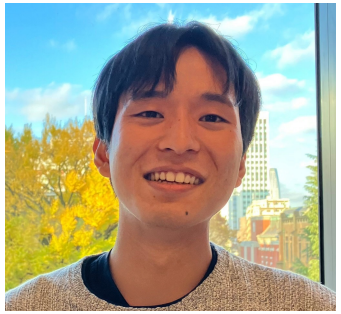
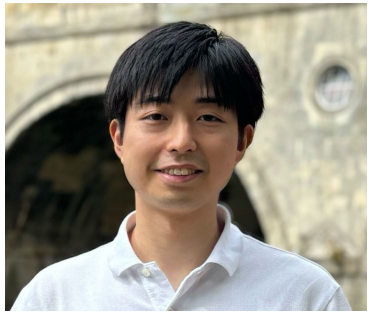
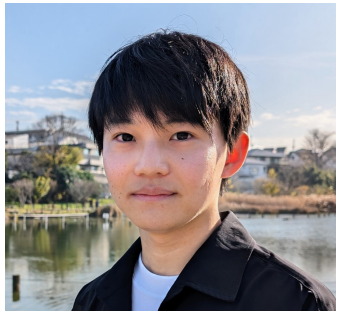
IEEE ICIP 2025

Keio University
1858
CALAMVS
GLADIO
FORTIOR

Event-based Egocentric Human Pose Estimation in Dynamic Environment


Wataru Ikeda† Masashi Hatano† Ryosei Hara† Mariko Isogawa†‡

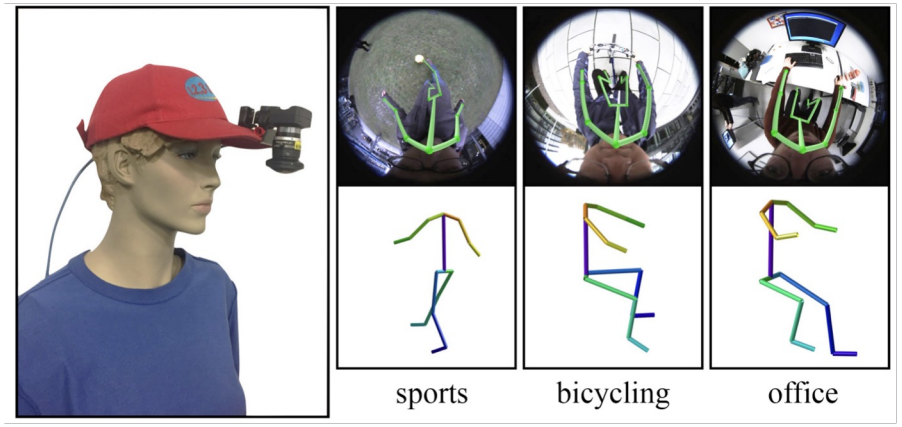
†Keio University, ‡JST Presto

Background: Egocentric Human Pose Estimation

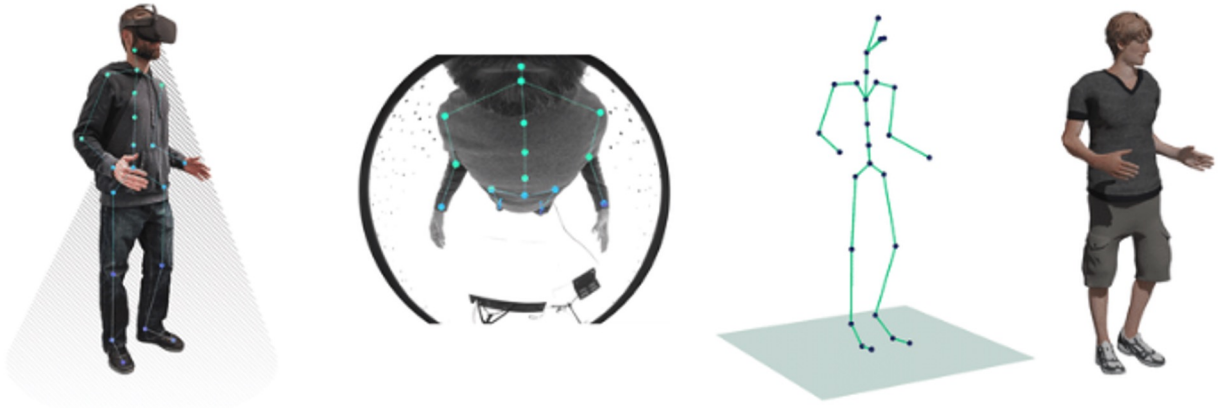
Human Pose Estimation from egocentric vision data captured by wearable devices

→ Applications in sports motion analysis, AR/VR avatar generation, healthcare

Challenges: Motion blur, low light performance



Mo²Cap² [1]



SelfPose [2]

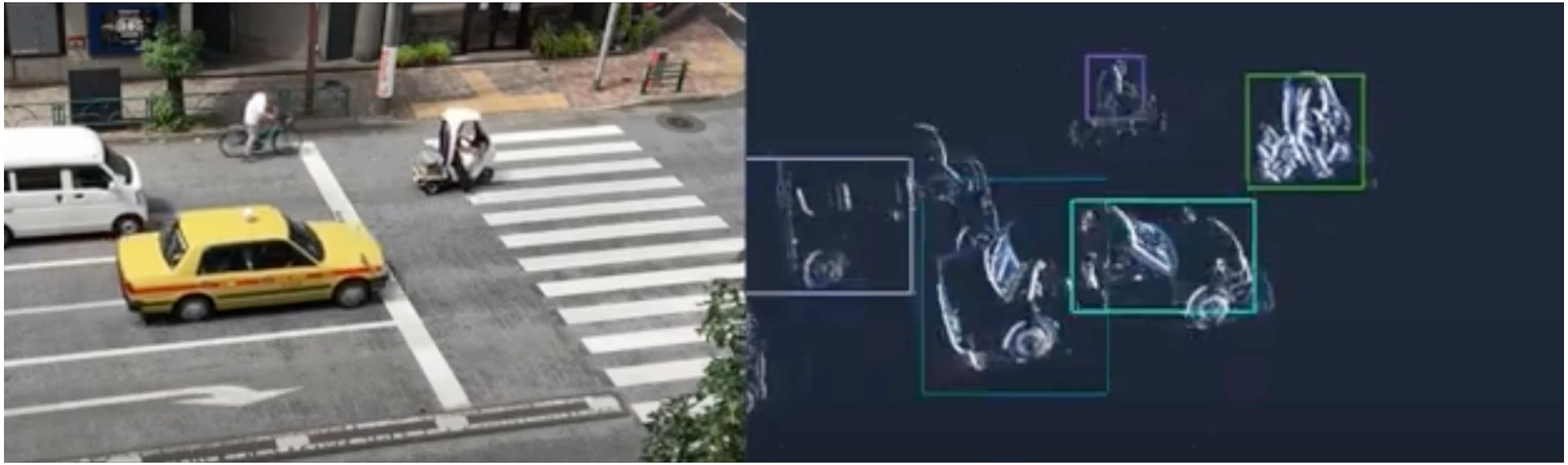
[1] Xu, et al., "Mo2Cap2 : Real-time mobile 3D motion capture with a cap-mounted fisheye camera". In TVCG, 2019.

