Egocentric Interaction Capture for Mixed Reality


Siwei Zhang[1]


Qianli Ma[1]


Yan Zhang[1]


Zhiyin Qian[1]


Taein Kwon[1]


Federica Bogo[2]


Marc Pollefeys[1,2]


Siyu Tang[1]

1 Computer Vision and Learning Group, ETH Zurich
2 Microsoft

ETH zürich

VLG Computer Vision and Learning Group

Microsoft

Swiss JRC Workshop
March 30, 2022

Video from Meta AI

Video from Microsoft Mixed Reality

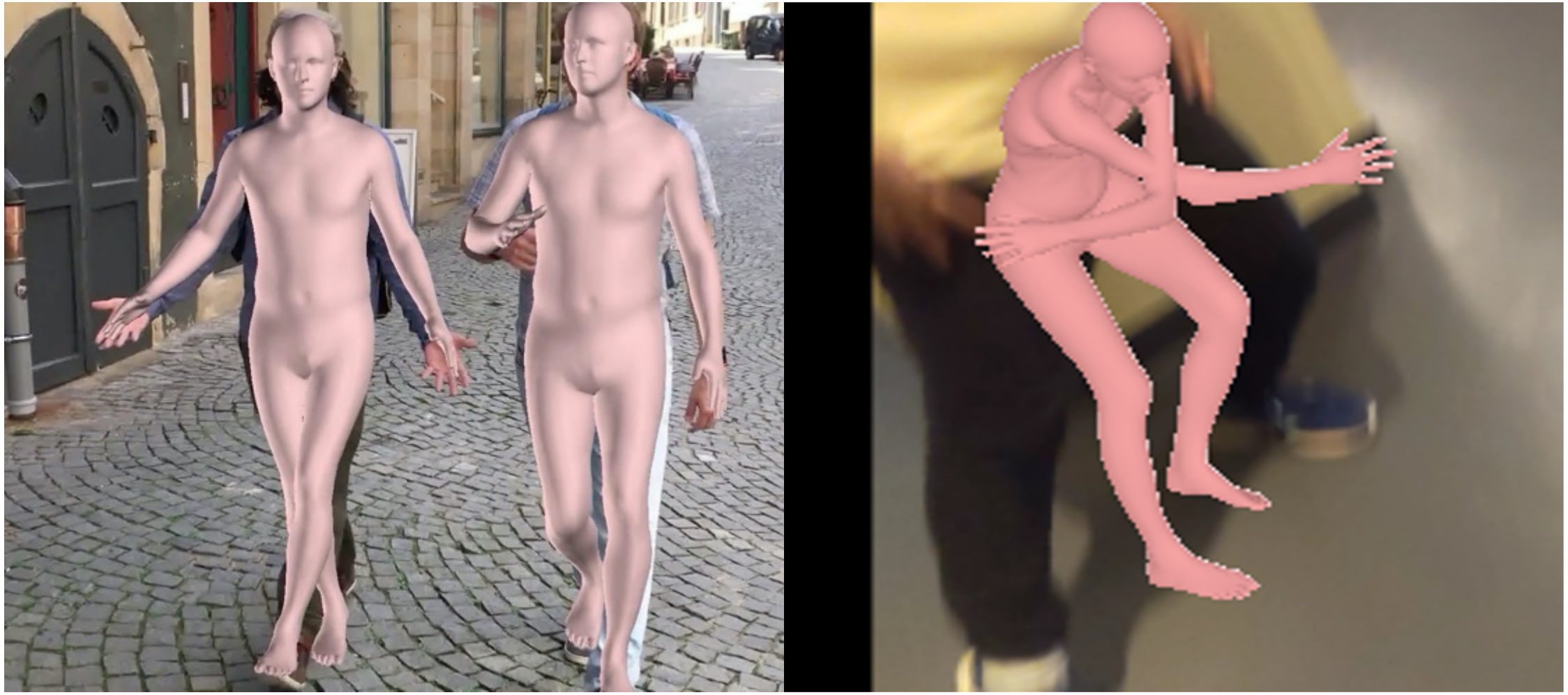
Third-person View

First-person / Egocentric View

Video from Meta AI

Third-person View

First-person / Egocentric View

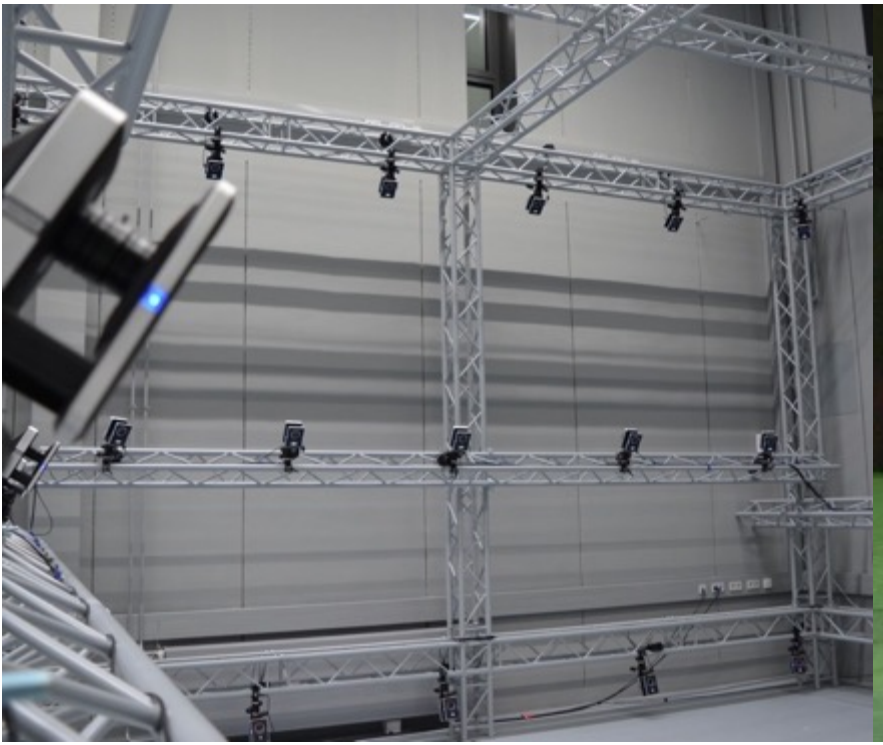


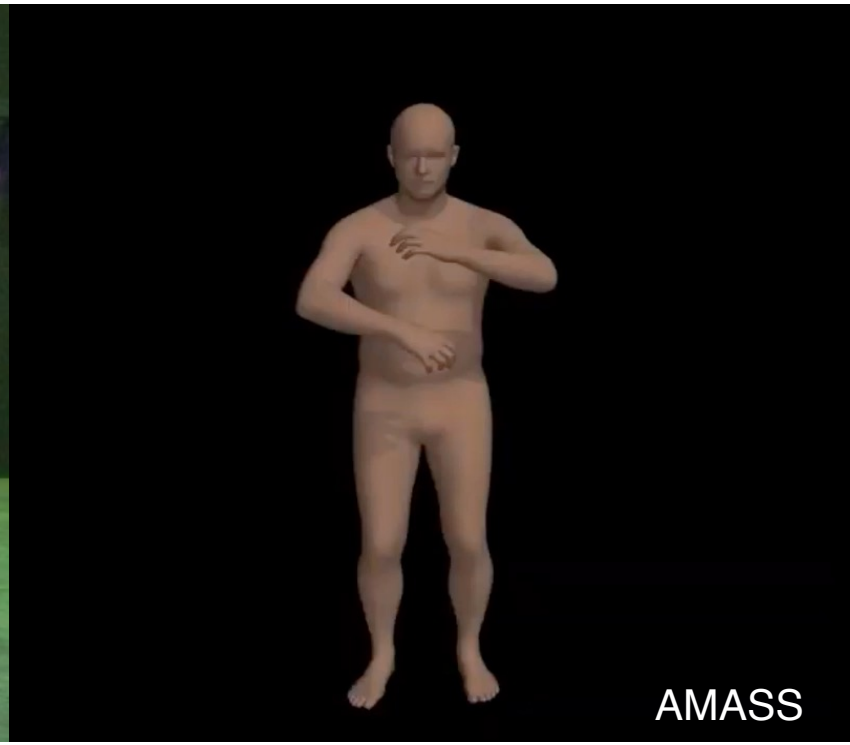
SPIN (Kolotouros et al.)


Egocentric View

How to capture 3D human pose, shape and motions in 3D scenes?

Marker-based Motion Caption System

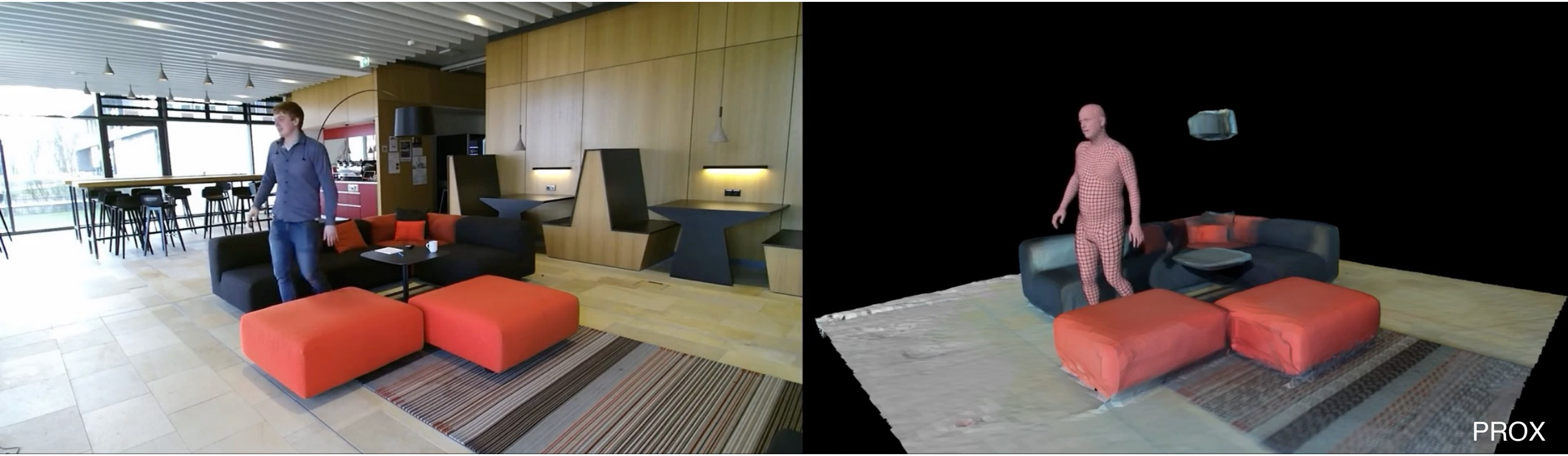




Expensive setup / expert knowledge

[1] https://ps.is.mpg.de/pages/motion-capture
[2] https://sentimentalflow.wordpress.com/2017/01/30/first-blog-post/
[3] AMASS: Archive of Motion Capture as Surface Shapes, Mahmood et al, ICCV 2019

Monocular RGB-D Kinect Setting



Lightweight, noisy motion reconstructions

[1] Hassan et al., Resolving 3D Human Pose Ambiguities with 3D Scene Constraints, 2019 ICCV

Monocular RGB-D Kinect Setting

How to reconstruct natural and smooth human motions in 3D scenes with a **lightweight setup**?

Lightweight, noisy motion reconstructions

[1] Hassan et al., Resolving 3D Human Pose Ambiguities with 3D Scene Constraints, 2019 ICCV

Outline

- **LEMO: Learning Motion Priors for 4D Human Body Capture in 3D Scenes**

  Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, Siyu Tang
  ICCV 2021, Oral presentation

- **EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices**

  Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, Siyu Tang

Outline

- **LEMO: Learning Motion Priors for 4D Human Body Capture in 3D Scenes**

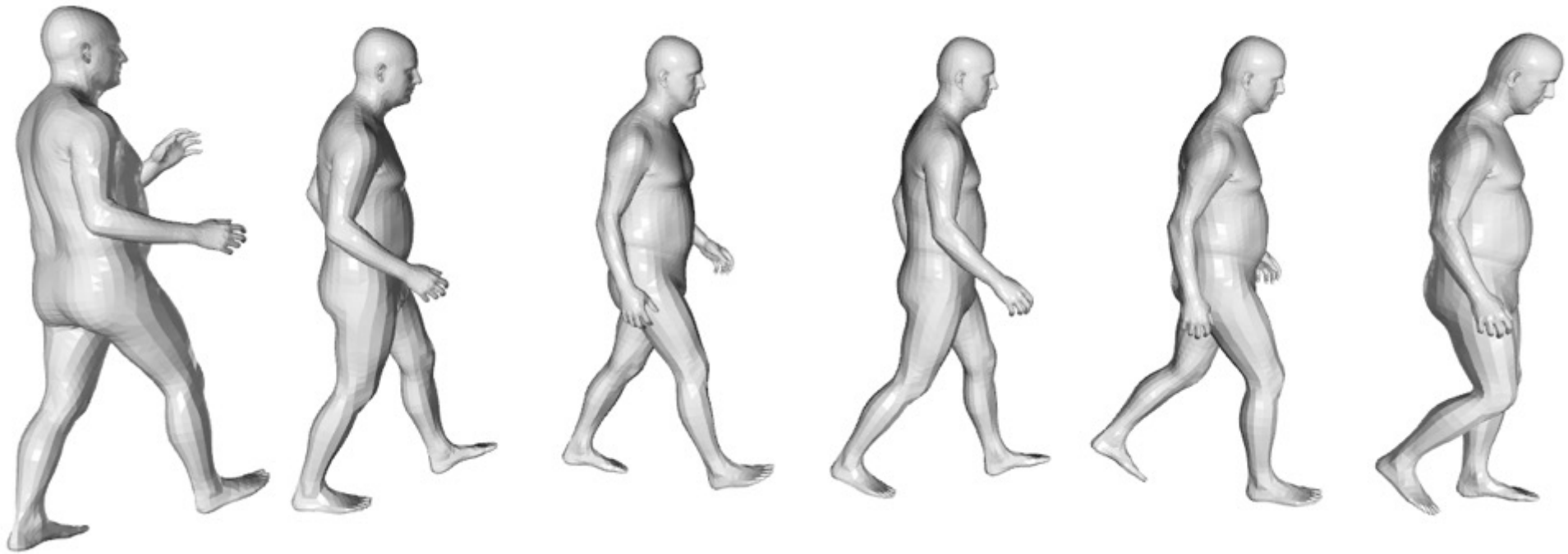
  Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, Siyu Tang
  ICCV 2021, Oral presentation

