南京航空航天大學
Nanjing University of Aeronautics and Astronautics

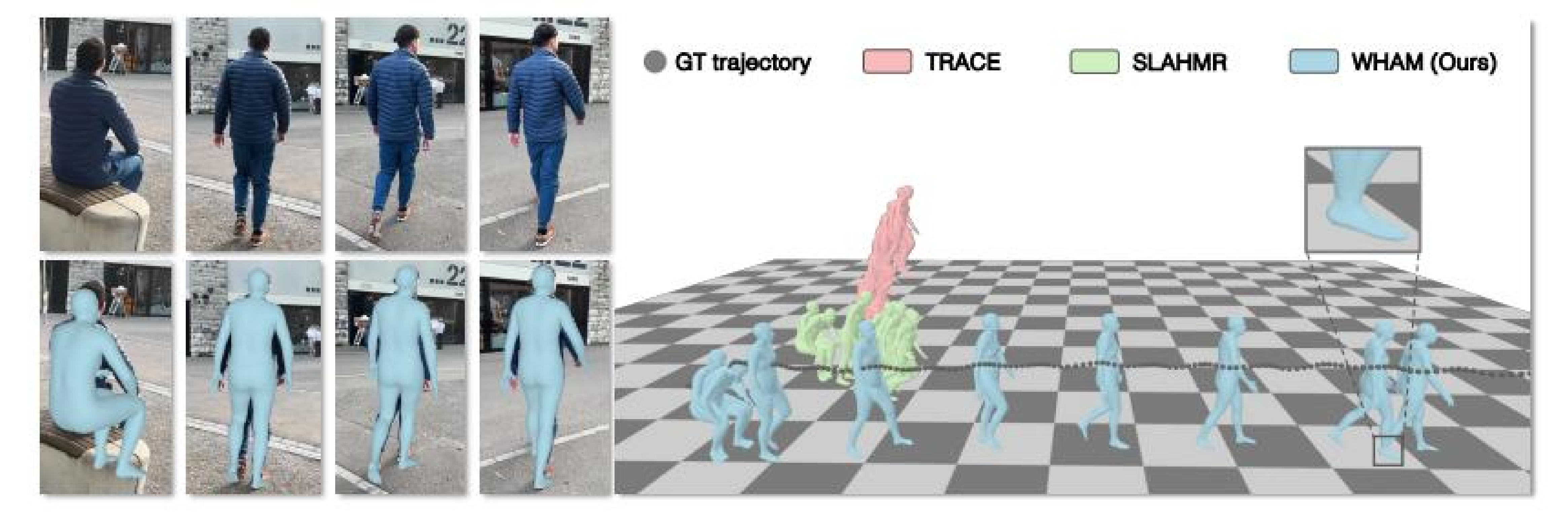
WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion

Soyong Shin[1] , Juyong Kim[1] , Eni Halilaj[1] , Michael J. Black[2]

Carnegie Mellon University[1]
Max Planck Institute for Intelligent System[2]

CVPR2024

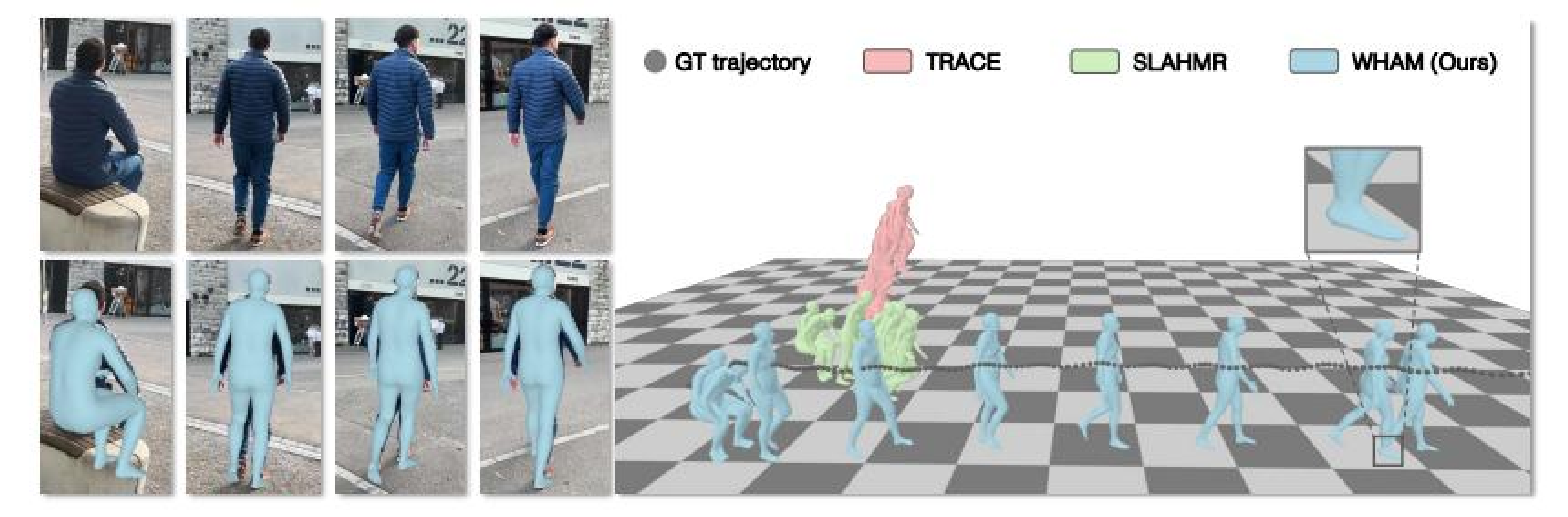
Introduction



The limitations of current methods:

- Most methods estimate the human in **camera coordinates**, and often have unsatisfactory performance in **global coordinates**.

