{10}(\beta_{pred_i} - \beta_{gt_i})^2 \tag{11}$$

$$\mathcal{L}_{pose2d} = \frac{1}{N \times 3 \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}\sum_{k=1}^{3}(\theta_{pred_{ijk}} - \theta_{gt_{ijk}})^2 \tag{12}$$

where β_{pred} and β_{gt} are the predicted and ground-truth shapes, respectively. θ_{pred} and θ_{gt} are the predicted and ground-truth poses, respectively.

The second loss function, $\mathcal{L}_{kp2d}$, is used to constrain the 2D joints of the human body. By comparing the estimated joints $KP2d_{pred}$ with the ground truth annotations $KP2d_{gt}$, this loss function encourages the regressor to accurately capture the spatial relationships and positions of the joints in the 2D image space.

$$\mathcal{L}_{kp2d} = \frac{1}{N \times 2}\sum_{i=1}^{N}\sum_{j=1}^{2}(KP2d_{pred_{ij}} - KP2d_{gt_{ij}})^2 \tag{13}$$

The 3D joints predicted by the 3D regressor are constrained by the loss $\mathcal{L}_{kp3d}$, which ensures that the regressor accurately captures the spatial relationships and positions of the joints by comparing predicted joints $KP3d_{pred}$ with the ground truth annotations $KP3d_{gt}$.

$$\mathcal{L}_{kp3d} = \frac{1}{N \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}(KP3d_{pred_{ij}} - KP3d_{gt_{ij}})^2 \tag{14}$$

In the second branch of the solver, the extracted features are fed into a RNN network, which is designed to generate the 3D human joints in the world coordinate system. To guide the training process and ensure appropriate joint prediction, we employ $\mathcal{L}^{W}_{joint}$ to encourage alignment between the predicted 3D joints Jt_{pred} and the ground truth labels Jt_{gt}.

$$\mathcal{L}^{W}_{joint} = \frac{1}{N \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}(Jt_{pred_{ij}} - Jt_{gt_{ij}})^2 \tag{15}$$

The third branch of our approach employs ST-GCN [21], where the fused features from the previous branches are utilized to predict the 3D human joints. To ensure accurate joint orientation, we apply $\mathcal{L}^{smpl}_{joint}$ to encourage alignment between the predicted joint orientations Jt_{pred} and the ground truth orientations Jt_{gt}.

$$\mathcal{L}^{smpl}_{joint} = \frac{1}{N \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}(Jt^{smpl}_{pred_{ij}} - Jt^{smpl}_{gt_{ij}})^2 \tag{16}$$

The outputs of this branch, which represent the predicted 3D joints, are then passed through a SMPL optimizer that converts the joint positions into human poses in axis-angle form. And the loss $\mathcal{L}^{smpl}_{pose}$ is employed to enforce alignment between predicted pose θ^{smpl}_{pred} with the ground truth poses θ^{smpl}_{gt}.

$$\mathcal{L}^{smpl}_{pose} = \frac{1}{N \times 3 \times 3}\sum_{i=1}^{N}\sum_{j=1}^{3}\sum_{k=1}^{3}(\theta^{smpl}_{pred_{ijk}} - \theta^{smpl}_{gt_{ijk}})^2 \tag{17}$$

All the aforementioned losses play a crucial role in our method to achieve accurate estimates of human pose.

I. Fusion Strategies

(a)ImmFusion-Based Strategy. For this strategy [2], as shown in (a) of Fig. 4, after the initial cross-attention operation that fuses the two modalities, it concatenates the fused features with the features from the two modalities. Subsequently, a series of self-attention and MLP layers are applied to obtain the final fusion features.

(b)FusionPose-Based Strategy. The FusionPose [4] strategy is depicted in Fig. 4 (b). After the initial cross-attention operation, the fused features are directly combined with the point cloud features. To further refine the fusion, it employs Feed-Forward network layers and residual structures, which could help in enhancing the fusion features by incorporating the point cloud information while preserving the original features from the cross-attention operation.

(c)MMCA *w/o* Multi-TE Strategy. Fig. 4 (c) illustrates a modified MMCA unit. We change the Transformer Encoder in MMCA from multiple layers to just one layer. Through this experimentation, we can determine whether the original RELI method's multi-layer transformer encoder stacking contributes to the feature extraction and fusion process.

(d)MMCA *w/o* Multi-CA&TE Strategy. As it is depicted in Fig. 4 (d), cross-attention is removed from MMCA. This strategy directly adds the point cloud and image features, and they are further fused with the initial features of the point cloud and image, which allows for the integration of the initial fusion results with the original features.

(e)MMCA + J&V (Joints&Vertices) Strategy. This strategy is depicted in Fig. 4 (e), which takes joints and vertices as *additional inputs*. It fuses other modalities with cross attention. During training, this strategy uses ground-truth joints and vertices as the additional input. During testing, this strategy uses the joints and vertices predicted by LiDARCap as inputs.

References

[1] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS Dataset and Benchmark Track*, June 2023. 2

[2] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2752–2758. IEEE, 2023. 1, 2, 5, 7, 8

[3] Jiaan Chen, Hao Shi, Yaozu Ye, Kailun Yang, Lei Sun, and Kaiwei Wang. Efficient human pose estimation via 3d event point cloud. In *3DV*, 2022. 5

[4] Peishan Cong, Yiteng Xu, Yiming Ren, Juze Zhang, Lan Xu, Jingya Wang, Jingyi Yu, and Yuexin Ma. Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In *AAAI*, pages 461–469. AAAI Press, 2023. 1, 2, 6, 7, 8

[5] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. Hsc4d: Human-centered 4d scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6792–6802, June 2022. 4

[6] Hehe Fan, Yi Yang, and Mohan S. Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14199–14208, 2021. 5

[7] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7:187–199, 2021. 5

[8] Sagi Katz, Ayellet Tal, and Ronen Basri. Direct visibility of point sets. In *ACM SIGGRAPH 2007 papers*, pages 24–es. 2007. 4

[9] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 3

[10] Misha Kazhdan and Hugues Hoppe. An adaptive multi-grid solver for applications in computer graphics. In *Computer graphics forum*, volume 38, pages 138–150. Wiley Online Library, 2019. 3

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

[12] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023. 1, 2

[13] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 1, 2

[14] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20502–20512, 2022. 1, 2, 3, 5

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015. 3

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[17] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5, 6

[18] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009. 4

[19] Ayça Takmaz, Jonas Schult, Irem Kaftan, Mertcan Akçay, Bastian Leibe, Robert Sumner, Francis Engelmann, and Siyu Tang. 3d segmentation of humans in point clouds with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1292–1304, 2023. 4

[20] Ming Yan, Yewang Chen, Yi Chen, Guoyao Zeng, Xiaoliang Hu, and Jixiang Du. A lightweight weakly supervised learning segmentation algorithm for imbalanced image based on rotation density peaks. *Knowledge-Based Systems*, 244:108513, 2022. 6

[21] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 8

[22] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11028–11039. IEEE, 2022. 1, 2

[23] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *ICCV*, pages 10976–10985, 2021. 1, 2

[24] Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, and Li Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10996–11005, 2021. 5