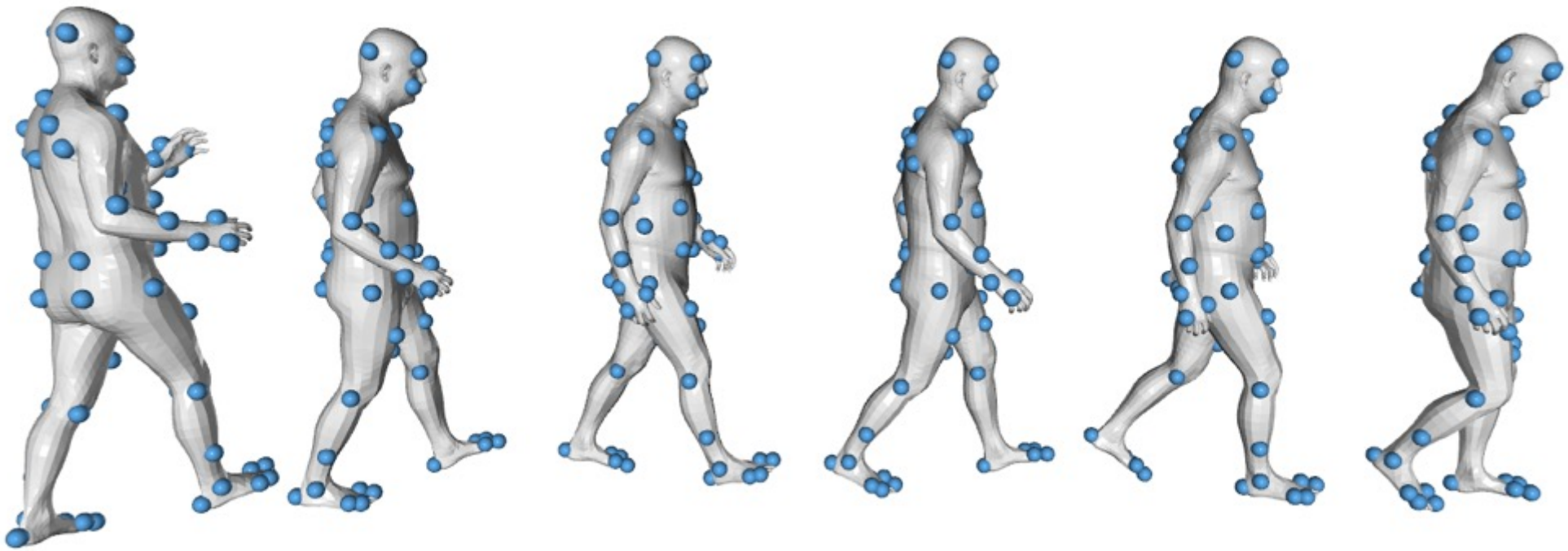
- **EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices**

  Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, Siyu Tang

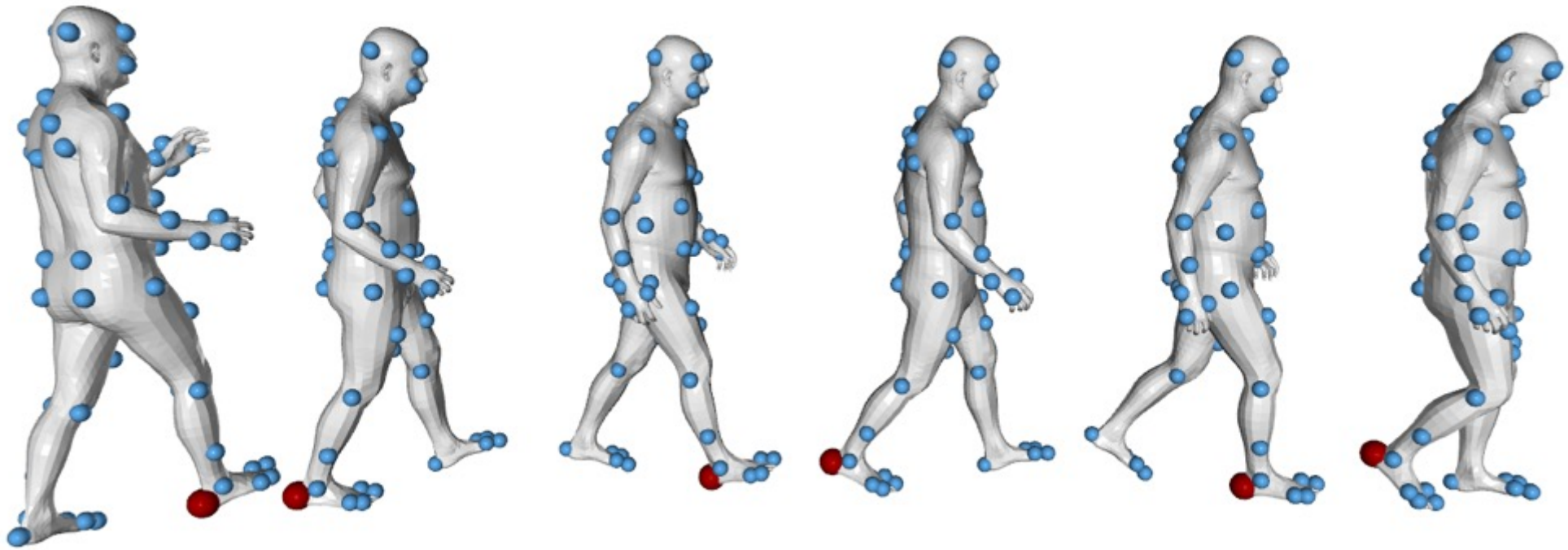
How to reconstruct natural and smooth human motions in 3D scenes
with a **lightweight setup**?

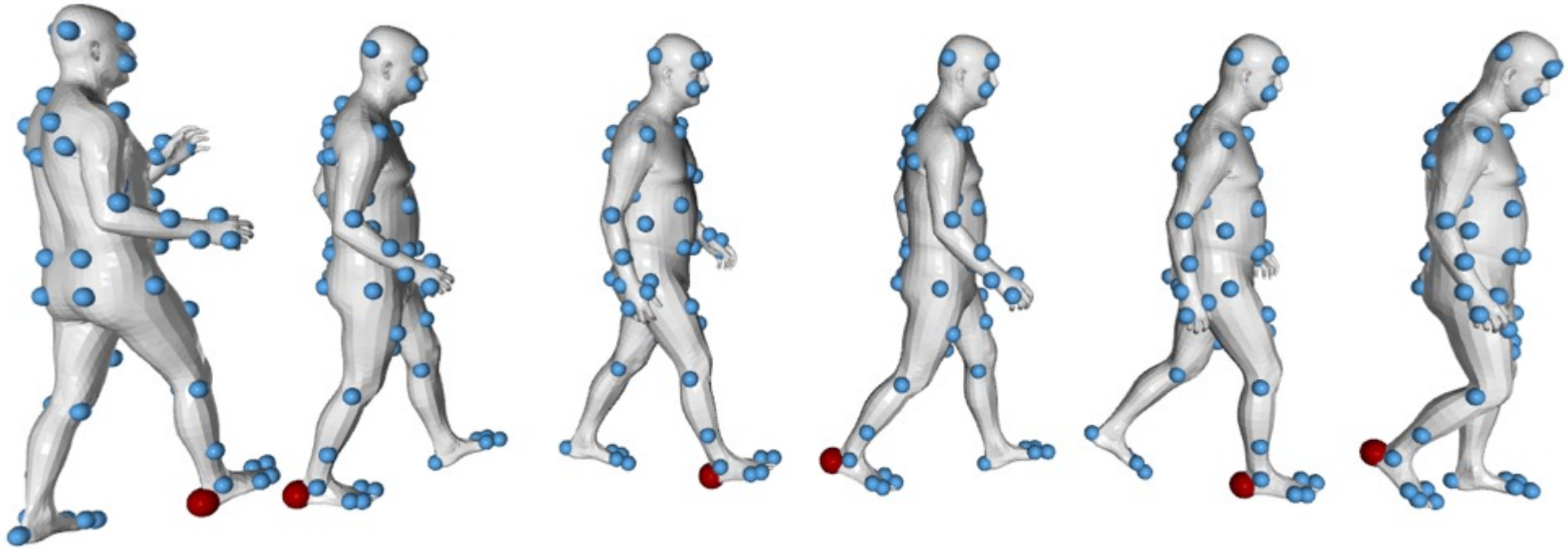
Key insight: learning motion priors from the high quality mocap dataset