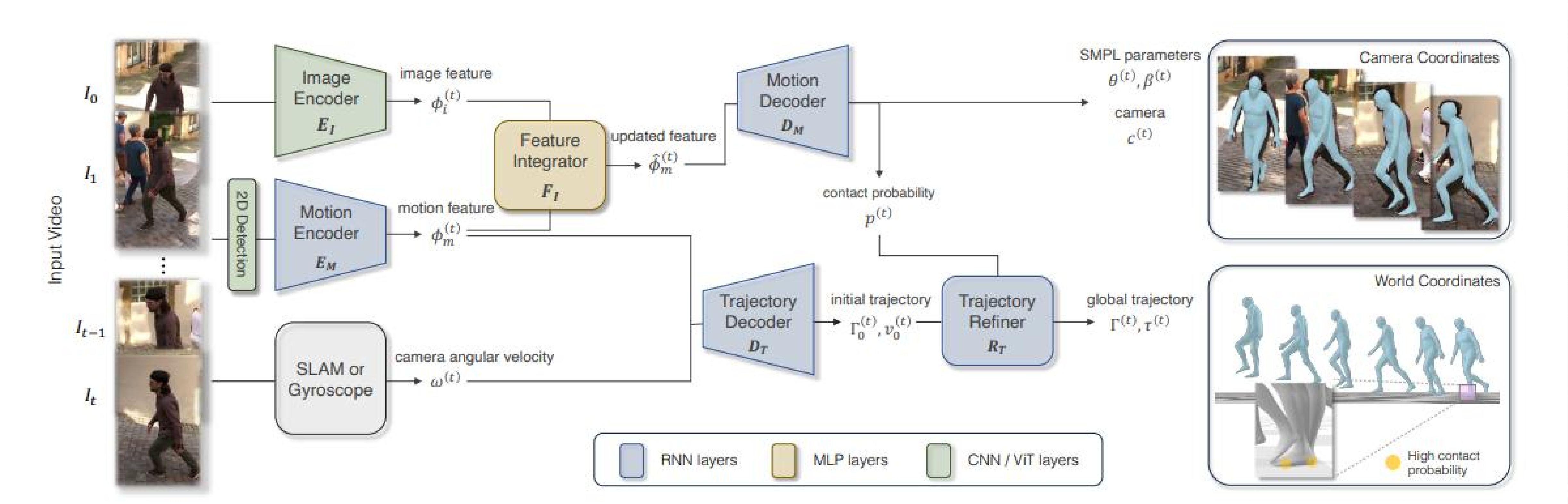
Introduction



The limitations of current methods:

- The most accurate methods rely on **computationally expensive optimization pipelines**, limiting their use to offline applications.
- Existing **video-based methods** are surprisingly less accurate than **single-frame methods**.

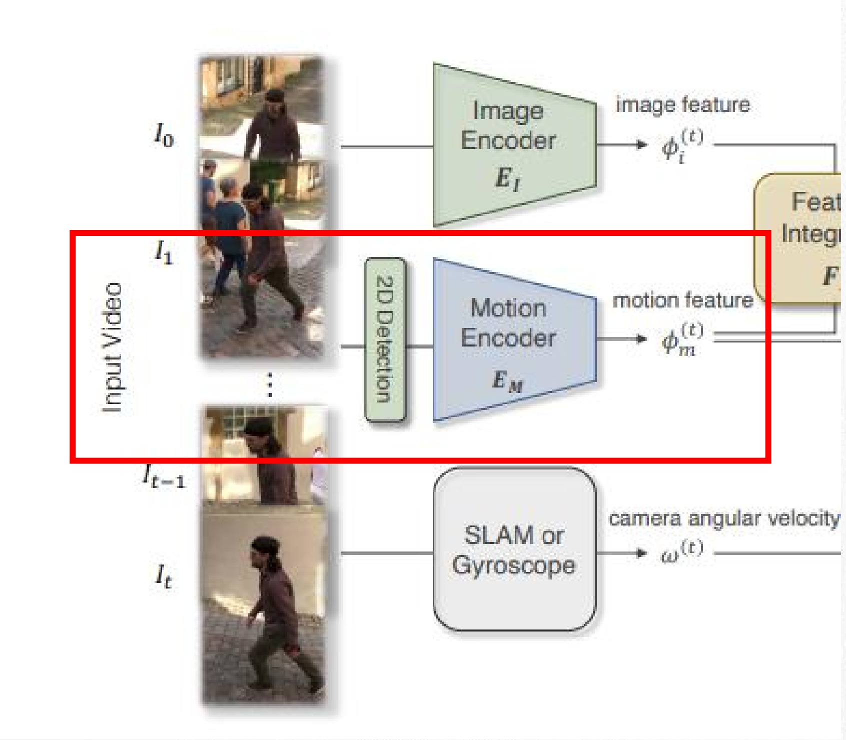
Method



Overview of WHAM

Method

Uni-directional Motion Encoder and Decoder:



Motion Encoder

$\{\boldsymbol{I}^{(t)}\}_{t=0}^{T}$: Input video data

$\{\boldsymbol{x}_{2D}^{(t)}\}_{t=0}^{T}$: 2D keypoints detected by ViTPose[1]