[2] Tome, et al., "SelfPose: 3d egocentric pose estimation from a headset mounted camera.", In TPAMI, 2020.

Background: Event-based Camera

Captures per-pixel brightness changes

Advantages: High dynamic range, high temporal resolution, etc.



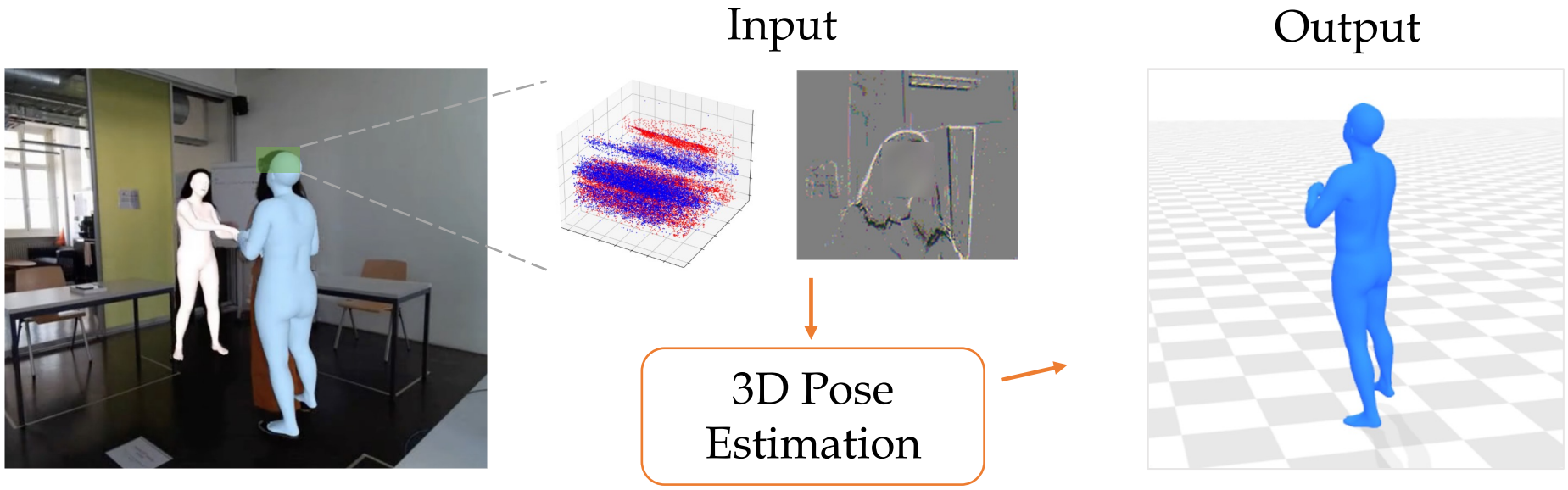
Video of a Moving Car (Left: RGB Camera, Right: Event Camera) [3]

[3] https://www.youtube.com/watch?v=on9rhTxbWPo, Accessed: July 29, 2025.

Our Goal: Event Egocentric HPE with front-facing camera

Novel task: Event-based Egocentric Human Pose Estimation using front-facing camera in Dynamic environment

→ Enhanced low-light capability and easy integration into wearable devices

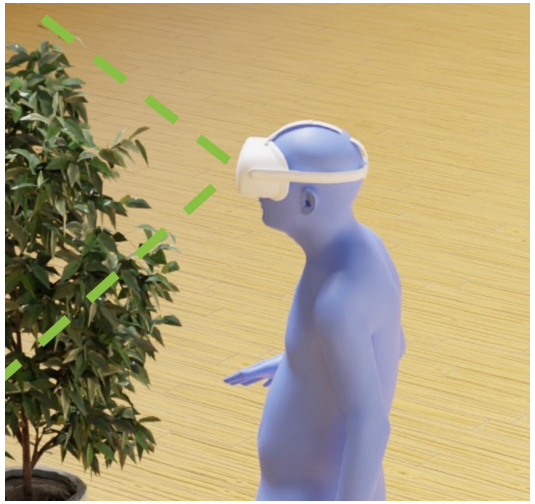


Input / Output in this task

Our Goal: Event Egocentric HPE with front-facing camera

Novel task: Event-based Egocentric Human Pose Estimation using front-facing camera in Dynamic environment

→ Enhanced low-light capability and easy integration into wearable devices



Related Works: HPE using head-mounted cameras

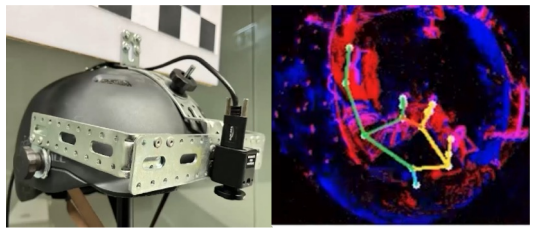
RGB

Event

Methods assuming body parts are visible

Selfpose [2]
etc.

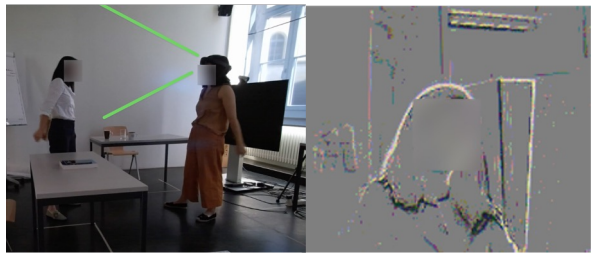
EventEgo3D++ [5]



Methods assuming the body is not visible

EgoPose [4]
etc.