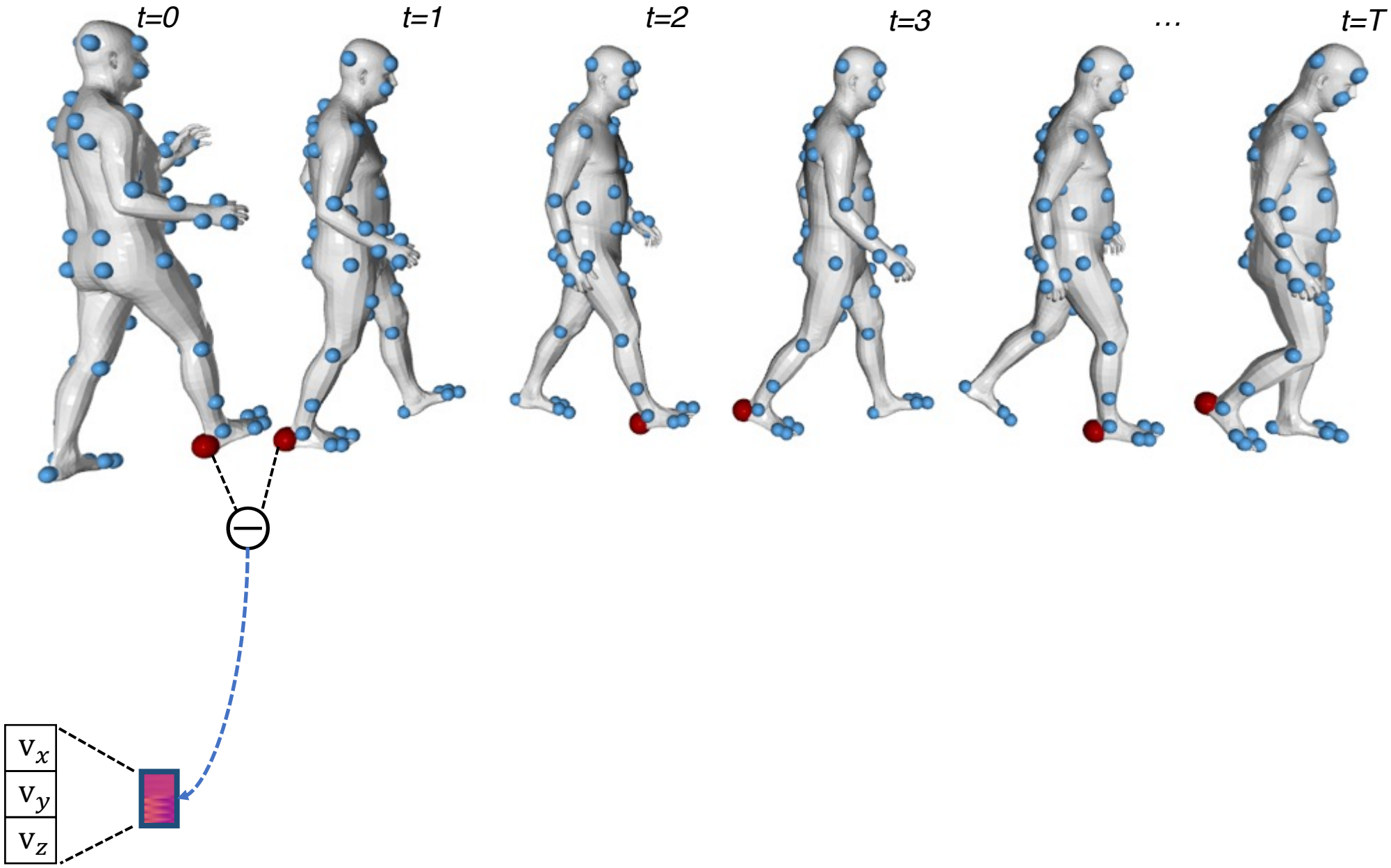


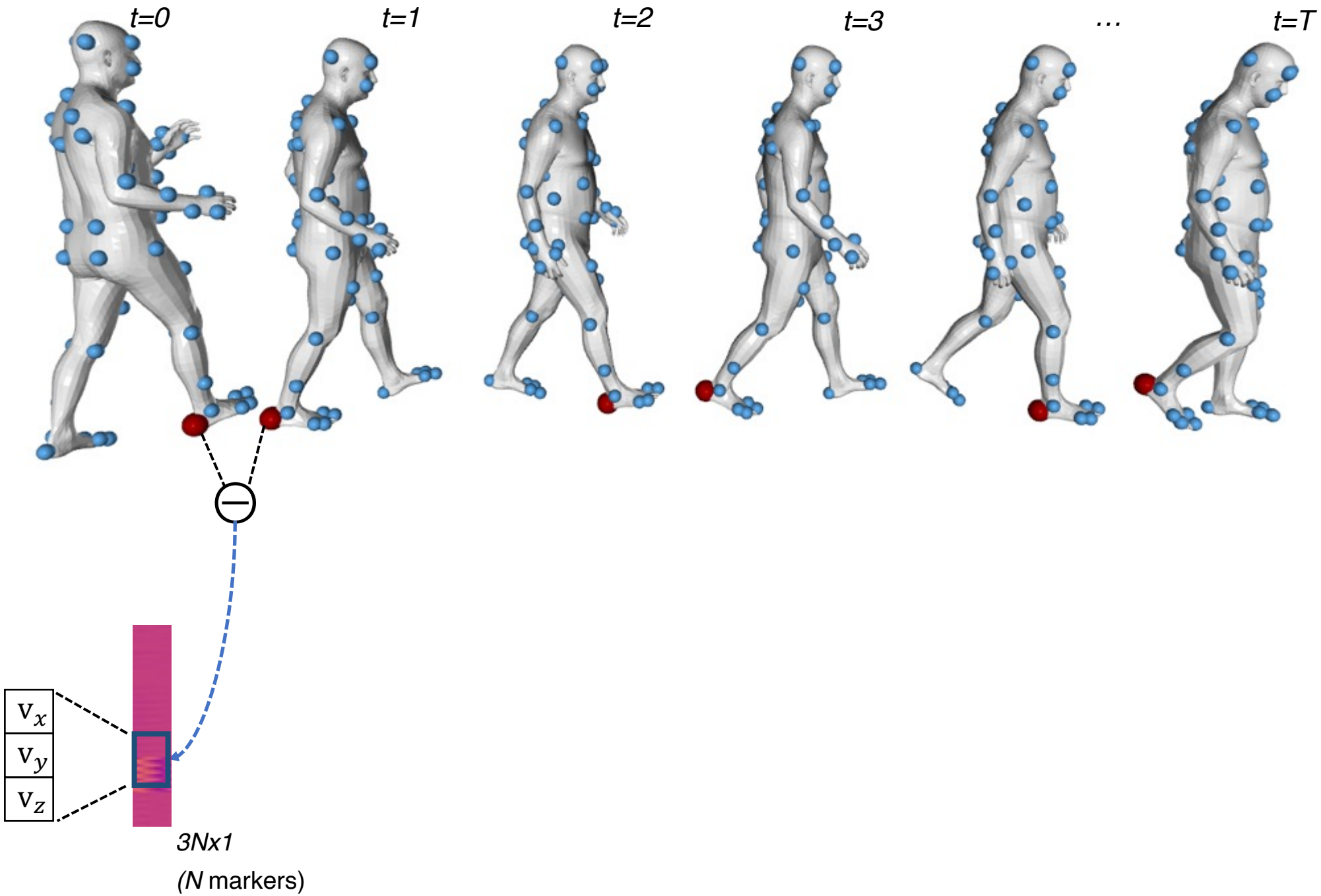


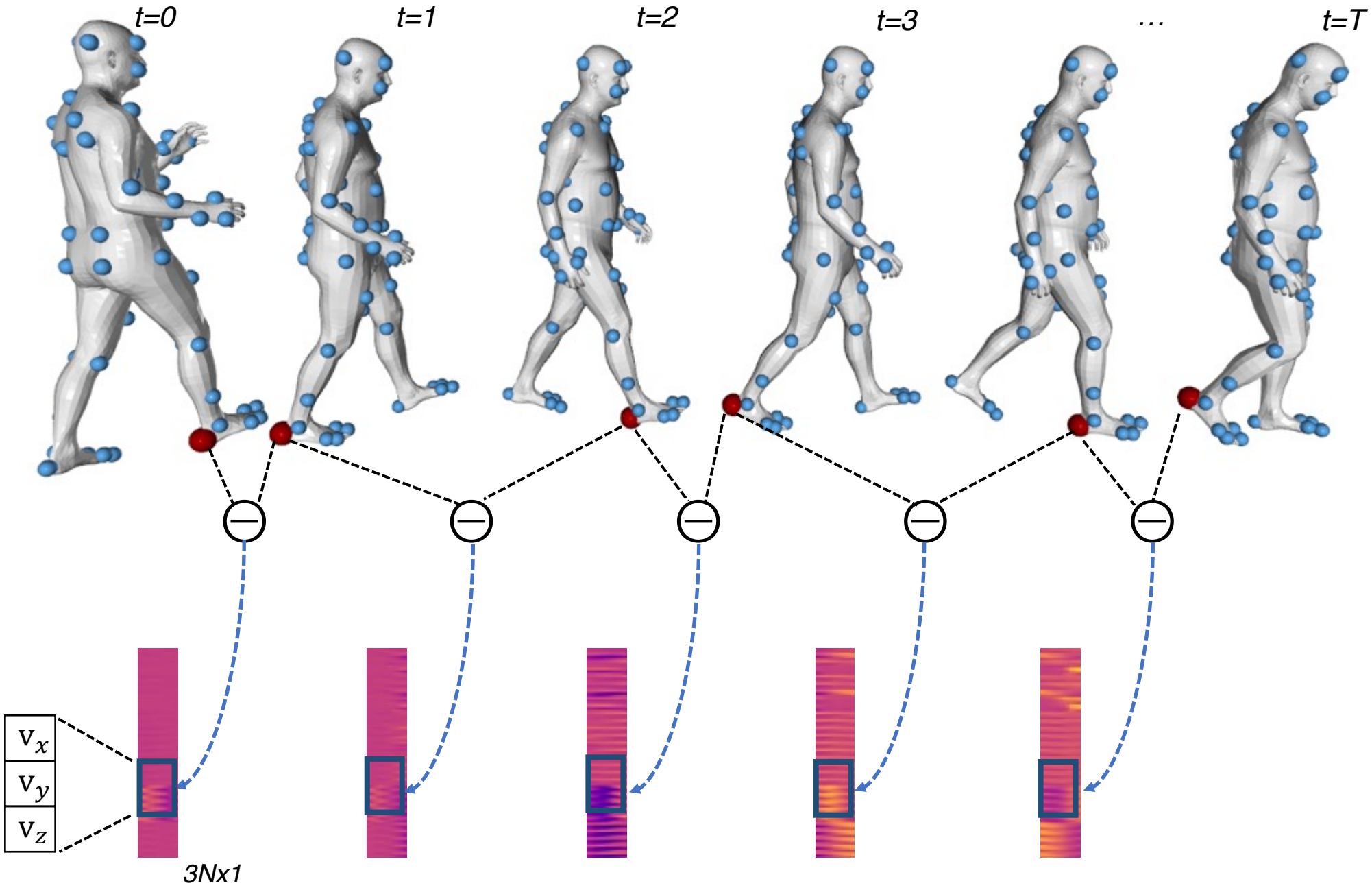
[1] We are More than Our Joints: Predicting how 3D Bodies Move, Zhang et al, CVPR 2021

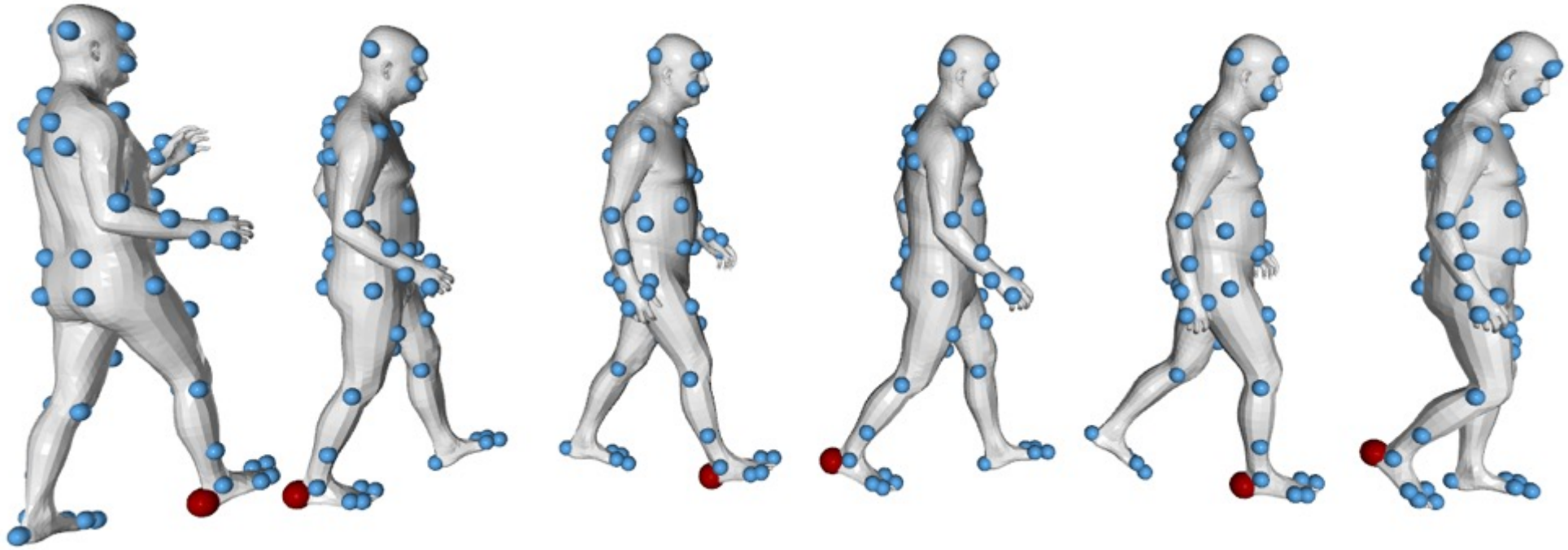


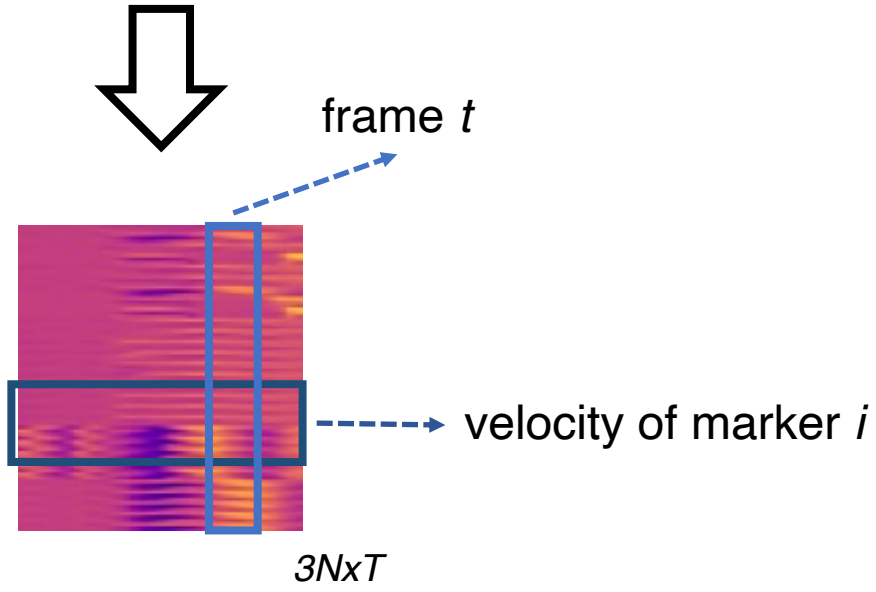


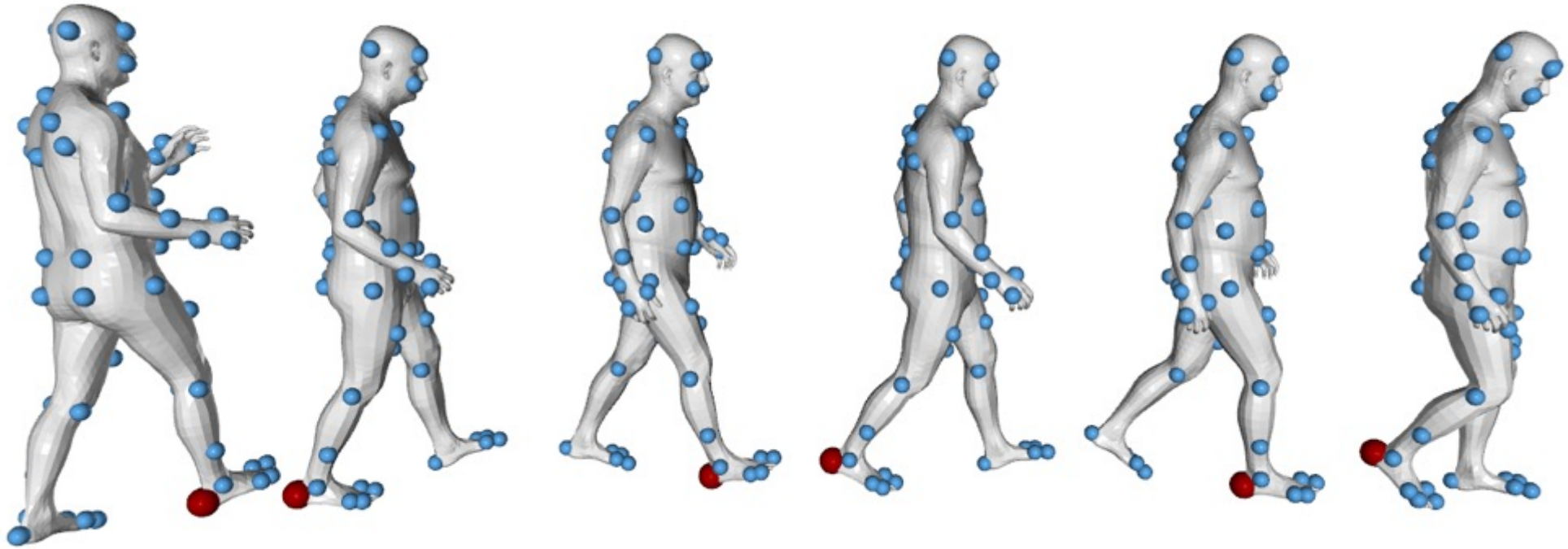

t=0
t=1
t=2
t=3
…
t=T
v_x
v_y
v_z

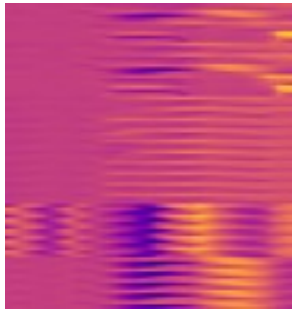

t=0
t=1
t=2
t=3
…
t=T
v_x
v_y
v_z
3Nx1
(N markers)