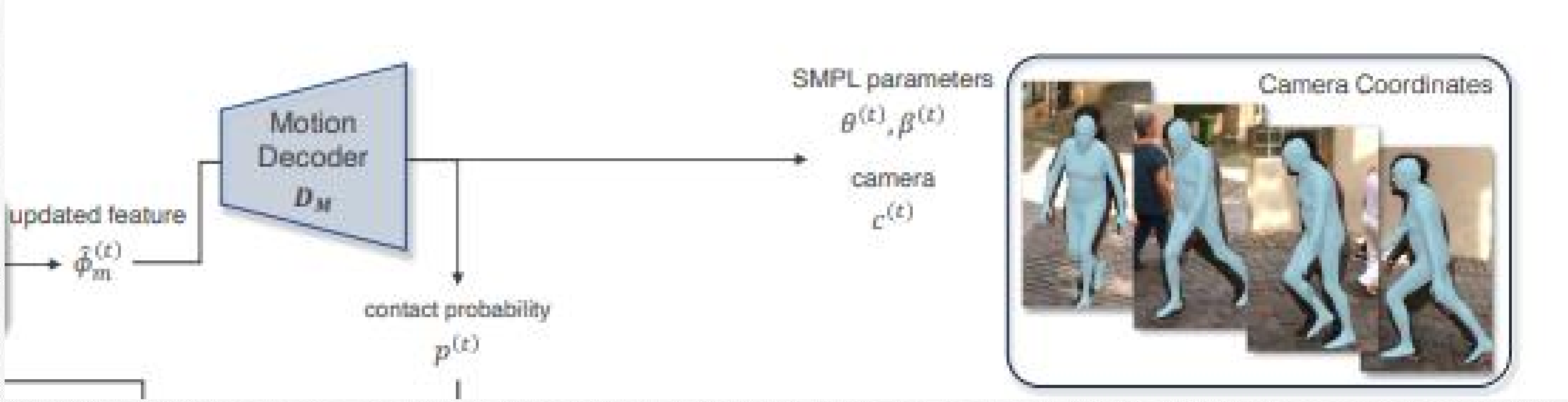
Motion features:

$$\phi_m^{(t)} = \boldsymbol{E}_M(\boldsymbol{x}_{2D}^{(0)}, \boldsymbol{x}_{2D}^{(1)}, \dots, \boldsymbol{x}_{2D}^{(t)} | \boldsymbol{h}_E^{(0)})$$

[1]:Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In Advances in Neural Information Processing Systems, 2022.

Method

Uni-directional Motion Encoder and Decoder:



Motion Decoder

$(\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)})$: SMPL parameters

Modeling parameters:

$$(\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{c}^{(t)}, \boldsymbol{p}^{(t)}) = \boldsymbol{D_M}(\hat{\phi}_m^{(0)}, \ldots, \hat{\phi}_m^{(t)} | \boldsymbol{h}_{\boldsymbol{D}}^{(0)})$$

$\boldsymbol{c}^{(t)}$: weak-perspective camera translation

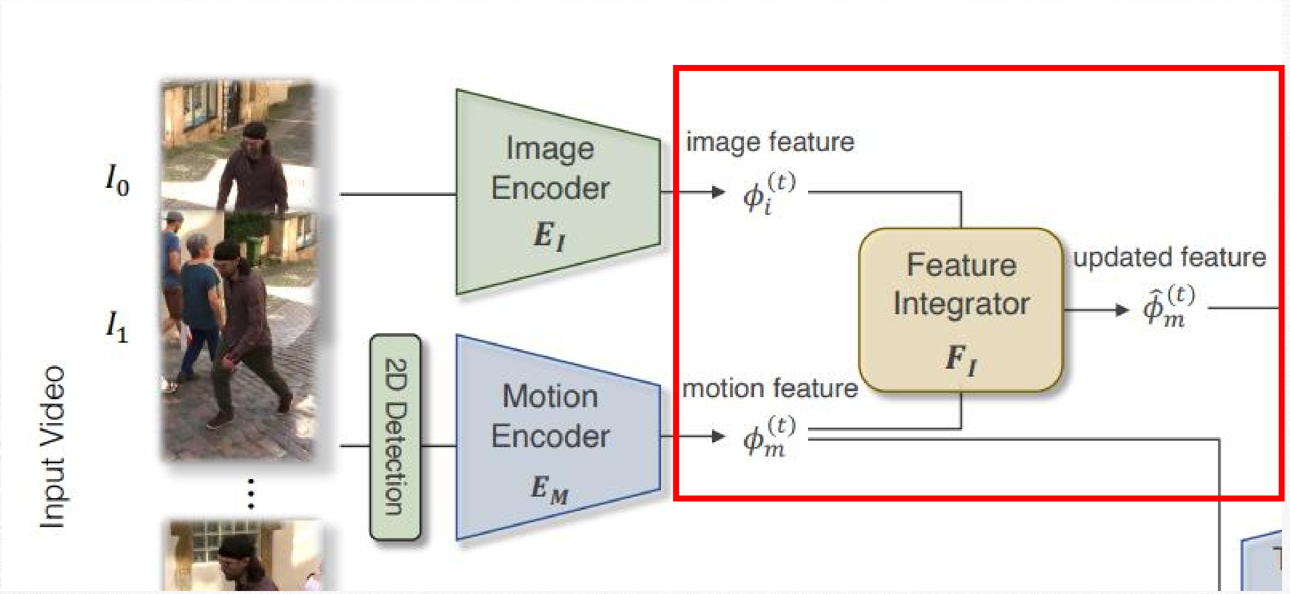
$\boldsymbol{p}^{(t)}$: foot-ground contact probability

Use Neural Initialization that uses MLP to initialize the hidden state, similar to PIP[2]

[2]:Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada,Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2022

Method

Motion and Visual Feature Integrator:



The feature of integration:

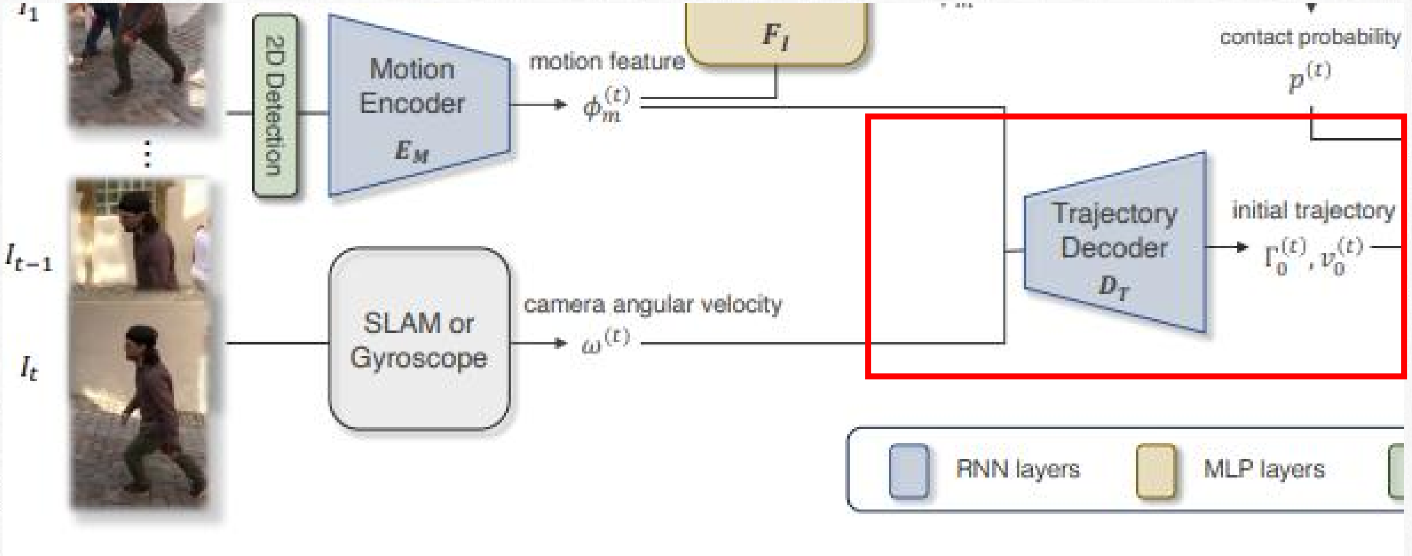
$$\hat{\phi}_m^{(t)} = \phi_m^{(t)} + \boldsymbol{F_I}(\boldsymbol{concat}(\phi_m^{(t)}, \phi_i^{(t)}))$$

$\phi_{\boldsymbol{i}}^{(t)}$: Image feature extracted by Image Encode $\boldsymbol{E_I}$

$\phi_{\boldsymbol{m}}^{(t)}$: Motion feature extracted by Motion Encode $\boldsymbol{E_M}$

Method

Global Trajectory Decode:



$$(\Gamma_0^{(t)}, v_0^{(t)}) = D_T(\phi_m^{(0)}, w^{(0)}, \ldots, \phi_m^{(t)}, w^{(t)})$$

$\phi_m^{(t)}$: Motion feature extracted by Motion Encode E_M

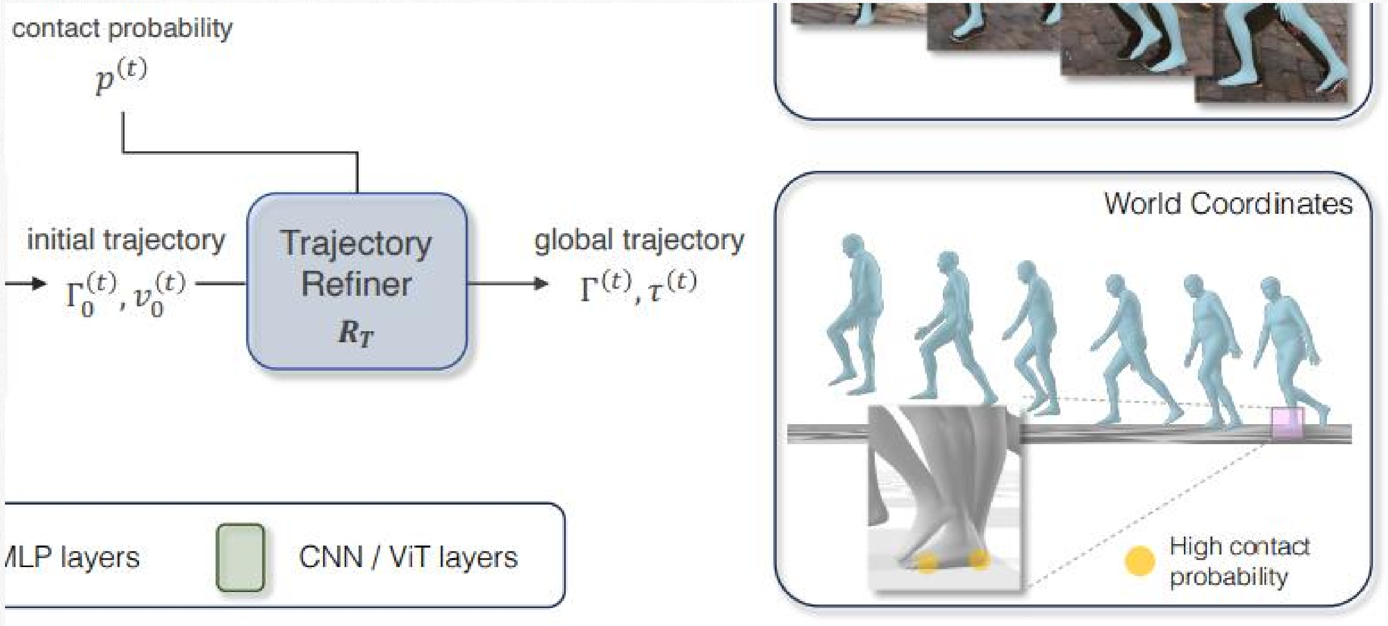
$w^{(t)}$: The angular velocity of the camera