Ours

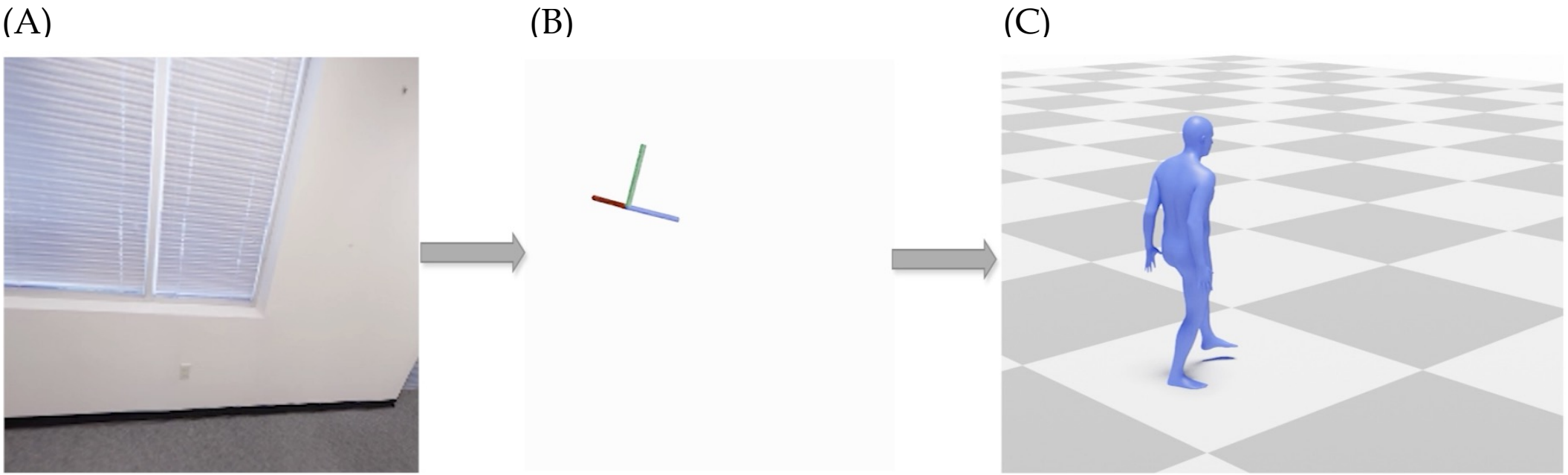


[4] Yuan, et al., “Ego-Pose Estimation and Forecasting as Real-Time PD Control”, In ICCV, 2019.

[5] Millerdurai et al., “Eventego3d++: 3d human motion capture from a head-mounted event camera”, In IJCV, 2025.

Baseline: EgoEgo [6]

- Full-body pose estimation from RGB video
- Uses head pose as an intermediate representation

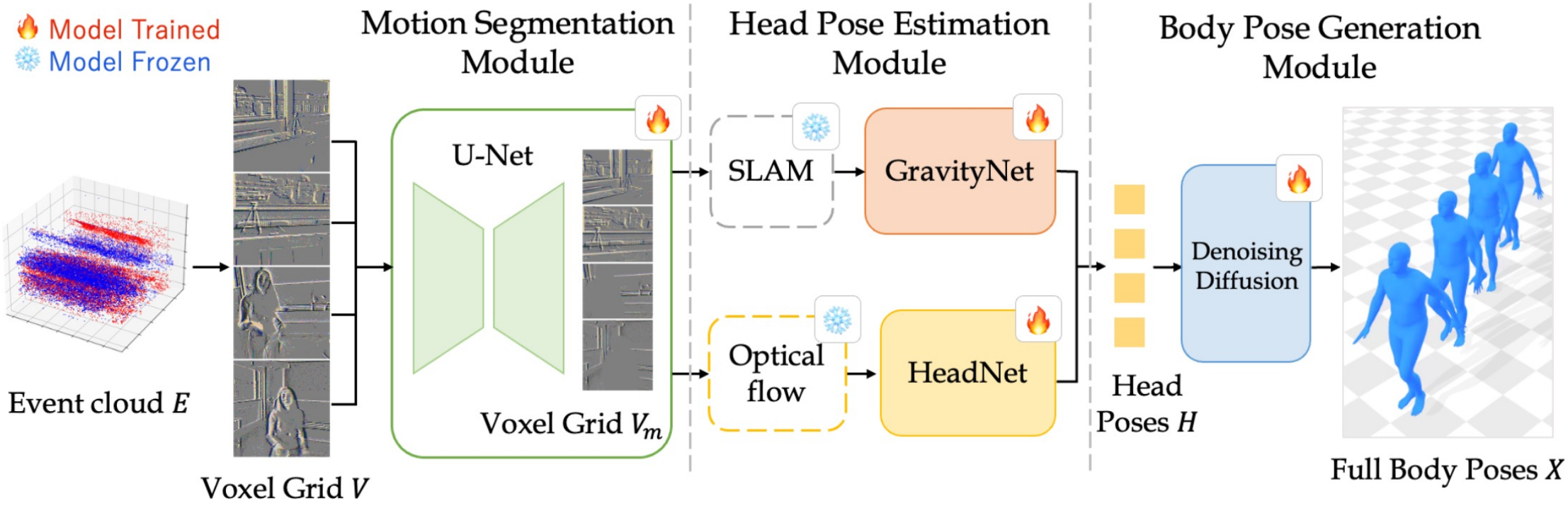


(A) Input RGB Video, (B) Head Pose Estimation, (C) Full-body Pose Estimation [6]

[6] Li, et al., "Ego-Body Pose Estimation via Ego-Head Pose Estimation", In CVPR, 2023.

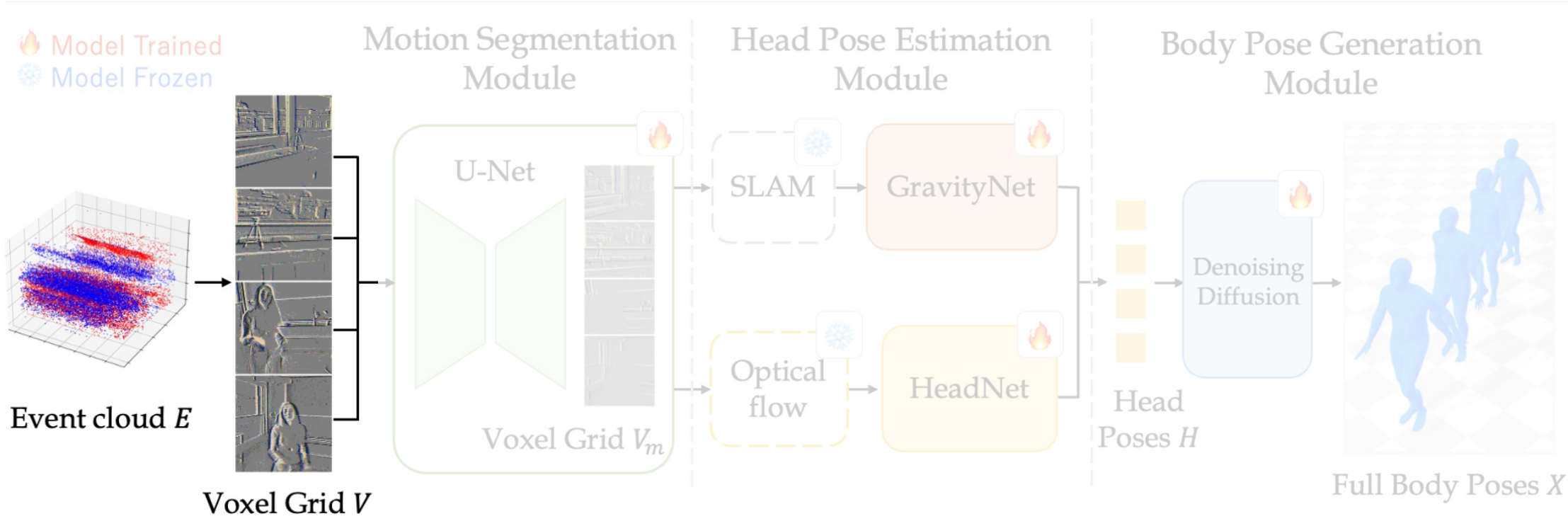
Method: Overview

- Proposed Framework Inspired by EgoEgo
- Includes a Motion Segmentation Module to handle dynamic environments