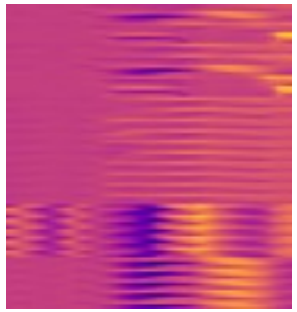

t=0
t=1
t=2
t=3
…
t=T
v_x
v_y
v_z
3Nx1

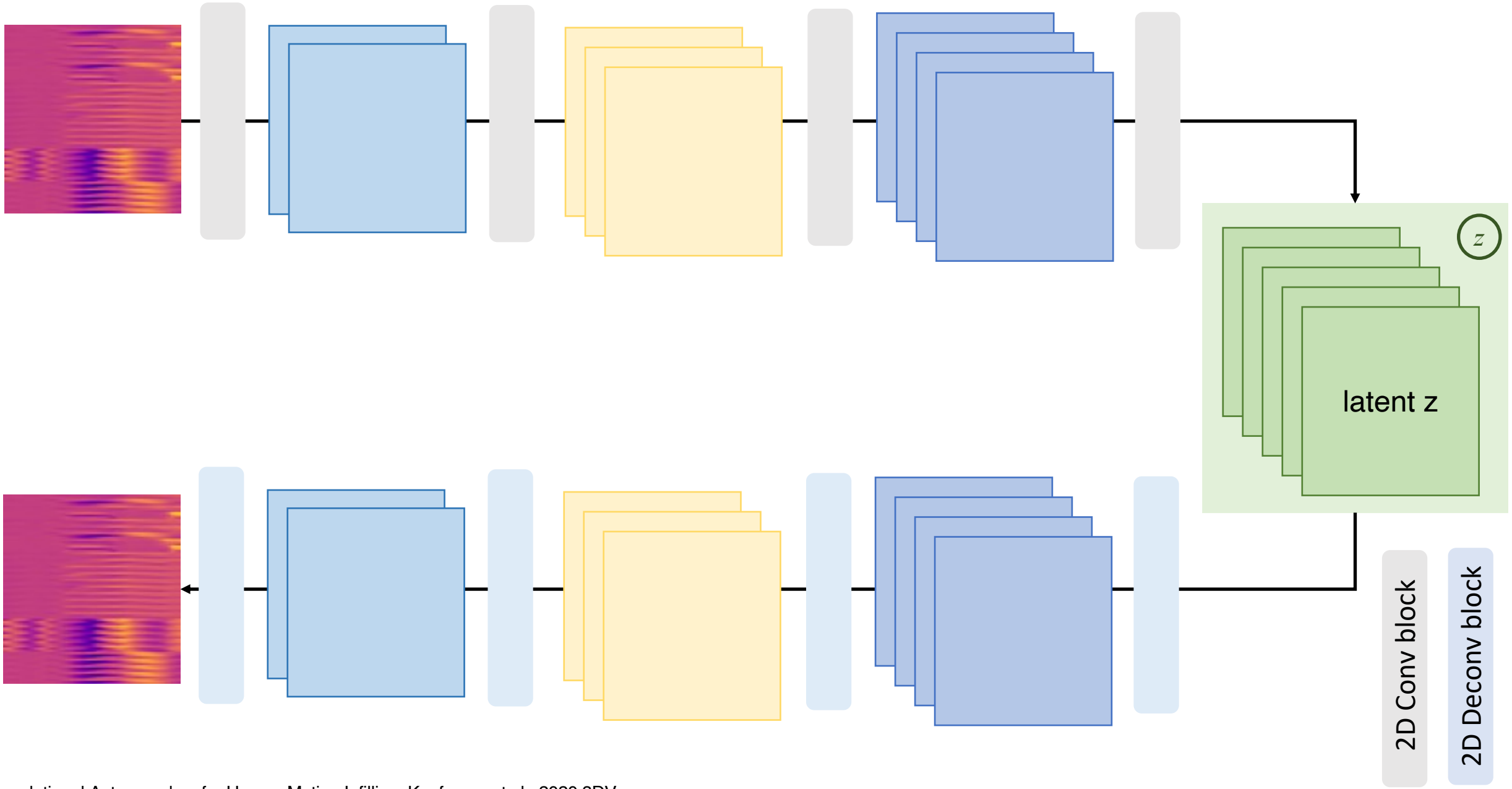



frame *t*
velocity of marker *i*
3NxT

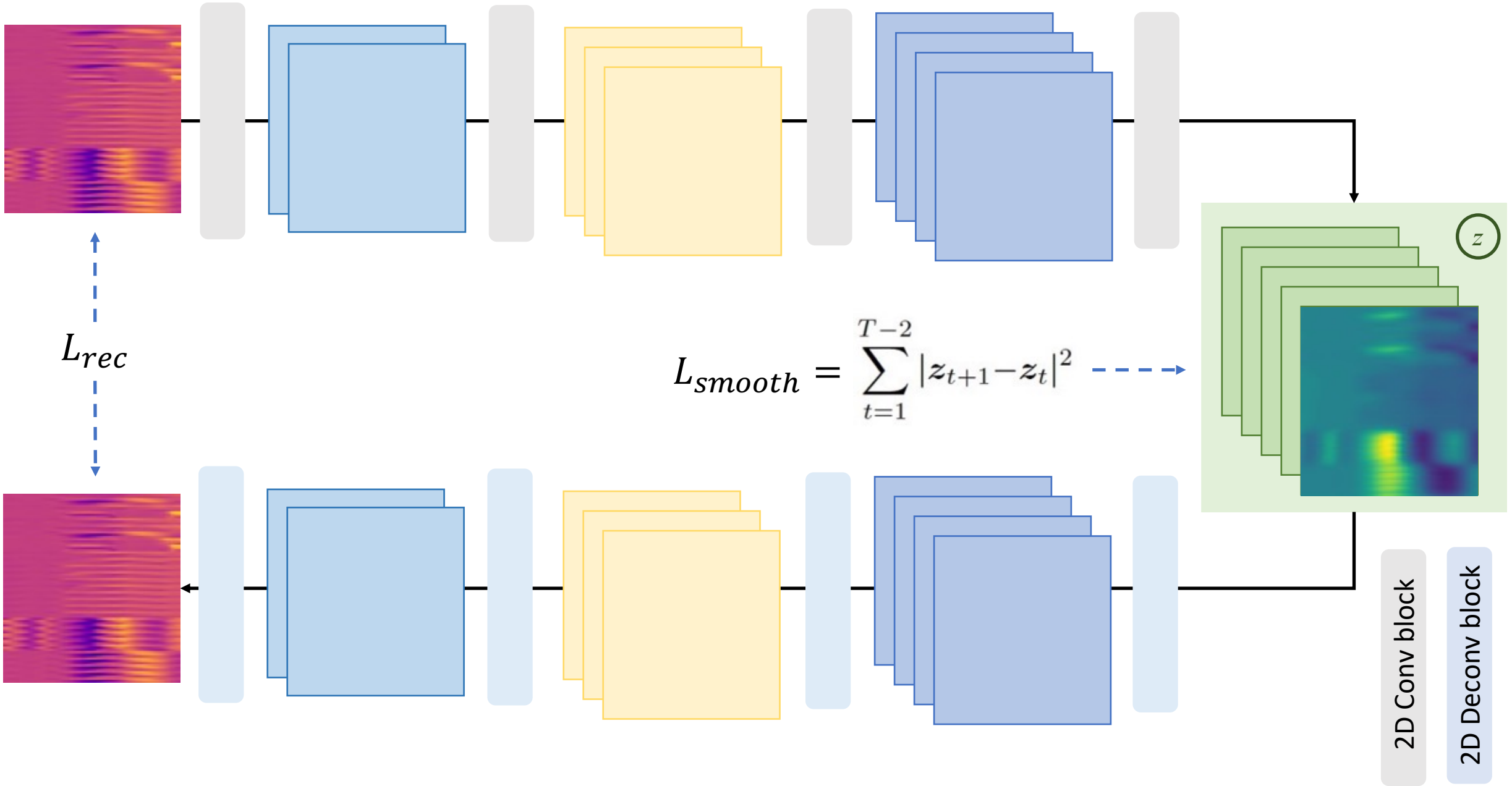


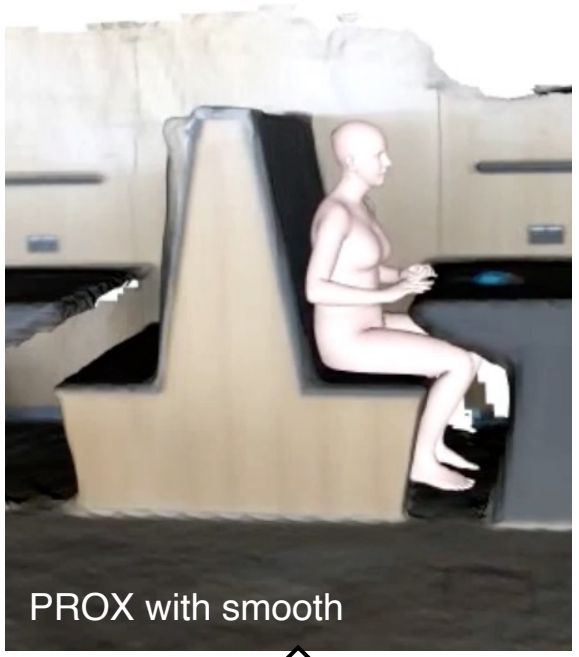


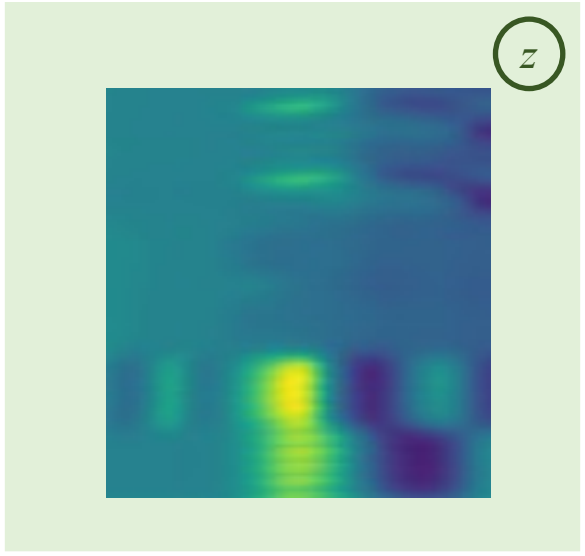


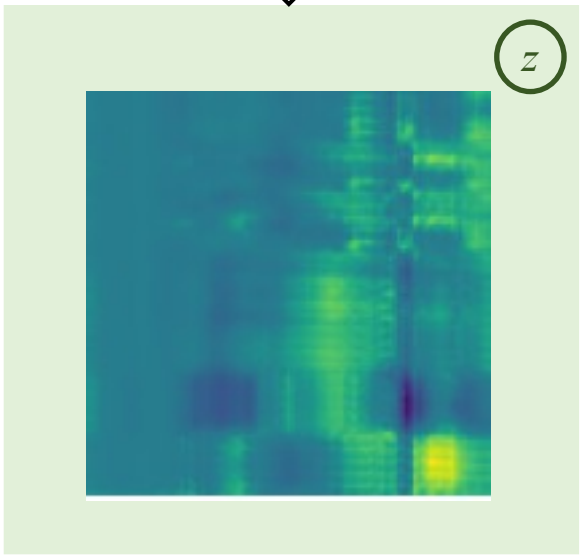


[1] Convolutional Autoencoders for Human Motion Infilling, Kaufmann et al., 2020 3DV

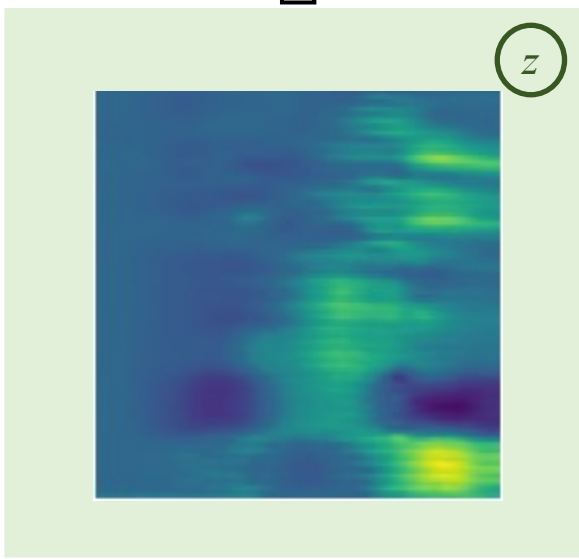

L_{rec}
$L_{smooth} = \sum_{t=1}^{T-2} |z_{t+1} - z_t|^2$
z
2D Conv block
2D Deconv block

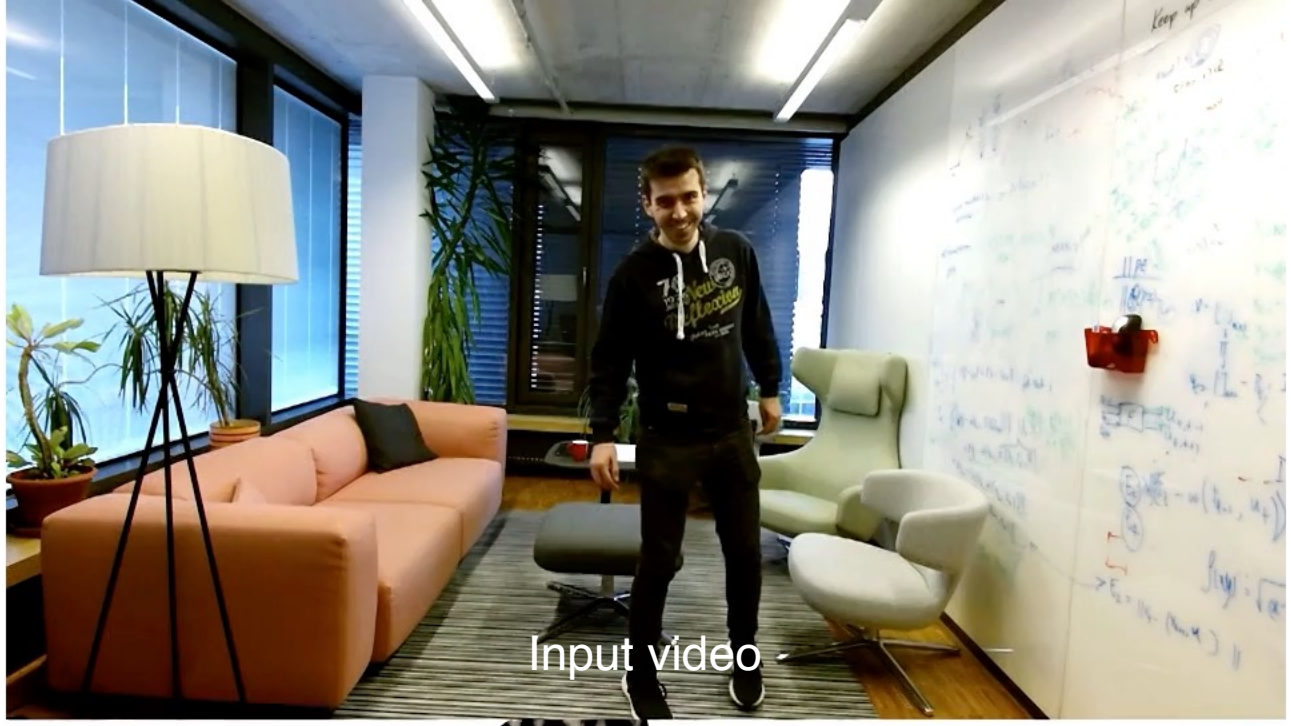






$$\min \sum_{t=1}^{T-2} |z_{t+1} - z_t|^2$$




Input video


PROX (Hassan et al.)