$\Gamma_0^{(t)}$: The rough global root orientation

$v_0^{(t)}$: The root velocity

Method

Contact Aware Trajectory Refinement:



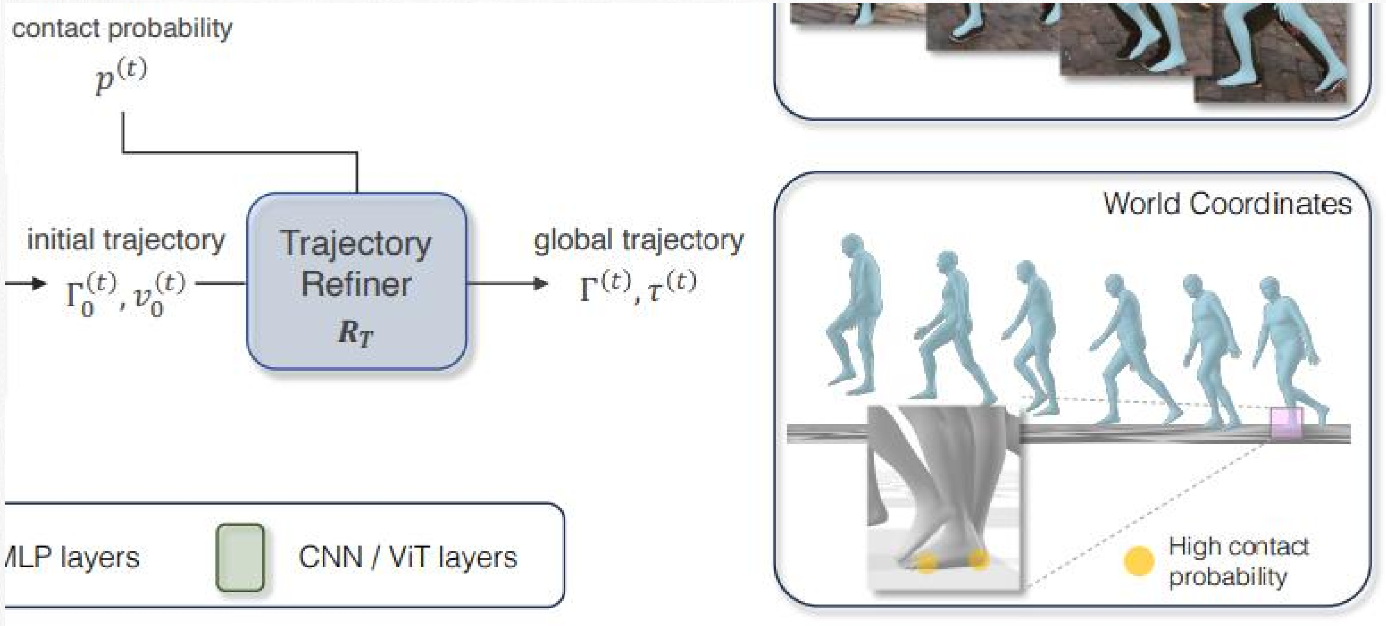
First, we adjust the ego-centric root velocity to $\tilde{\boldsymbol{v}}^{(t)}$ to minimize foot sliding

$$\tilde{\boldsymbol{v}}^{(t)} = \boldsymbol{v}_{\mathbf{0}}^{(t)} - (\Gamma_{\mathbf{0}}^{(t)})^{-1} \overline{\boldsymbol{v}}_{\boldsymbol{f}}^{(t)}$$

Where $\overline{\boldsymbol{v}}_{\boldsymbol{f}}^{(t)}$ is the averaged velocity of the toes and heels in world coordinates when their contact probability $\boldsymbol{p}^{(t)}$ is higher than a threshold.

Method

Contact Aware Trajectory Refinement:



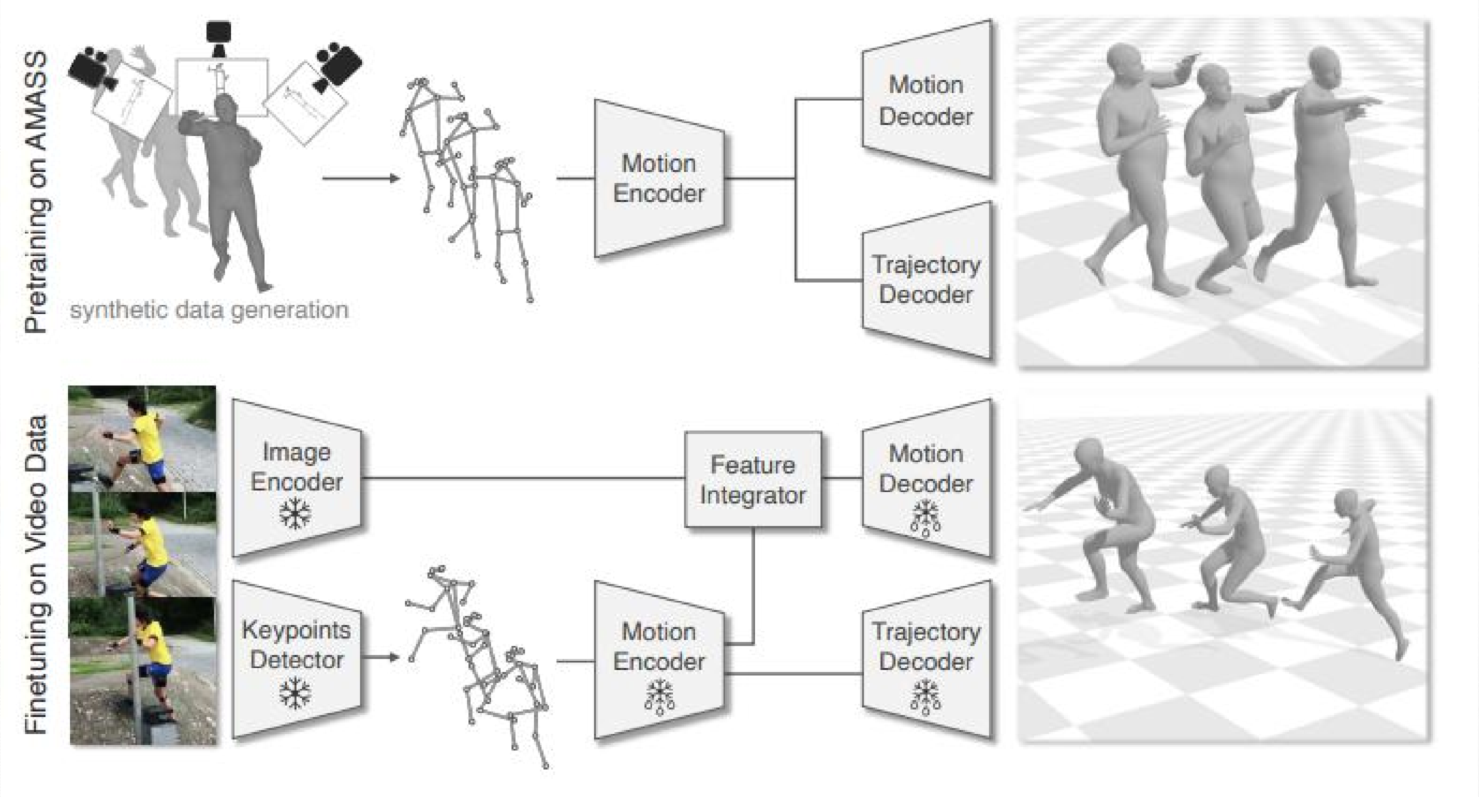
$$(\Gamma^{(t)}, \boldsymbol{v}^{(t)}) = \boldsymbol{R_T}(\phi_{\boldsymbol{m}}^{(\boldsymbol{0})}, \Gamma_{\boldsymbol{0}}^{(\boldsymbol{0})}, \widetilde{\boldsymbol{v}}^{(\boldsymbol{0})}, \ldots, \phi_{\boldsymbol{m}}^{(\boldsymbol{t})}, \Gamma_{\boldsymbol{0}}^{(\boldsymbol{t})}, \widetilde{\boldsymbol{v}}^{(\boldsymbol{t})})$$

$$\tau^{(\boldsymbol{t})} = \sum_{\boldsymbol{i}=\boldsymbol{0}}^{\boldsymbol{t}-\boldsymbol{1}} \Gamma^{(\boldsymbol{i})} \boldsymbol{v}^{(\boldsymbol{i})}$$