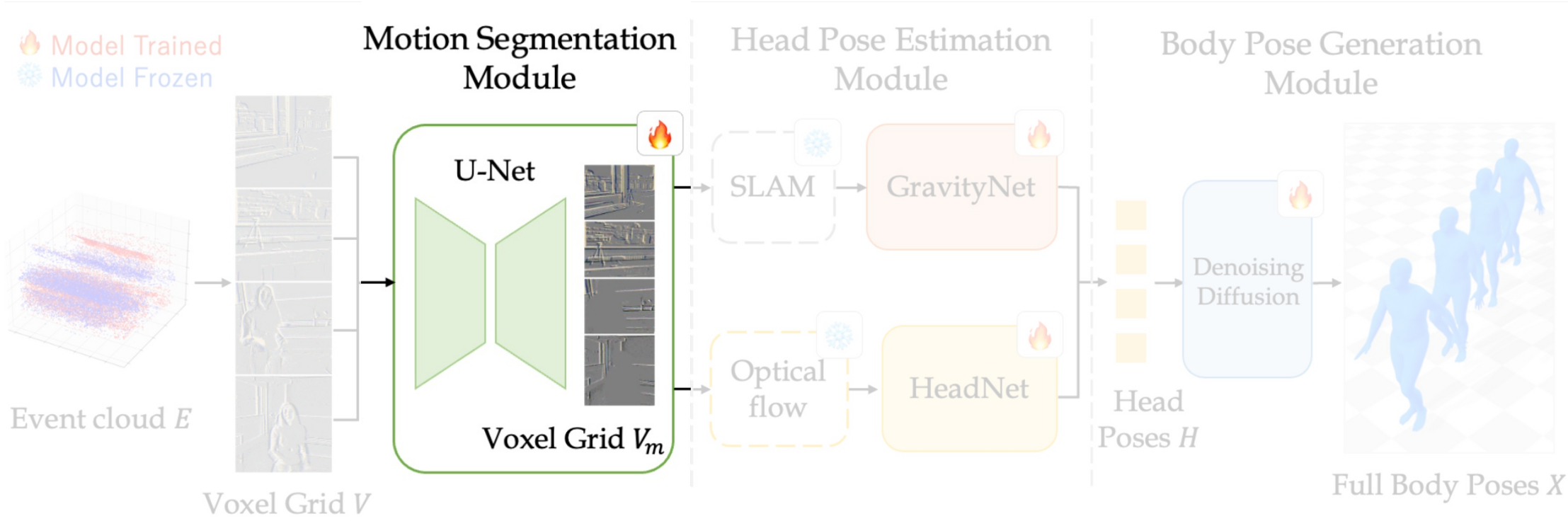
Method: Voxelization

- **Converts input event data into voxel grids**
- To reduce computation while preserving some temporal information



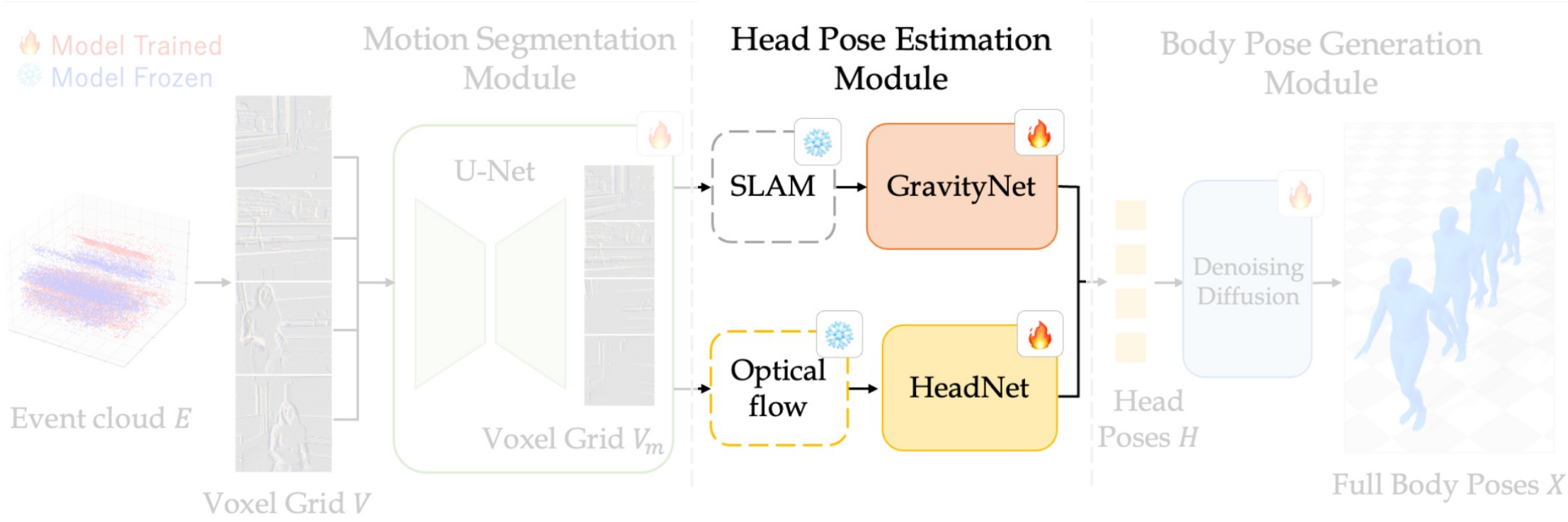
Method: Motion Segmentation Module

- Removes dynamic objects and extract background information
- Uses U-Net to learn masks of dynamic objects in the voxel grid