LEMO (Ours)

Outline

- **LEMO: Learning Motion Priors for 4D Human Body Capture in 3D Scenes**

  Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, Siyu Tang
  ICCV 2021, Oral presentation

- **EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices**

  Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, Siyu Tang

Introducing: EgoBody Dataset

Third-Person Views,
taken with Kinects

Egocentric View,
taken with HoloLens2

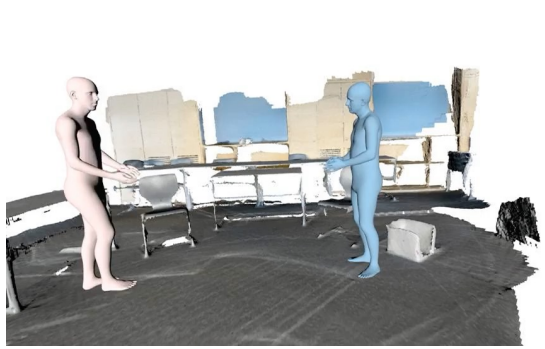
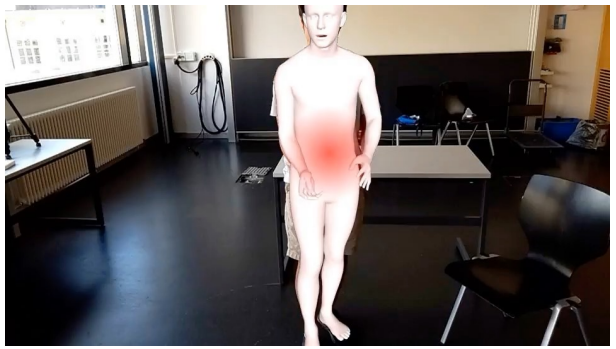
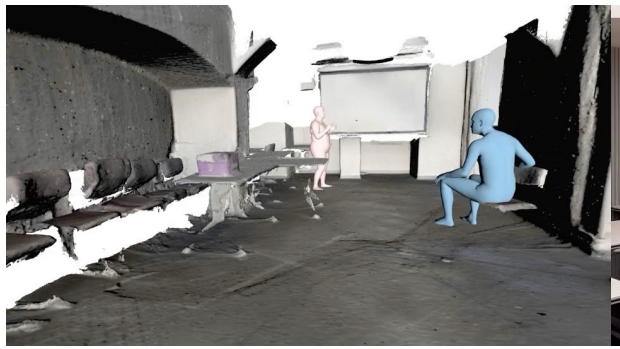
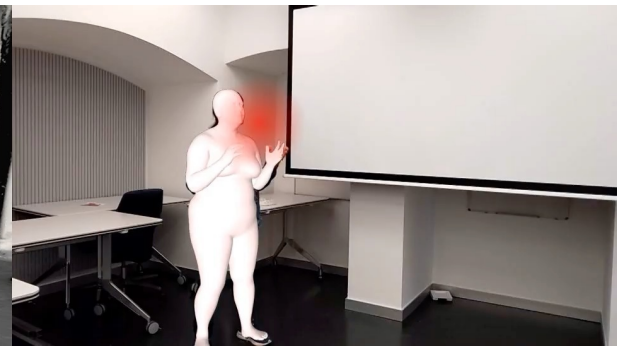
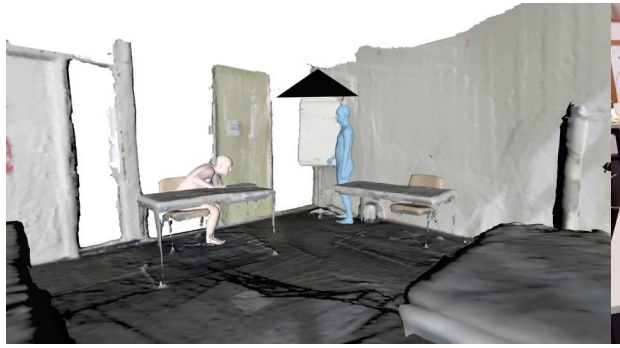
Introducing: EgoBody Dataset

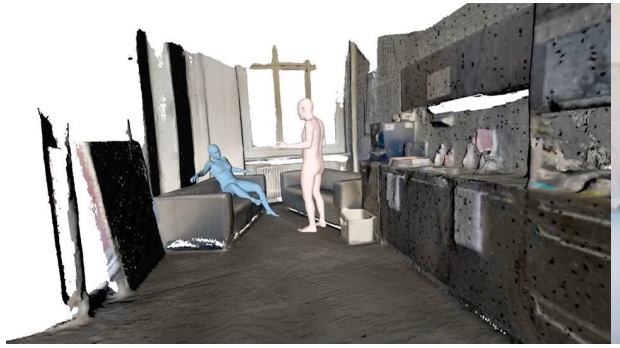
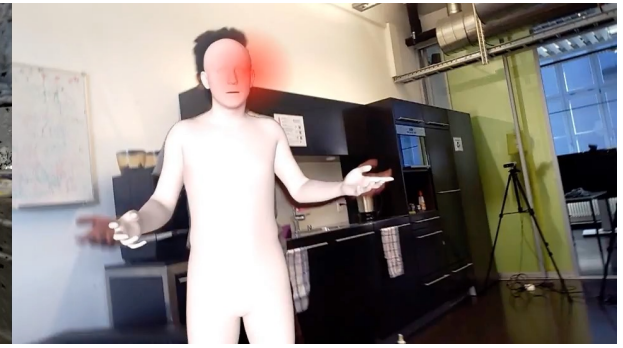
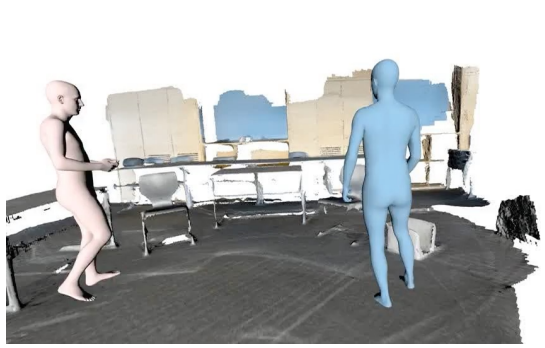
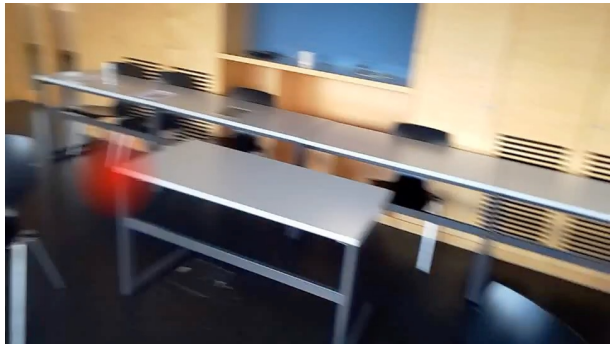
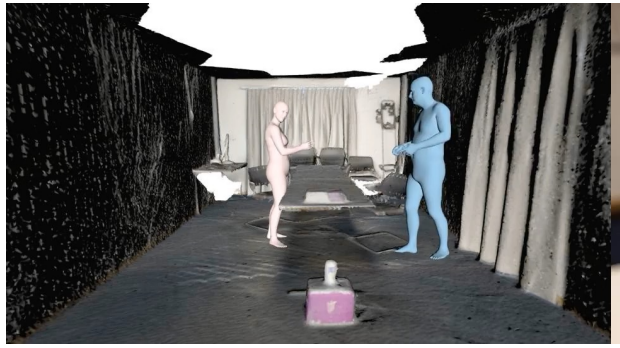
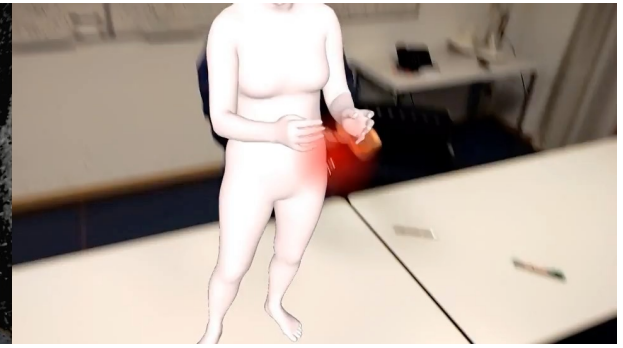
Third-Person View

+ ground truth body annotation of
the camera wearer & the second person
+ scene reconstruction