$\Gamma^{(t)}$: Global root orientation

$\tau^{(t)}$: Global root velocity

Method



WHAM’ s Two-Stage Training Scheme

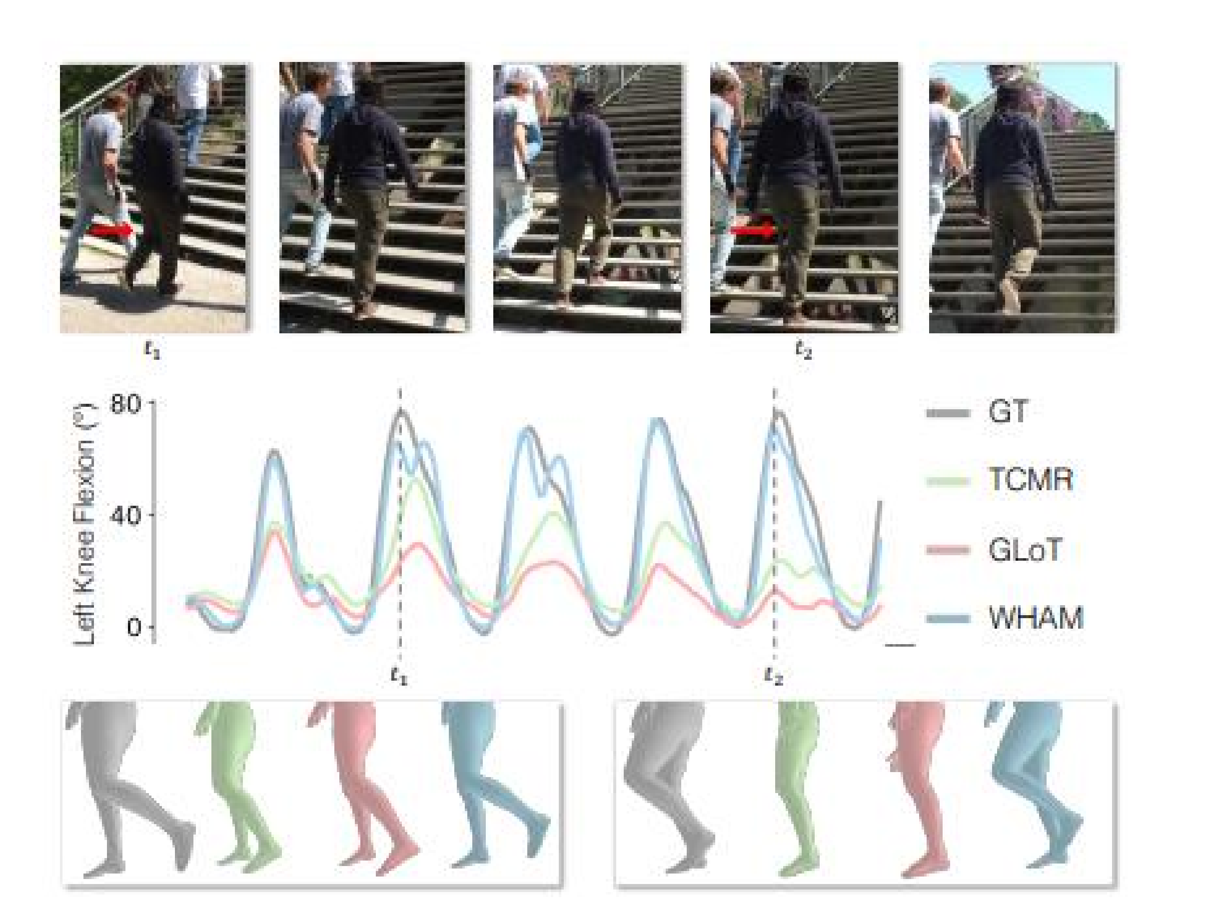
Experiments

Quantitative (3D Human Motion Recovery):

	Models	3DPW (14)				RICH (24)				EMDB (24)			
		PA-MPJPE	MPJPE	PVE	Accel	PA-MPJPE	MPJPE	PVE	Accel	PA-MPJPE	MPJPE	PVE	Accel
per-frame	SPIN [20]	59.2	96.9	112.8	31.4	69.7	122.9	144.2	35.2	87.1	140.7	166.1	41.3
	PARE* [17]	46.5	74.5	88.6	–	60.7	109.2	123.5	–	72.2	113.9	133.2	–
	CLIFF* [24]	43.0	69.0	81.2	22.5	56.6	102.6	115.0	22.4	68.3	103.5	123.7	24.5
	HybrIK* [22]	41.8	71.6	82.3	–	56.4	96.8	110.4	–	65.6	103.0	122.2	–
	HMR2.0 [6]	44.4	69.8	82.2	18.1	48.1	96.0	110.9	18.8	60.7	98.3	120.8	19.9
	ReFit* [49]	40.5	65.3	75.1	18.5	47.9	80.7	92.9	17.1	58.6	88.0	104.5	20.7
temporal	TCMR* [5]	52.7	86.5	101.4	6.0	65.6	119.1	137.7	5.0	79.8	127.7	150.2	5.3
	VIBE* [16]	51.9	82.9	98.4	18.5	68.4	120.5	140.2	21.8	81.6	126.1	149.9	26.5
	MPS-Net* [50]	52.1	84.3	99.0	6.5	67.1	118.2	136.7	5.8	81.4	123.3	143.9	6.2
	GLoT* [42]	50.6	80.7	96.4	**6.0**	65.6	114.3	132.7	5.2	79.1	119.9	140.8	5.4
	GLAMR [55]	51.1	–	–	8.0	79.9	–	–	107.7	73.8	113.8	134.9	33.0
	TRACE* [43]	50.9	79.1	95.4	28.6	–	–	–	–	71.5	110.0	129.6	25.5
	SLAHMR [53]	55.9	–	–	–	52.5	–	–	9.4	69.7	93.7	111.3	7.1
	PACE [19]	–	–	–	–	49.3	–	–	8.8	–	–	–	–
	WHAM (Res)*	40.2	62.7	75.1	6.3	51.8	89.5	103.2	**5.0**	57.8	84.0	99.7	**5.2**
	WHAM (HR)*	39.0	62.6	74.8	6.4	49.1	84.6	96.4	5.2	57.1	85.7	103.2	5.6
	WHAM (ViT)*	**35.9**	**57.8**	**68.7**	6.6	**44.3**	**80.0**	**91.2**	5.3	**50.4**	**79.7**	**94.4**	5.3

Experiments

Qualitative (3D Human Motion Recovery):