Method: Head Pose Estimation Module

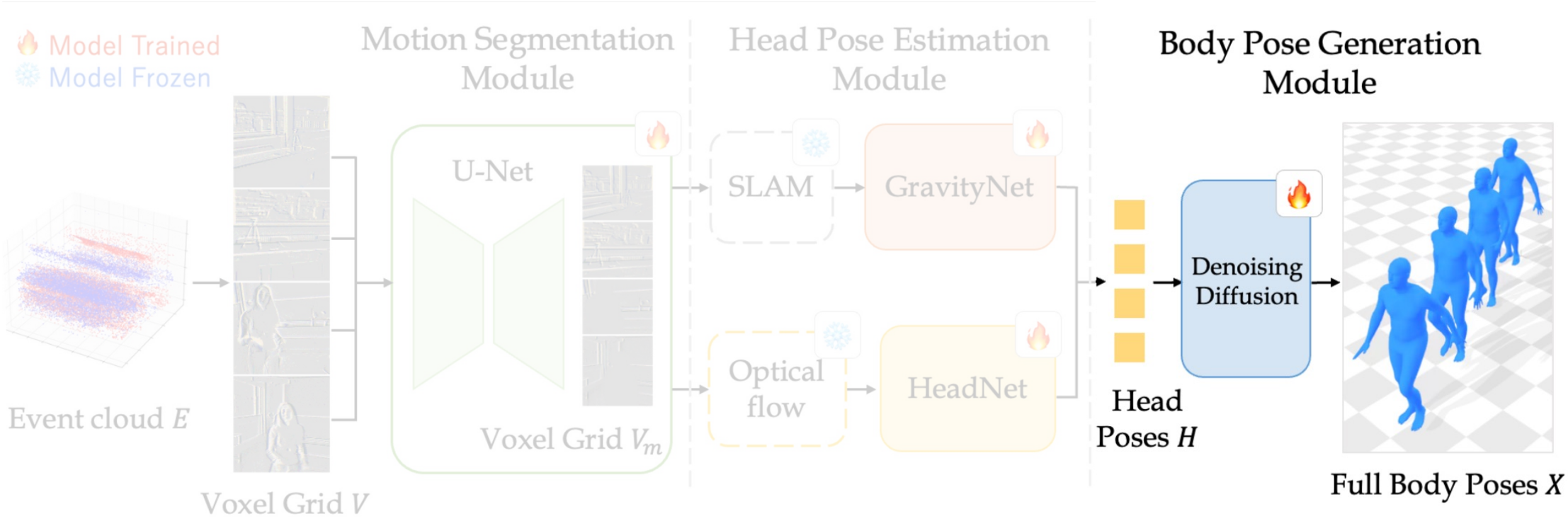
- Camera pose estimated by monocular SLAM [7]
- HeadNet and GravityNet interpolate movement distance and gravity direction



[7] Teed, et all., "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," in NeurIPS, 2021.

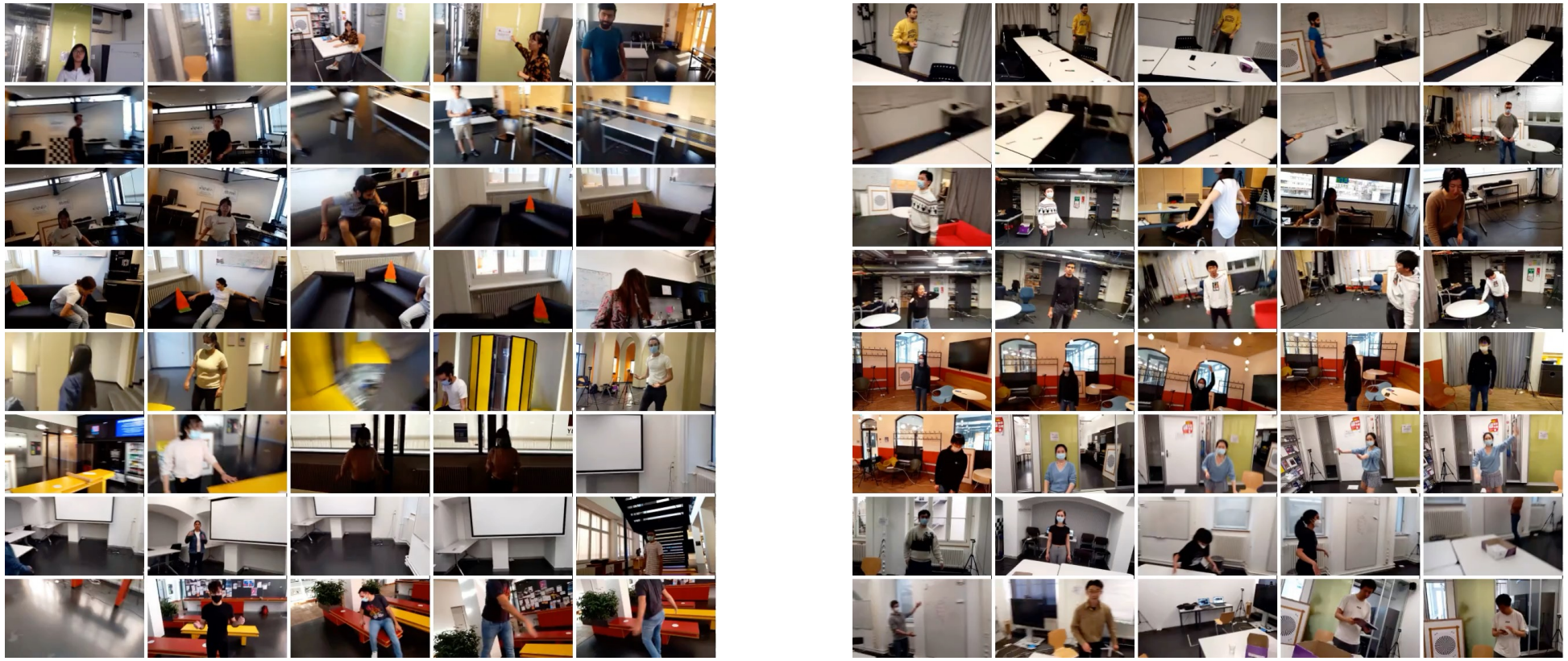
Method: Body Pose Generation Module

- Generates full-body pose conditioned on head pose
- Uses Diffusion Model (DDPM)



Dataset: Overview

Event data generated from EgoBody [8] RGB dataset in dynamic scenes



Egocentric videos from the EgoBody dataset

[8] Zhang, et al., "Egobody: Human body shape and motion of interacting people from head-mounted devices", In ECCV, 2022.

Dataset: Detailed Information

Generate events from over 190,000 RGB frames using DVS-Voltmeter [9]