Egocentric View

+ ground truth body annotation
+ eye gaze / attention

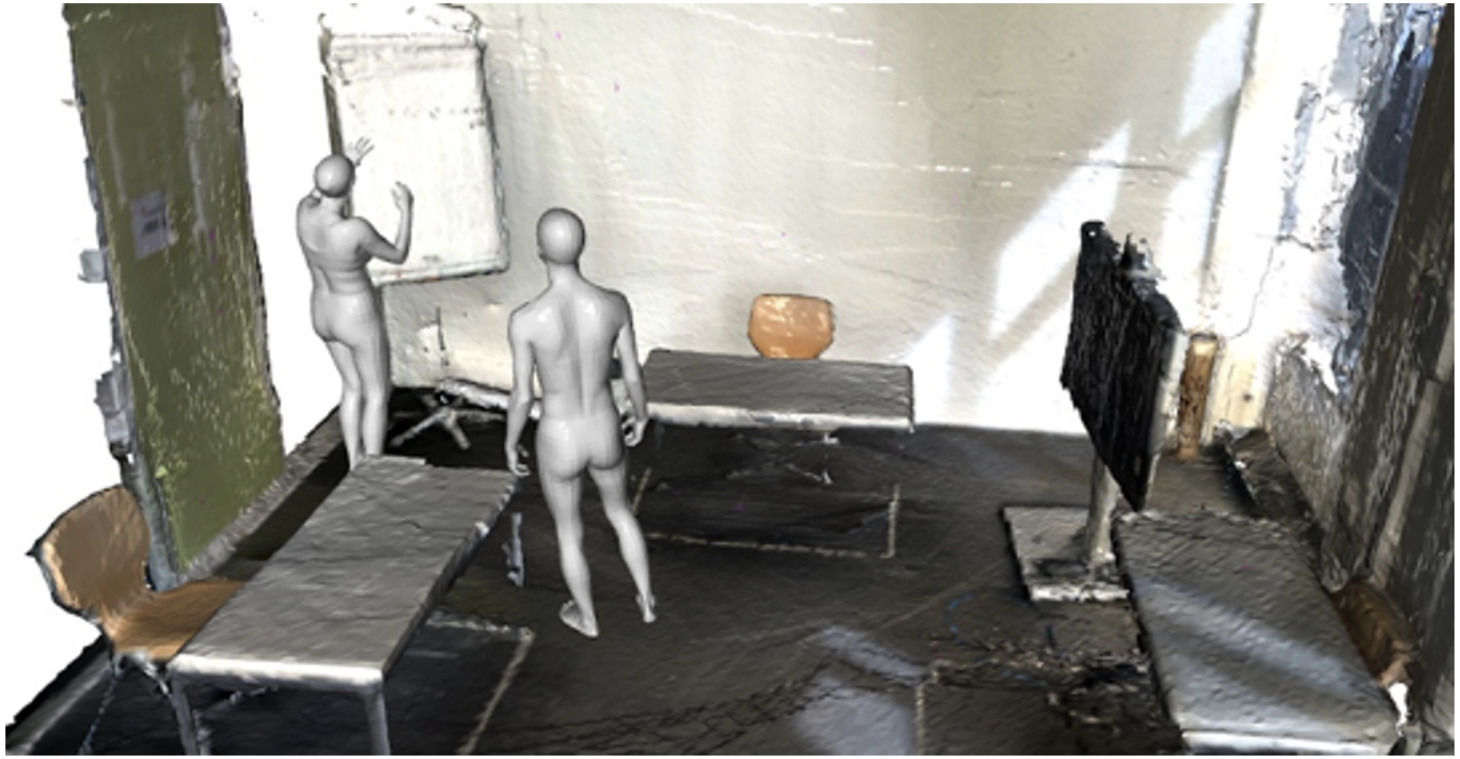






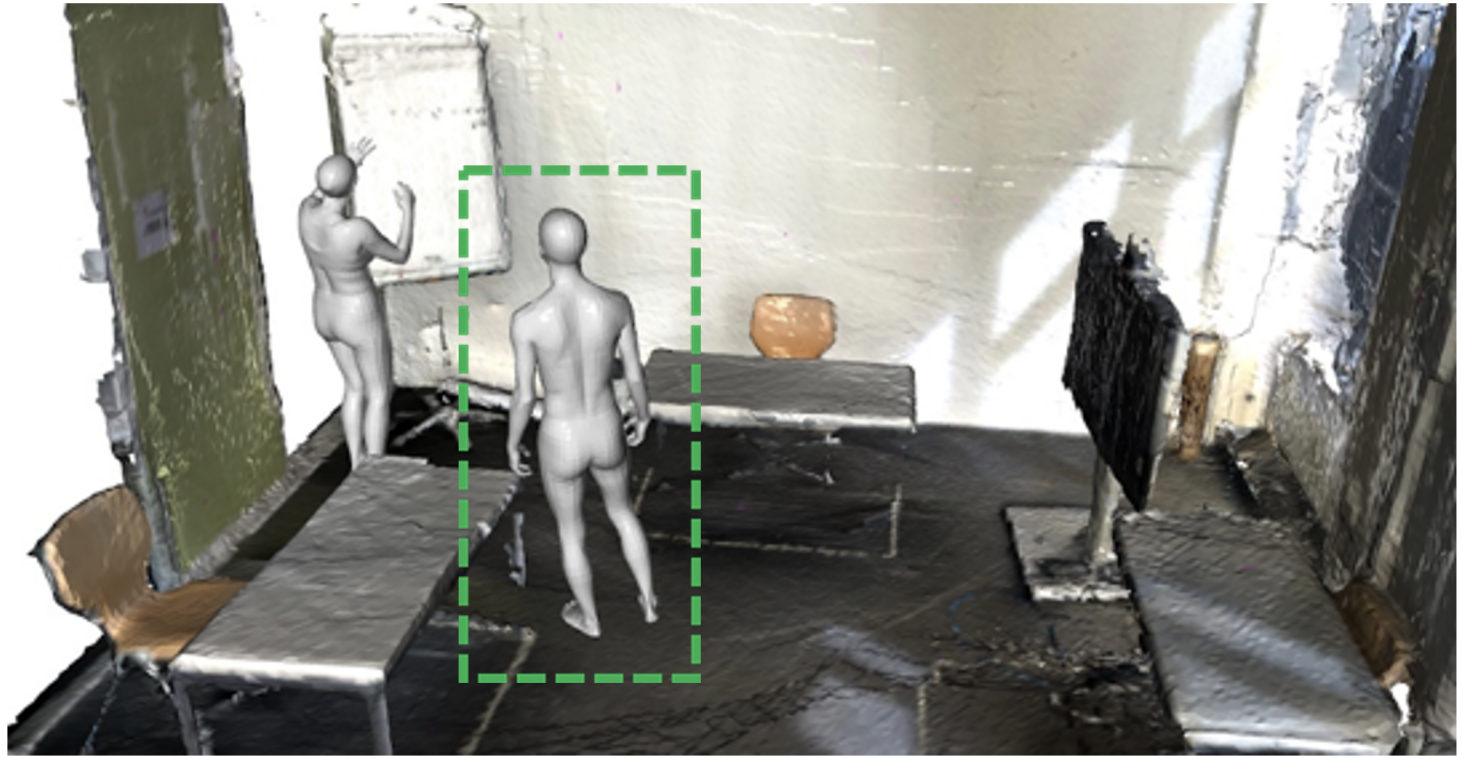







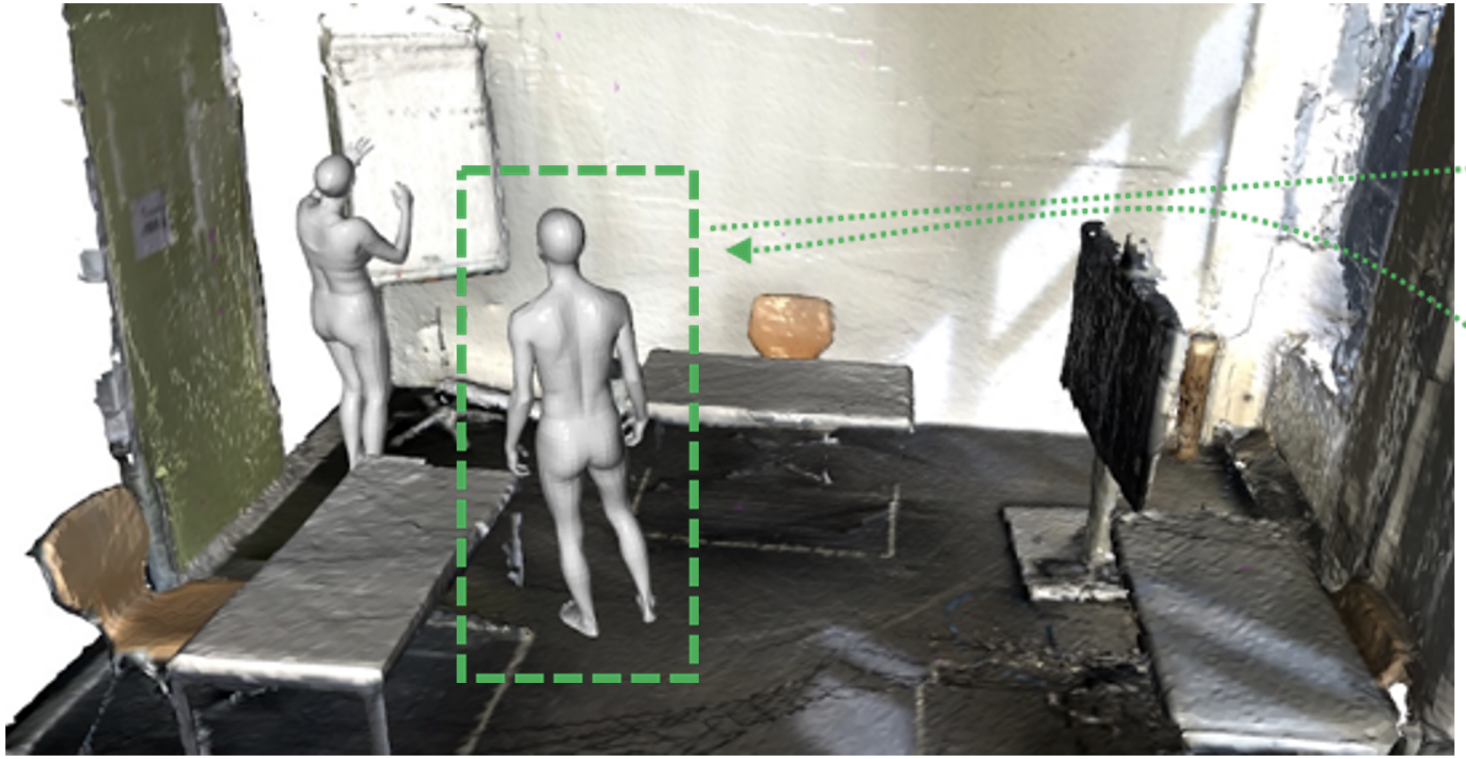
Dataset Overview

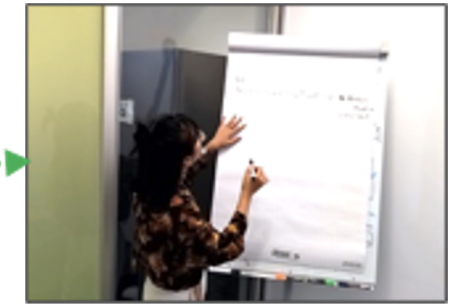
- **68** sequences
- **20** subjects
- **9** indoor scenes
- **153k multi-view third-person view RGBD frames** from Azure Kinects
- **139k egocentric view RGB frames** from HoloLens2
- **Eye gaze, hand/head tracking** from HoloLens2
- **3D human shape and motion annotations** for both interacting subjects