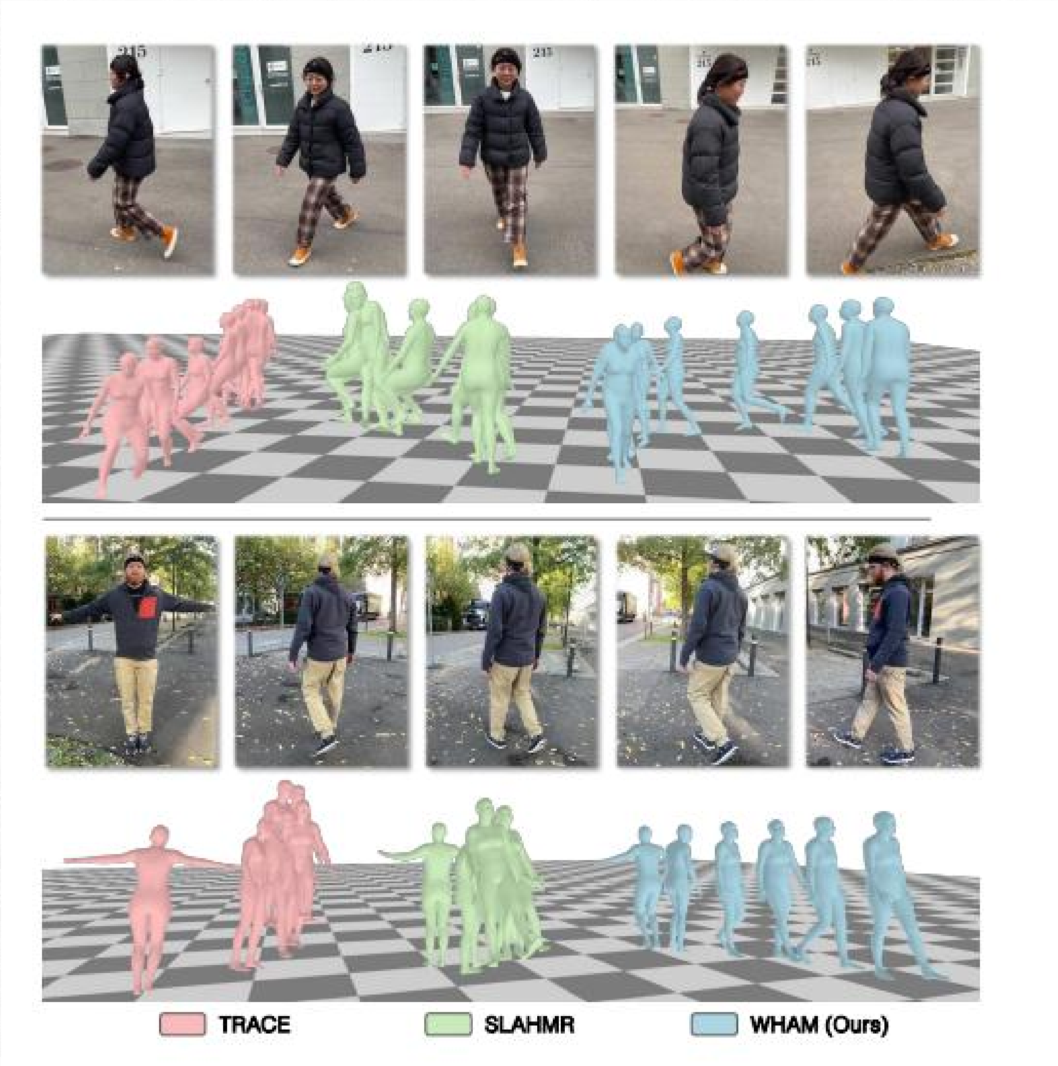
Experiments

Quantitative (3D Global Trajectory Recovery):

Models	EMDB 2				
	WA-MPJPE_{100}	W–MPJPE_{100}	RTE	Jitter	FS
DPVO (+ HMR2.0) [6, 45]	647.8	2231.4	15.8	537.3	107.6
GLAMR [55]	280.8	726.6	11.4	46.3	20.7
TRACE [43]	529.0	1702.3	17.7	2987.6	370.7
SLAHMR [53]	326.9	776.1	10.2	31.3	14.5
WHAM (w/DPVO [45])	135.6	354.8	6.0	22.5	4.4
WHAM (w/DROID [44])	133.3	343.9	4.6	21.5	**4.4**
WHAM (w/ GT gyro)	**131.1**	**335.3**	**4.1**	**21.0**	4.4

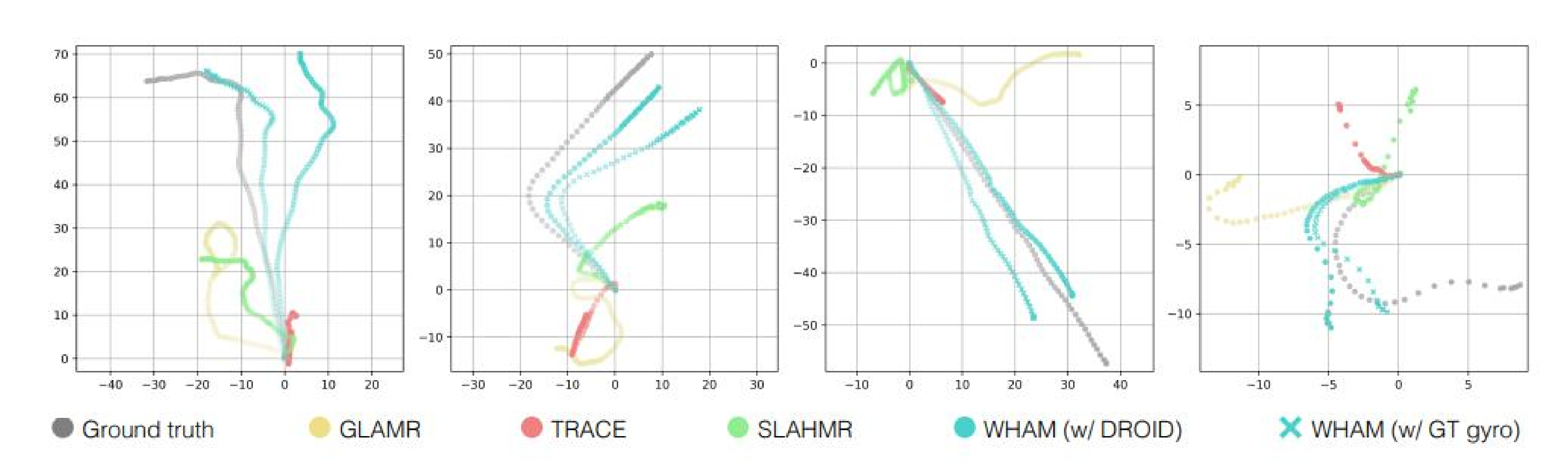
Experiments

Qualitative (3D Global Trajectory Recovery):



Experiments

Qualitative (3D Global Trajectory Recovery):



Experiments

Ablation Study

Models	EMDB 2						
	PA–MPJPE	MPJPE	WA–MPJPE$_{100}$	W–MPJPE$_{100}$	RTE	Jitter	FS
w/o F_I	44.2	69.0	147.6	377.9	6.3	23.1	5.5
w/o lifting	60.3	83.0	238.0	693.0	11.5	24.5	5.0
w/o NI	40.4	66.3	142.7	368.1	6.8	22.3	4.6
w/o ω	39.1	62.0	156.5	422.0	10.1	22.1	5.0
w/o traj. ref.	**38.2**	**59.3**	154.7	407.5	6.3	**18.8**	6.5
WHAM (Ours)	**38.2**	**59.3**	**135.6**	**354.8**	**6.0**	22.5	**4.4**

Thank you for watching!