→ Collect event data, dynamic masks, and 3D human pose information

Event Voxel Grid

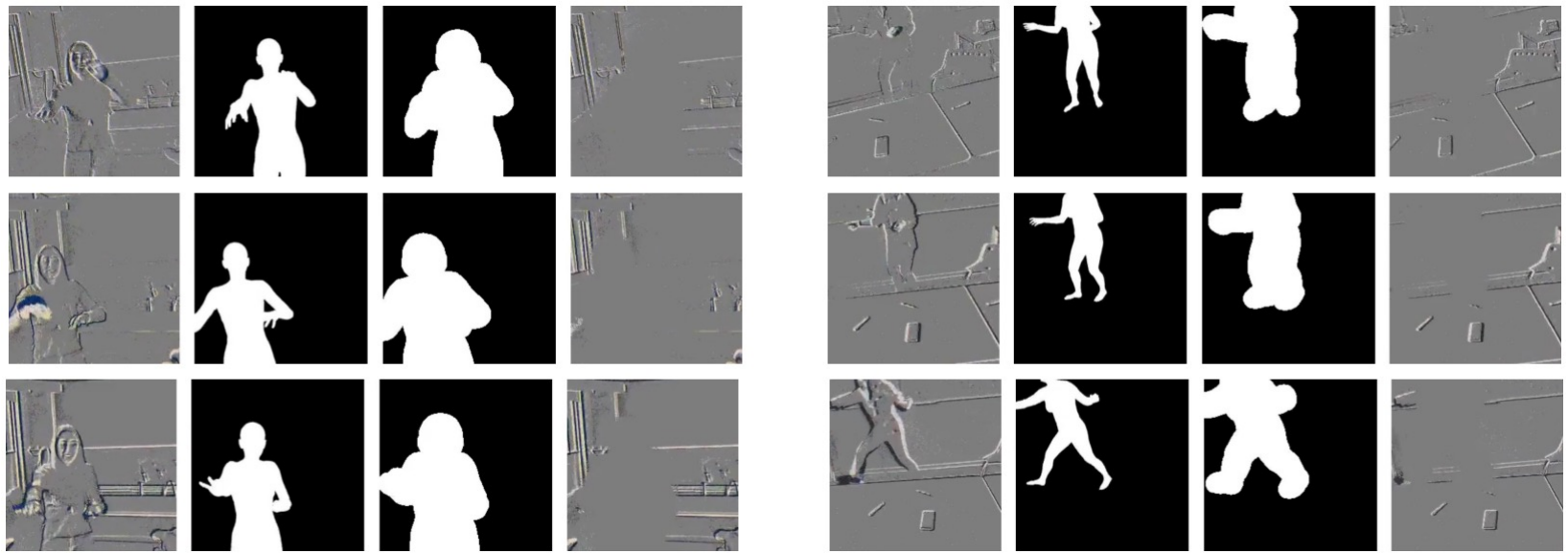
Dynamic mask

SMPL-X

[9] Lin, et al., "DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors", In ECCV, 2022.

Dataset: Dynamic Mask Generation

1. Project the ground-truth mesh onto the 2D image
2. Apply dilation (kernel size 5, one iteration). Event regions appear larger than the actual object, so dilation helps ensure proper removal.



Dataset constructed for this task

Implementation Details

Motion Segmentation Module:

- Optimizer: AdamW [10], LR: 1.0×10^{-5}, Batch size: 32
- Trained on RTX 4090, ~23 hours to converge

HeadNet

- Optimizer: AdamW [10], LR: 1.0×10^{-4}, Batch size: 64
- Trained on A6000, ~1 hour

GravityNet

- Optimizer: AdamW [10], LR: 1.0×10^{-4}, Batch size: 256
- Trained on A6000, ~1 hour

Denoising Difffusion

- Optimizer: AdamW [10], LR: 1.0×10^{-4}, Batch size: 32
- Trained on RTX 6000 Ada, ~32 hours
- Fine-tuned from a pre-trained EgoEgo model

[10] Loshchilov, et all., “Decoupled weight decay regularization”, In arXiv, 2017.

Evaluation Metrics

$MPJPE$ (Mean Per Joint Position Error)

Average Euclidean distance between predicted and ground-truth joint positions

O_{head} (Head Orientation Error)

Head rotation error, measured by Frobenius norm of the difference between predicted and ground-truth rotation matrices

T_{head} (Head Translation Error)

Average Euclidean distance between predicted and ground-truth head positions

$Accel$

Difference in joint accelerations between prediction and ground truth

FS (Foot Skating)

Amount of foot sliding, based on foot joint velocity on the xy-plane (NeMF [11])

[11] He, et all., "Nemf: Neural motion fields for kinematic animation", In NeurIPS, 2022.

Quantitative Evaluation: Full-body Pose Estimation

Our approach outperformed against baselines on 4 of 5 evaluation metrics

Quantitative Results for Full-body Pose Estimation

Method	O_{head} ↓	T_{head} [mm]↓	MPJPE [mm]↓	Accel [mm/s²]↓	FS [mm]↓
EgoEgo*	0.293	126.6	**119.5**	2.87	0.79
	− 3.8%	− 3.8%	+ 1.7%	− 6.2%	− 19.0%
Ours	**0.282**	**121.8**	121.5	**2.69**	**0.64**

* EgoEgo using Event Voxel Grid as input

Quantitative Evaluation: Head Pose Impact

Our approach outperformed against baselines on both evaluation metrics

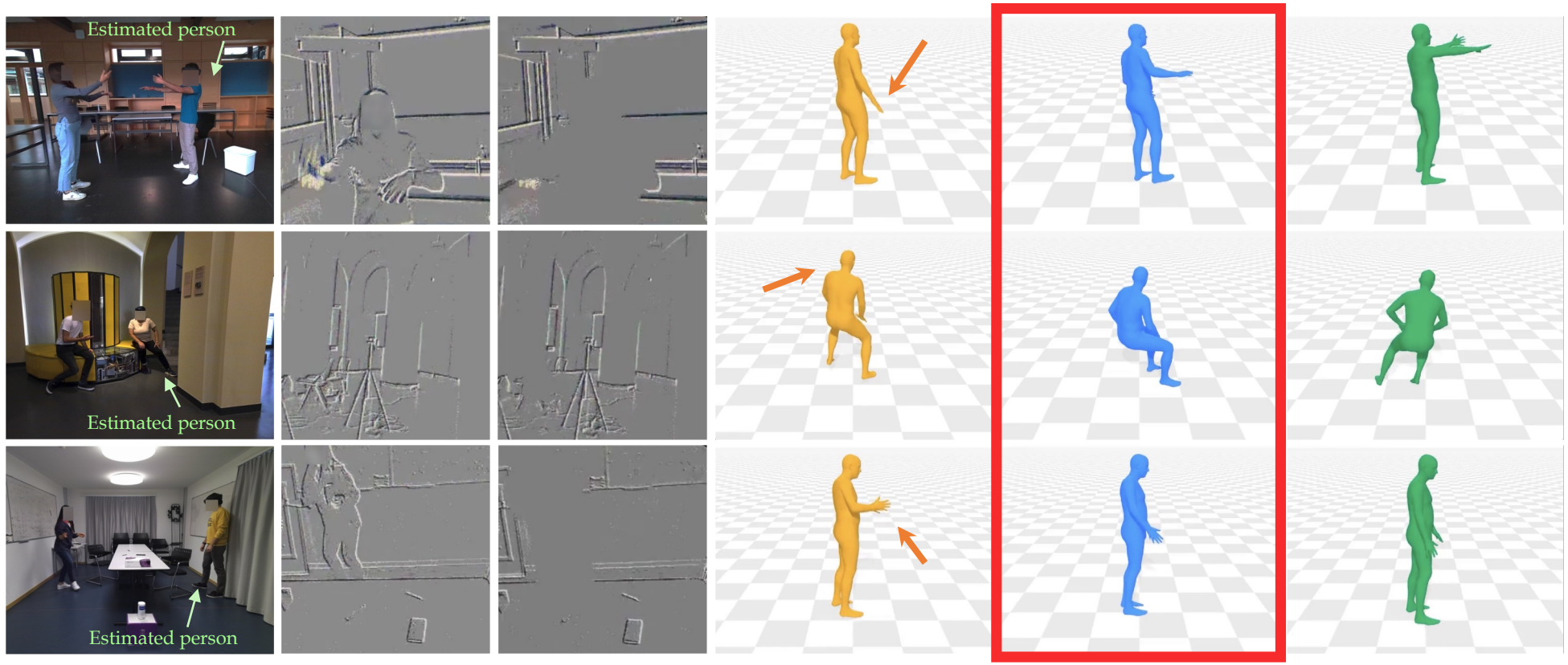
Quantitative Results on Head Pose

Method	O_{head} ↓		T_{head} [mm]↓	
EgoEgo*	0.286	– 3.1%	122.57	– 2.7%
Ours	**0.277**		**119.23**	