Capture Setup



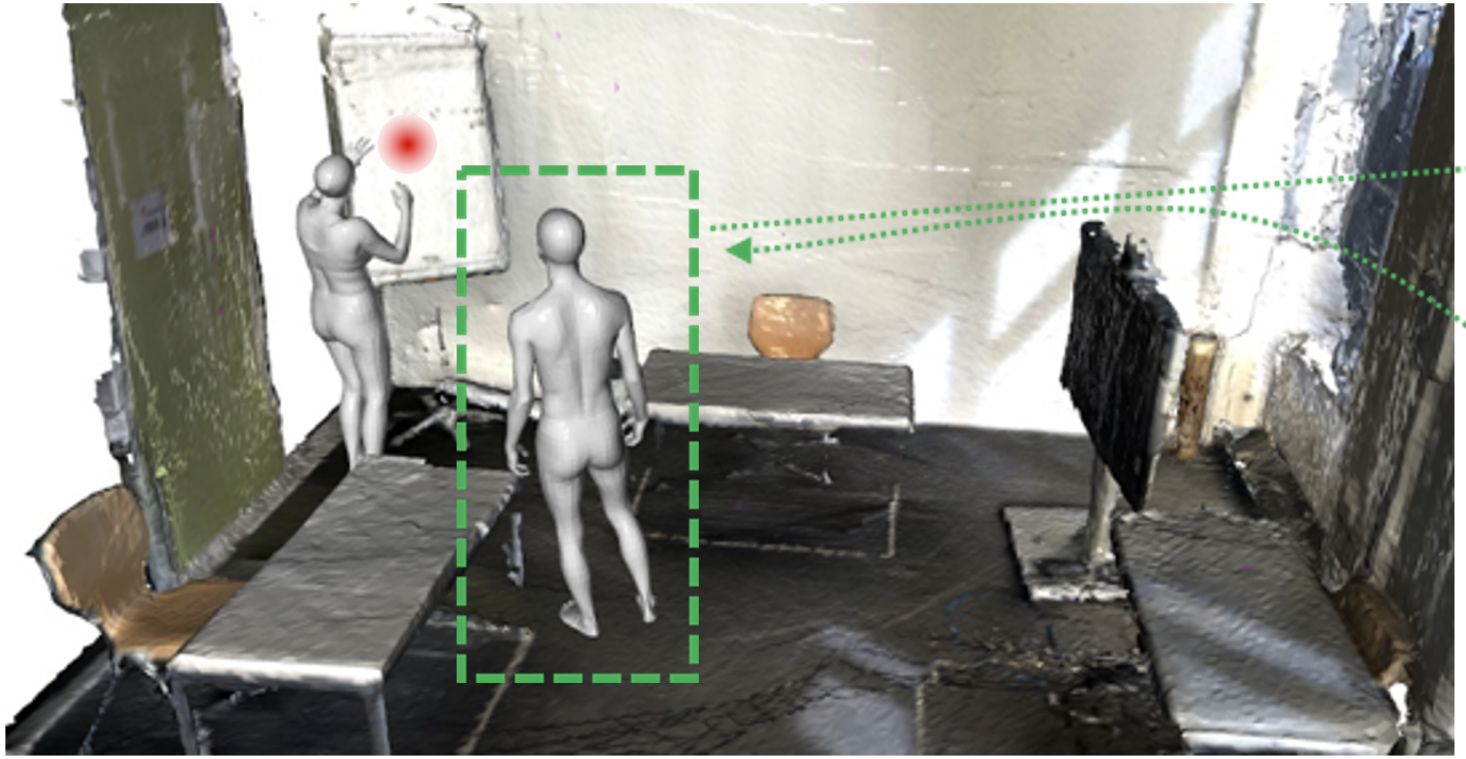






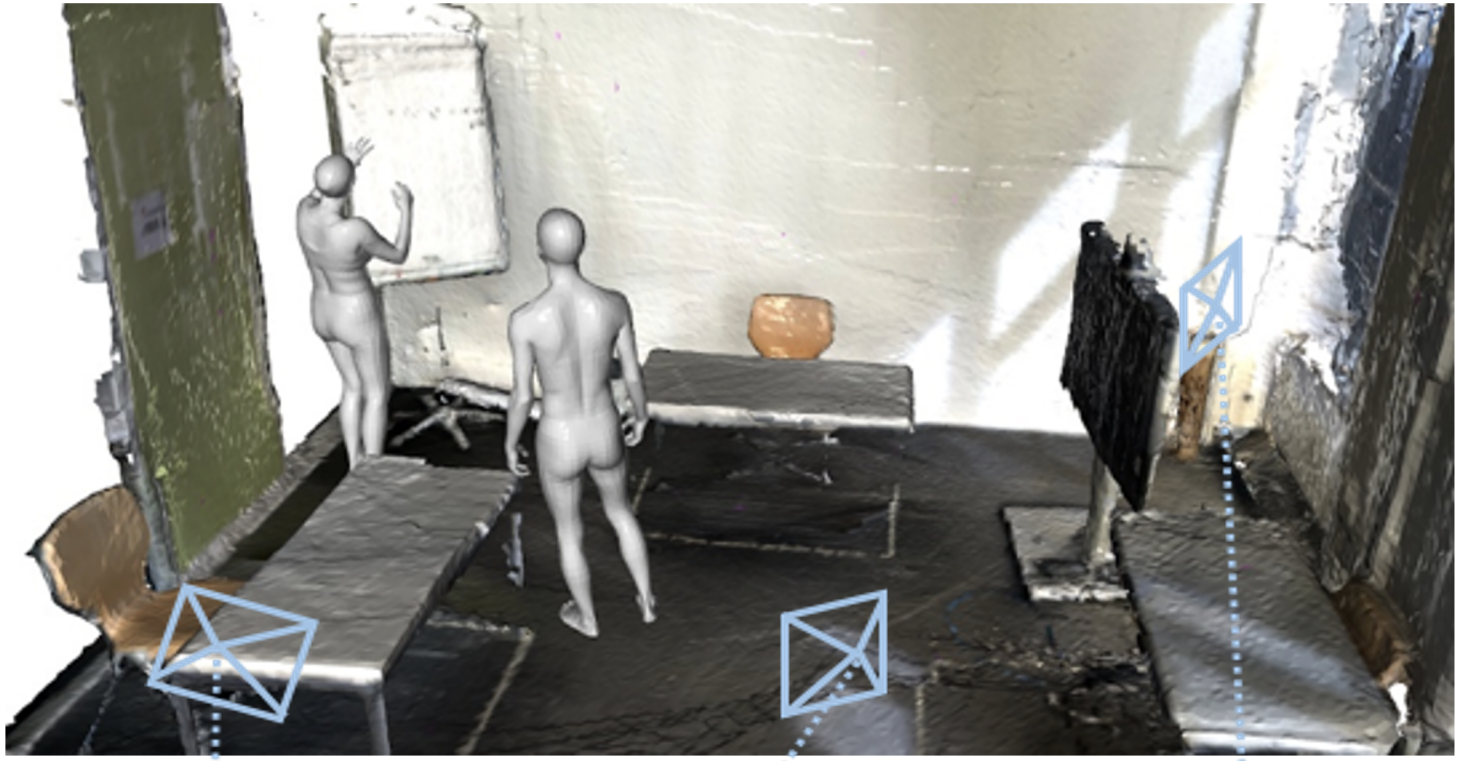
First-Person View

HoloLens2



First-Person View

HoloLens2



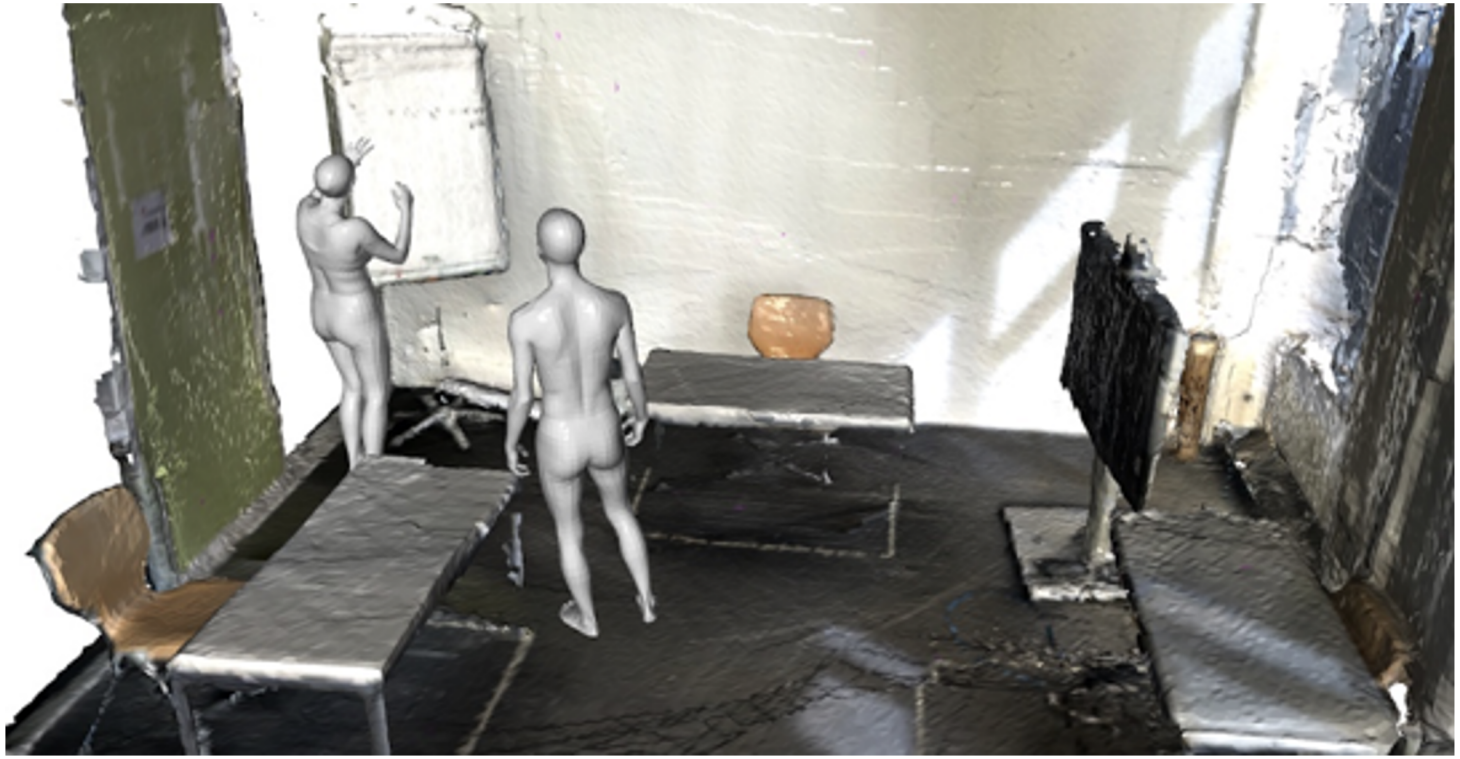
Kinect view 1

Kinect view 2

Kinect view 3

Third-Person View

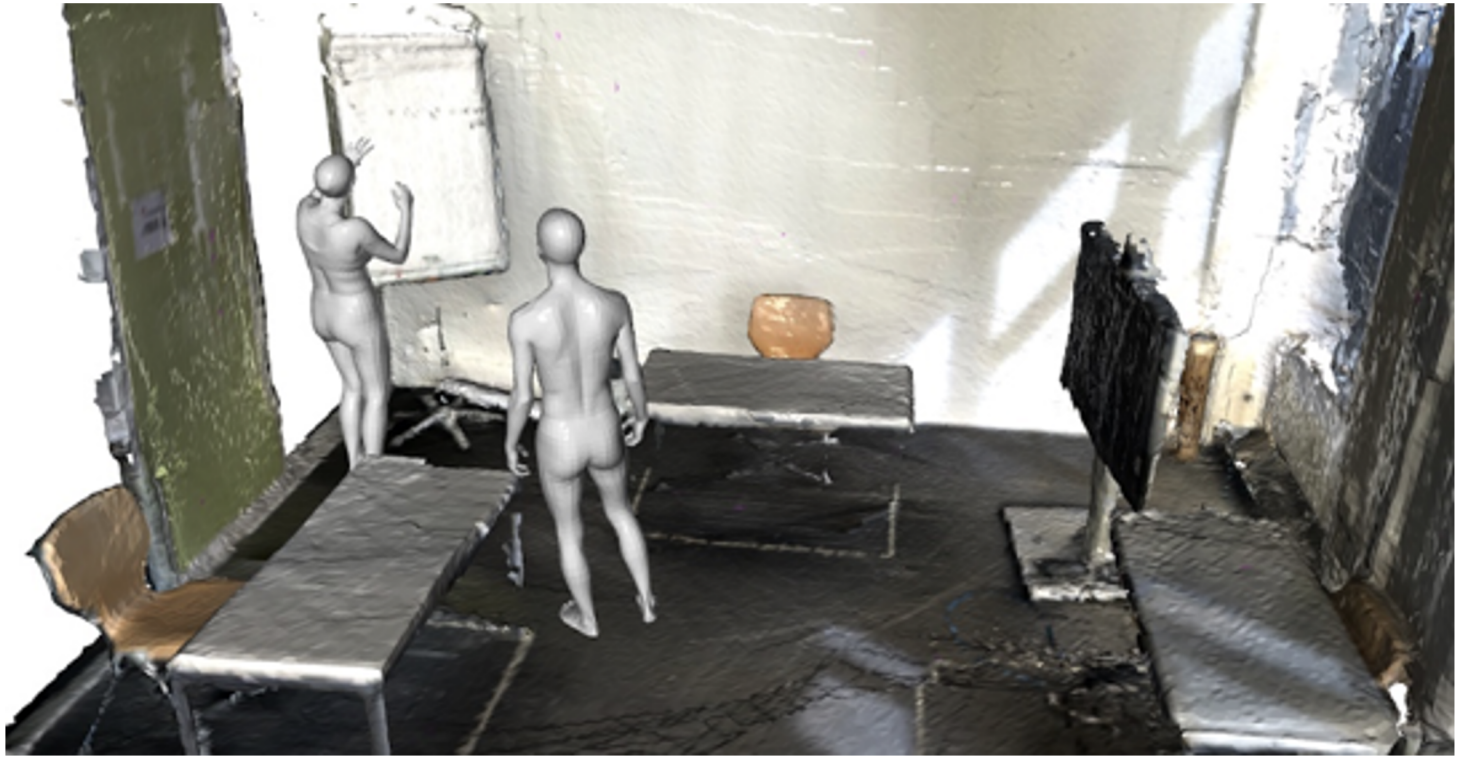
Azure Kinect



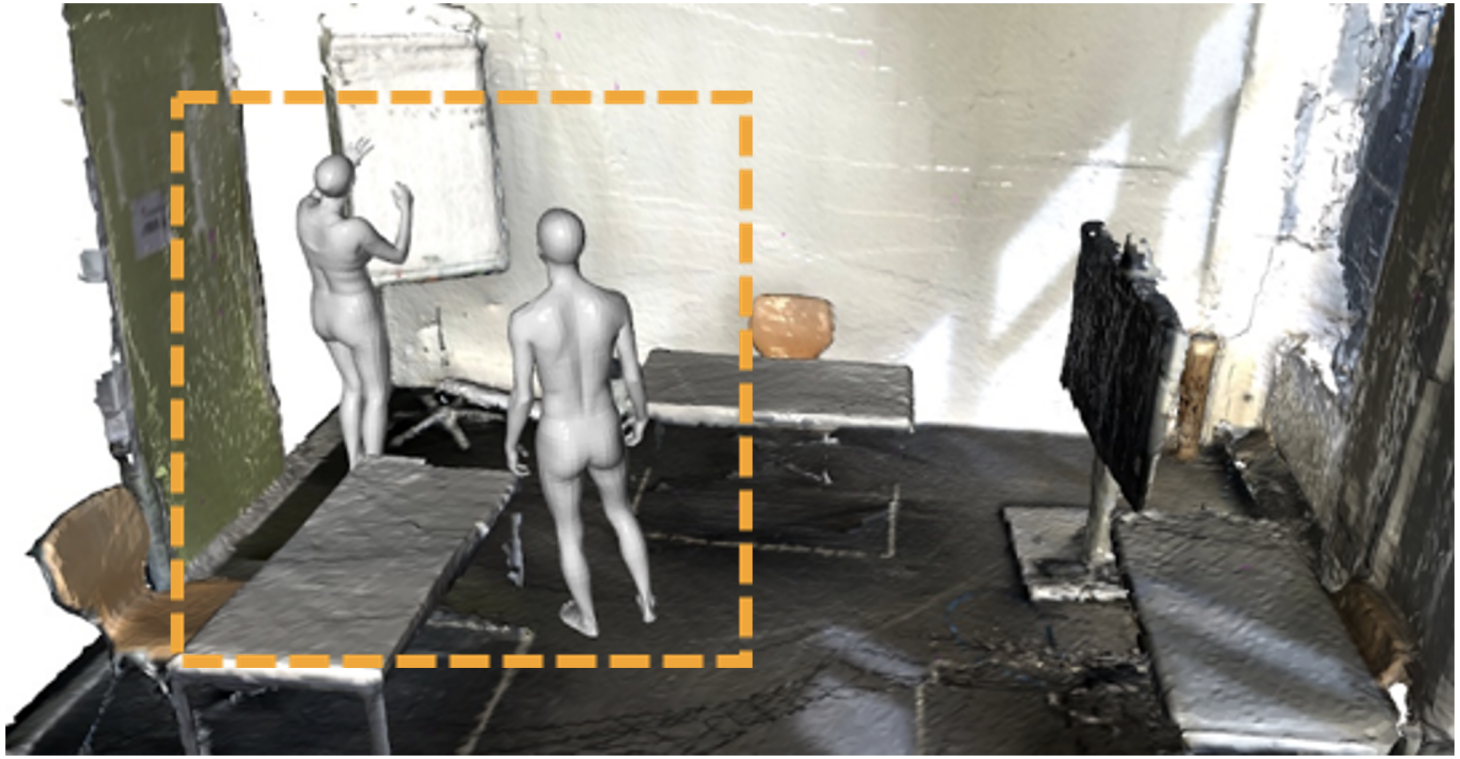
Kinect view 1

Kinect view 2

Kinect view 3

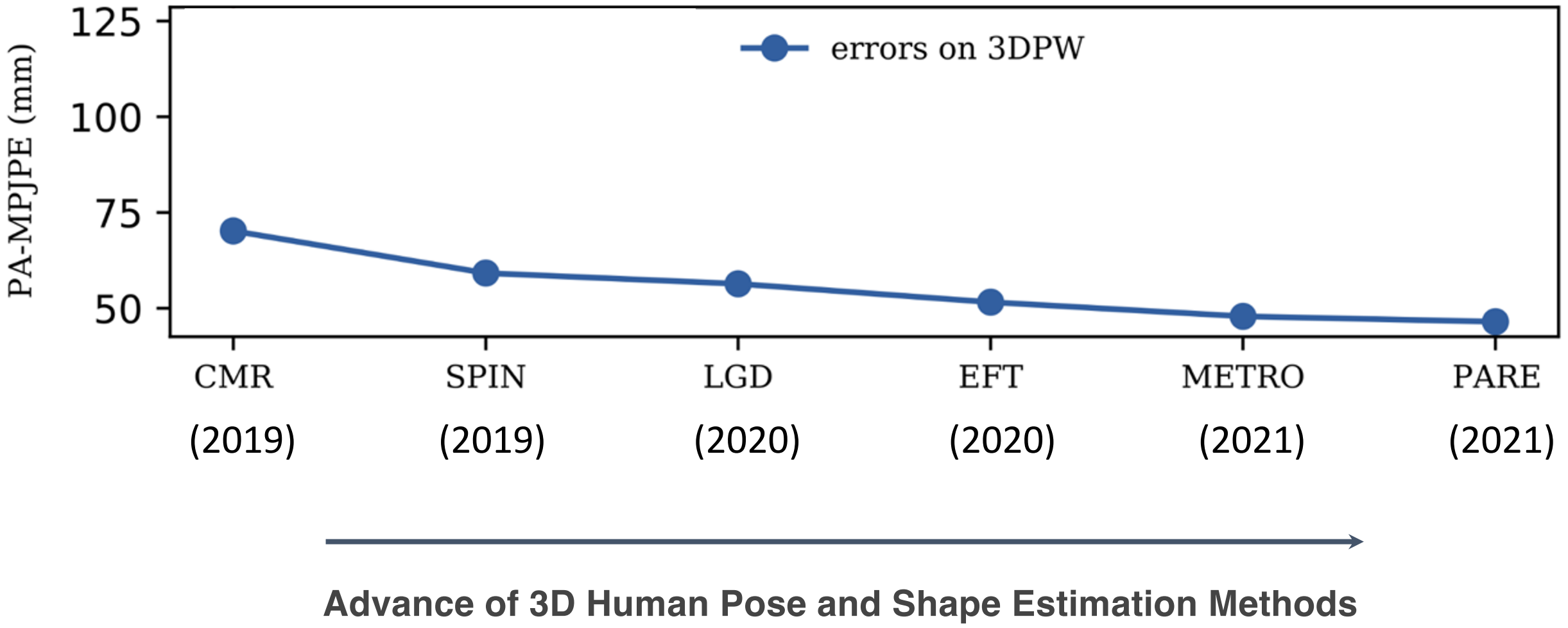


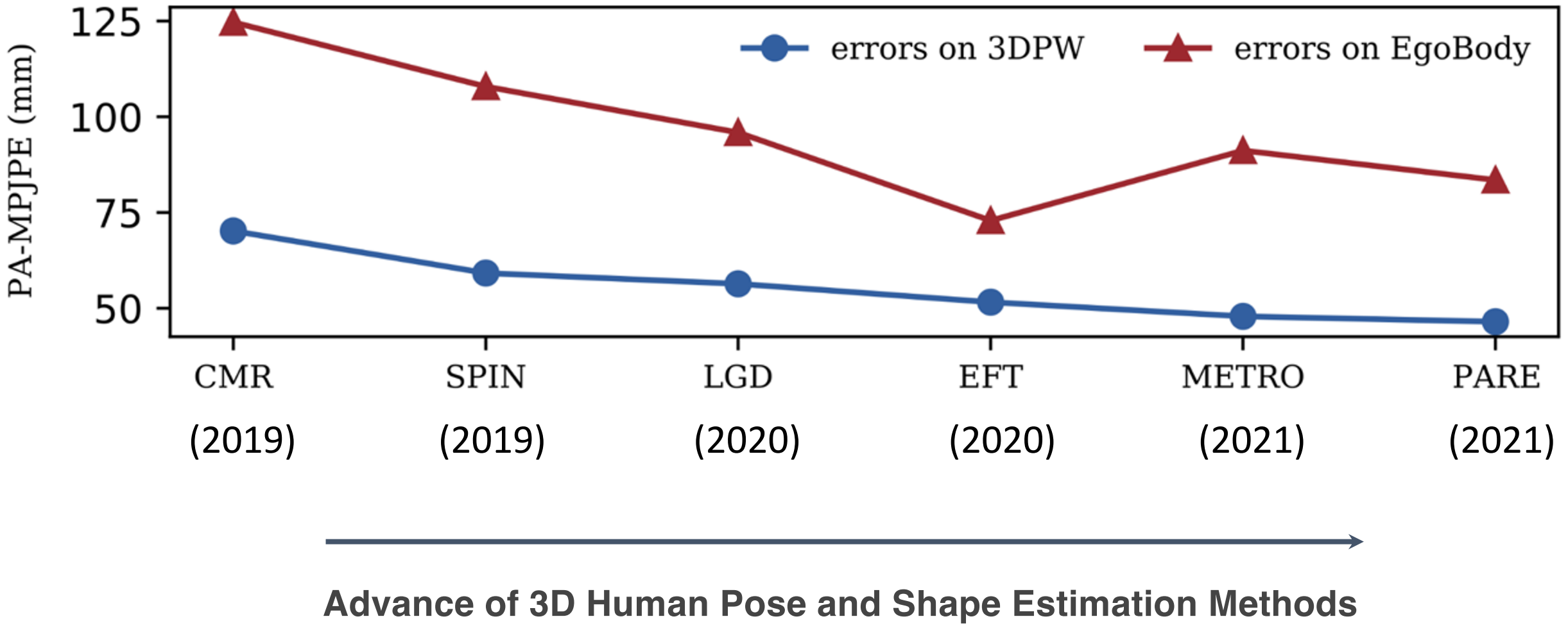
Kinect view 1 Kinect view 2 Kinect view 3



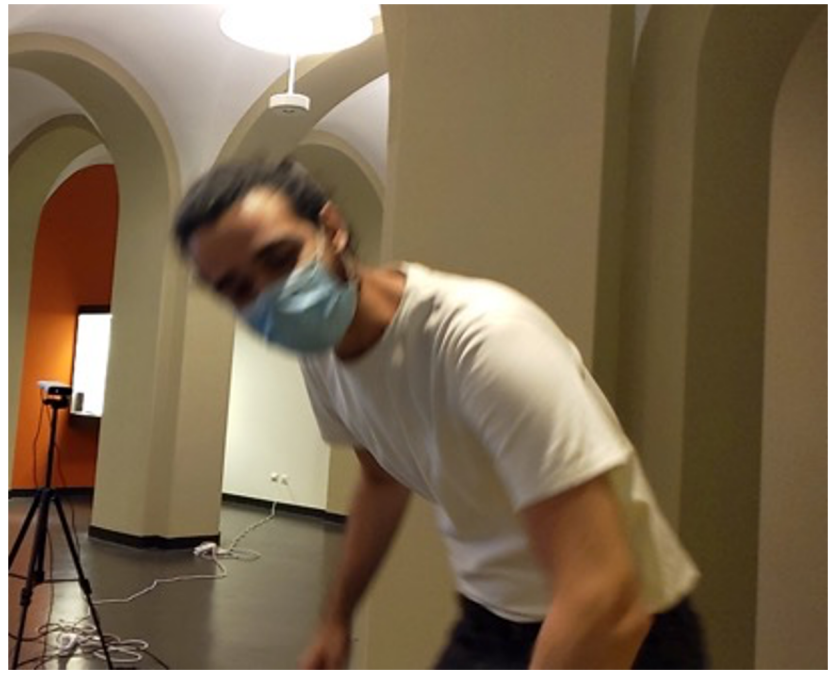
Kinect view 1 Kinect view 2 Kinect view 3

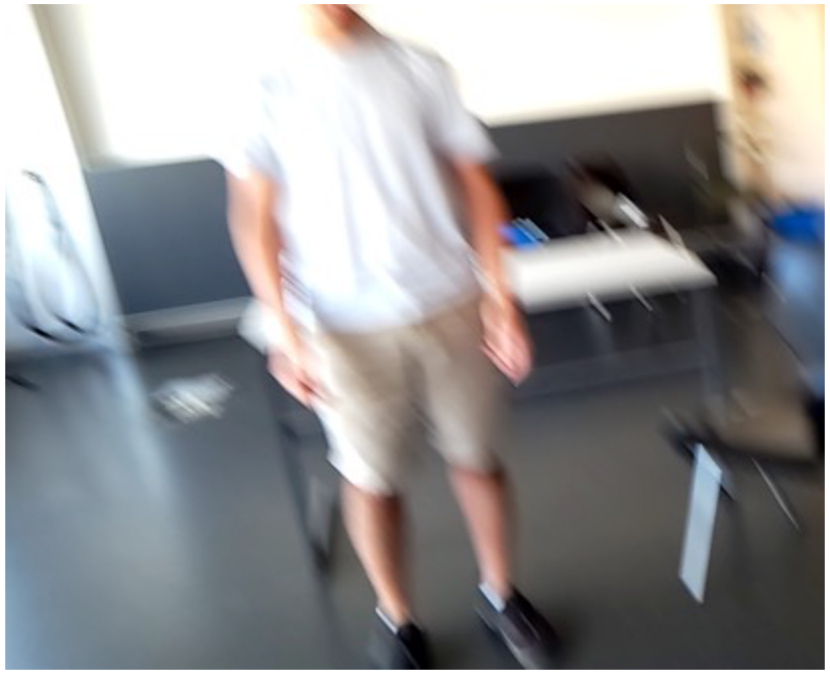
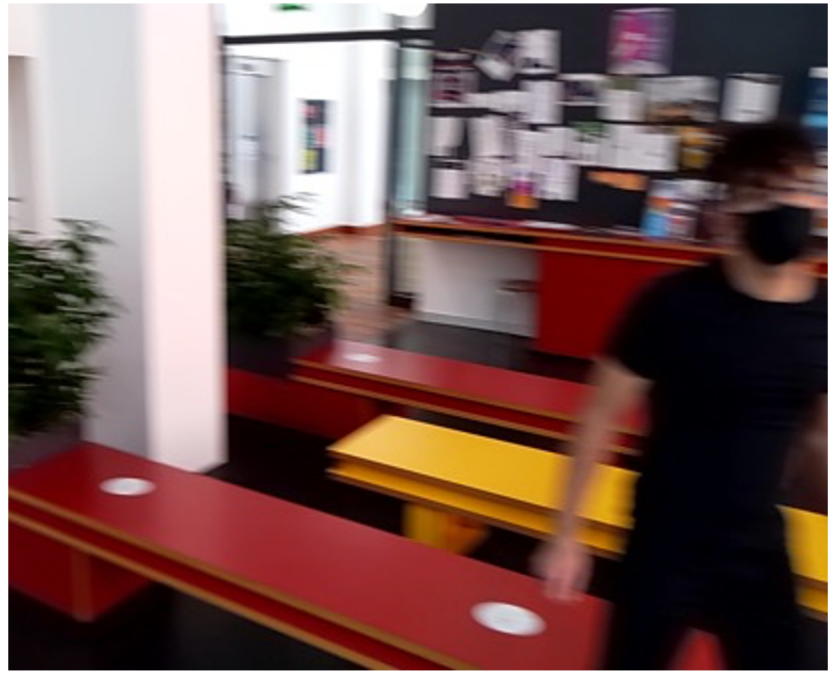