* EgoEgo using Event Voxel Grid as input

Qualitative Evaluation: Overview

Improved head pose accuracy leads to better estimation of hand position and head height

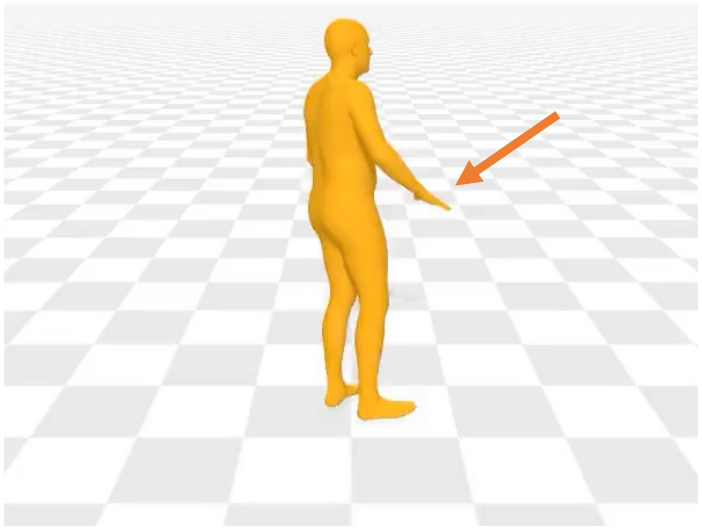


* EgoEgo using Event Voxel Grid as input

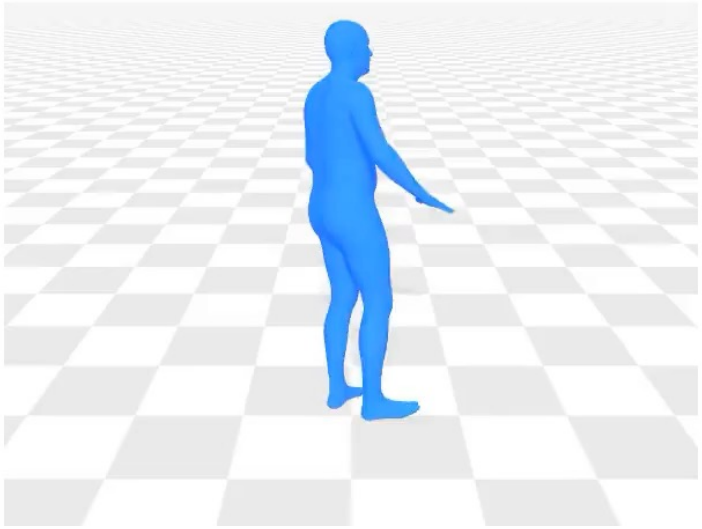
Moving hands

(Ref) Third Person View

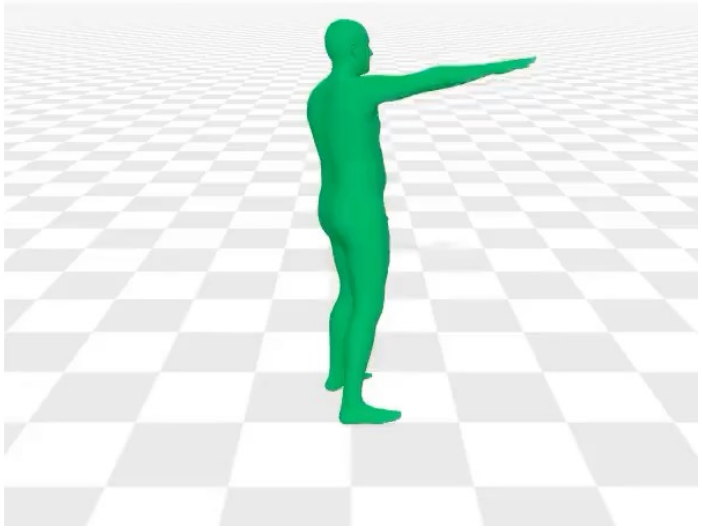
Voxel Grid



EgoEgo*



Ours

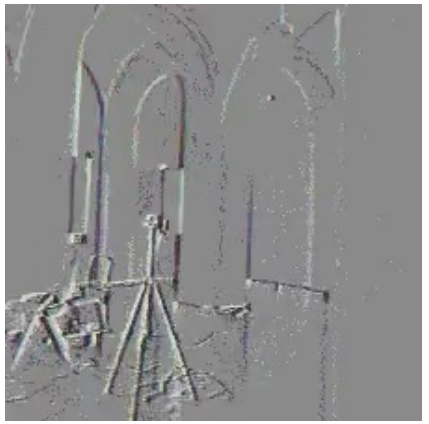


GT

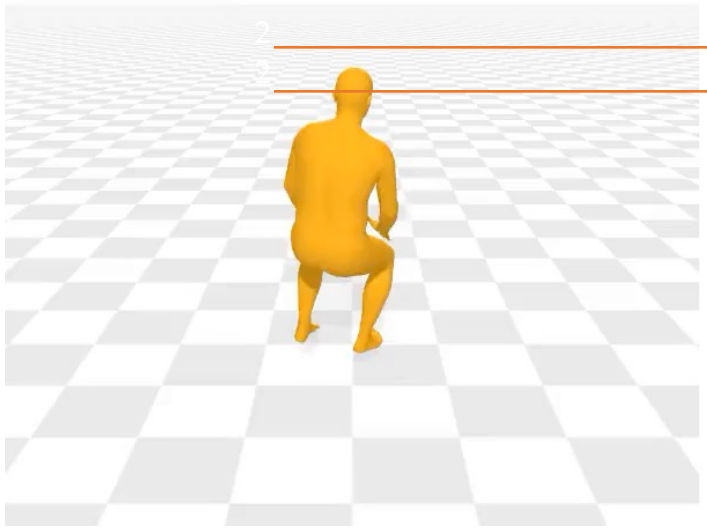
* EgoEgo using Event Voxel Grid as input

Sitting down

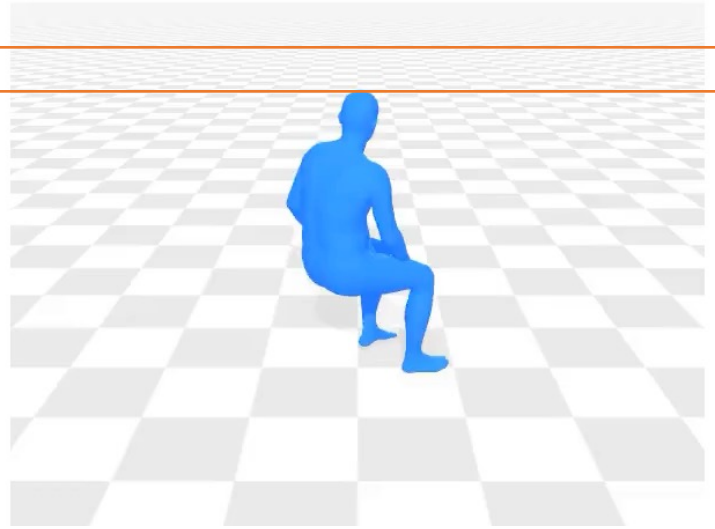
(Ref) Third Person View



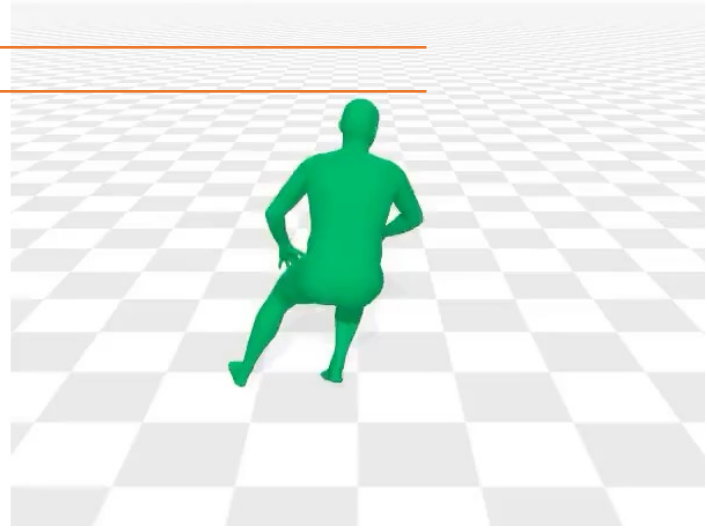
Voxel Grid



EgoEgo*



Ours

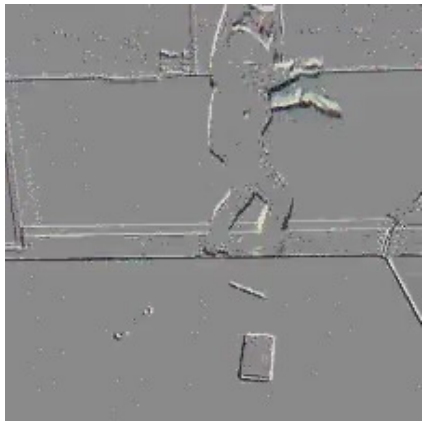


GT

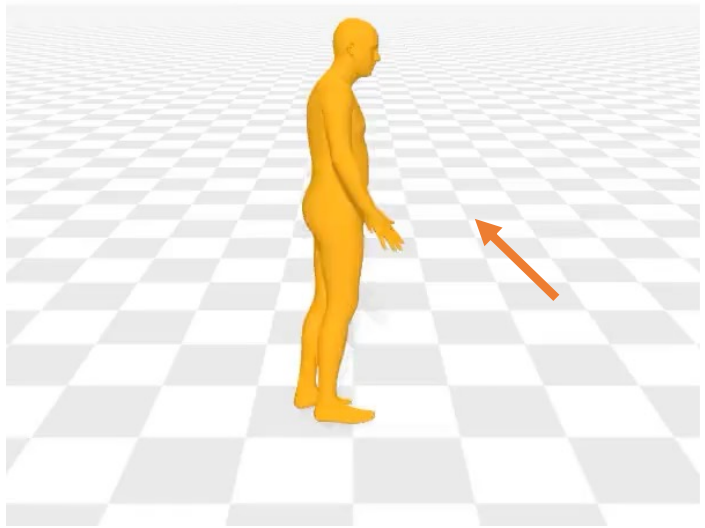
* EgoEgo using Event Voxel Grid as input

Taking with gestures

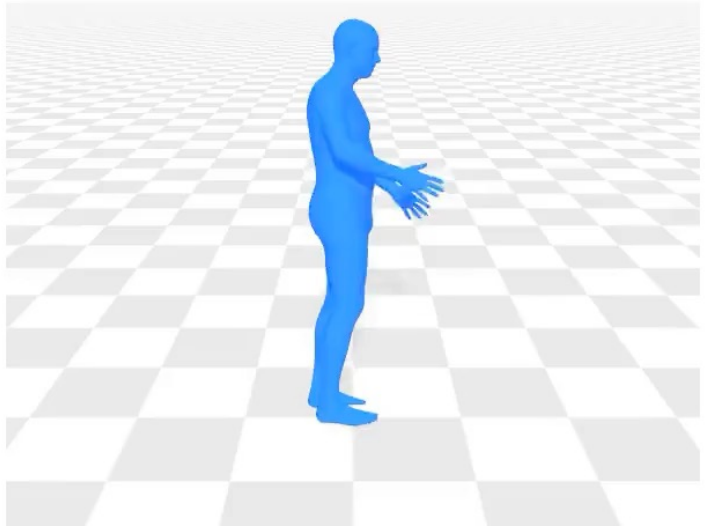
(Ref) Third Person View



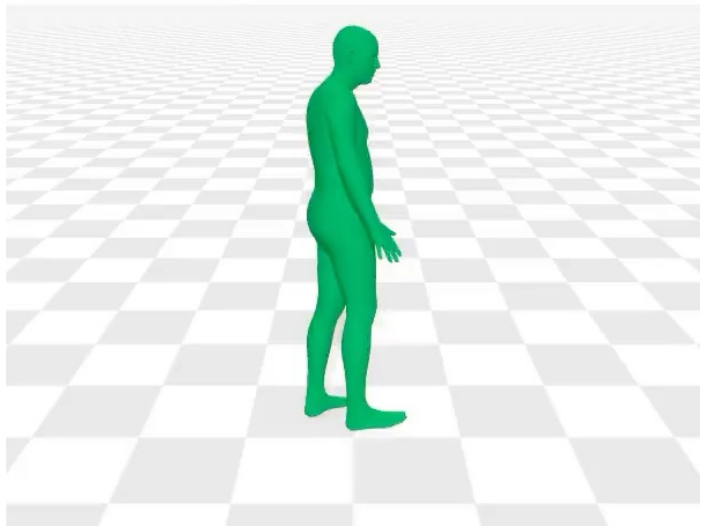
Voxel Grid



EgoEgo*



Ours



GT

* EgoEgo using Event Voxel Grid as input

Summary

- Proposed a **novel task and framework** for human pose estimation using a **head-mounted, front-facing egocentric event camera**
- Introduced a **Motion Segmentation Module** to remove dynamic objects and extract background information in **dynamic environments**
- Built an **original dataset** using an event simulator — the **first egocentric front-facing event dataset** for this task
- Experiments show our method **outperforms baseline approaches** in both **quantitative and qualitative evaluations**

Appendix: First Author Biography

Research Interests:
Human Motion, Egocentric Vision, Event-based Camera

Apr. 2024 – Present (Jul. 2025): Keio University, Yokohama, Japan

- M.S. in Engineering
  - Advisor: Dr. Mariko Isogawa

Apr. 2020 - Mar. 2024: Keio University, Yokohama, Japan

- B.S. in Engineering
  - Advisor: Dr. Mariko Isogawa