Benchmark:
3D Human Pose and Shape Estimation From Egocentric Images


PA-MPJPE (mm)
125
100
75
50
errors on 3DPW
CMR
(2019)
SPIN
(2019)
LGD
(2020)
EFT
(2020)
METRO
(2021)
PARE
(2021)
Advance of 3D Human Pose and Shape Estimation Methods


PA-MPJPE (mm)
125
100
75
50
errors on 3DPW
errors on EgoBody
CMR
SPIN
LGD
EFT
METRO
PARE
(2019)
(2019)
(2020)
(2020)
(2021)
(2021)
Advance of 3D Human Pose and Shape Estimation Methods



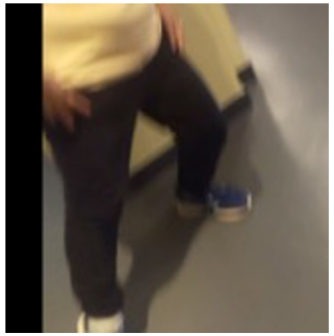


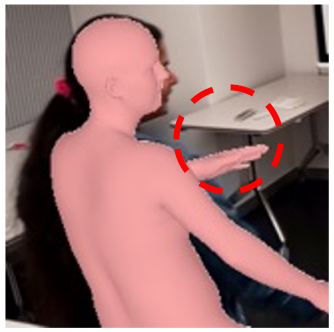
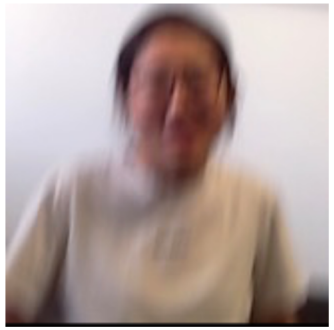
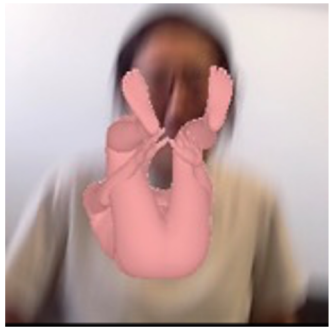
Input Image

SPIN
(Kolotouros et al.)

EFT
(Joo et al.)



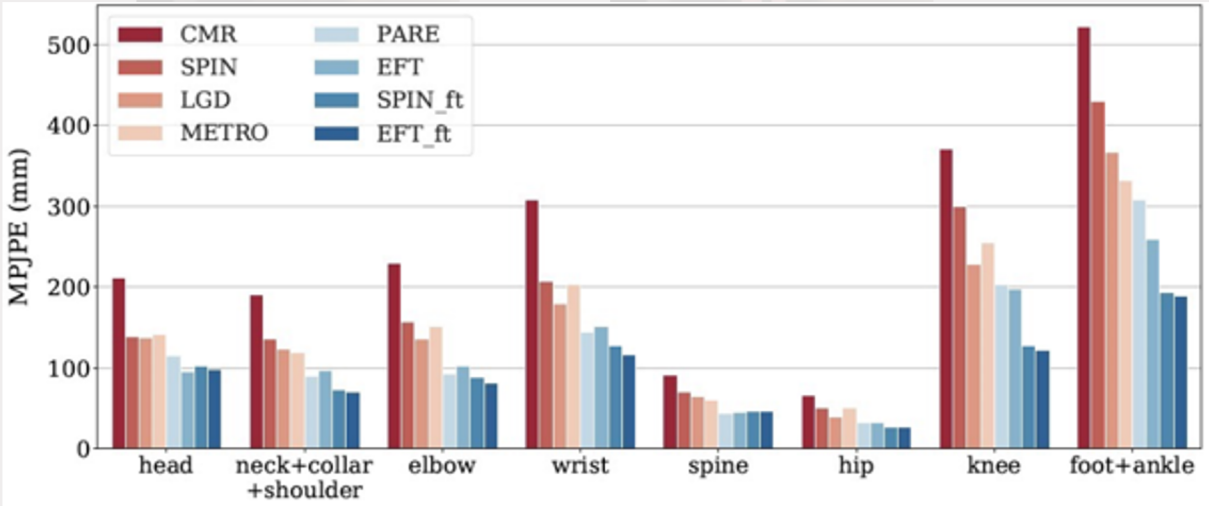


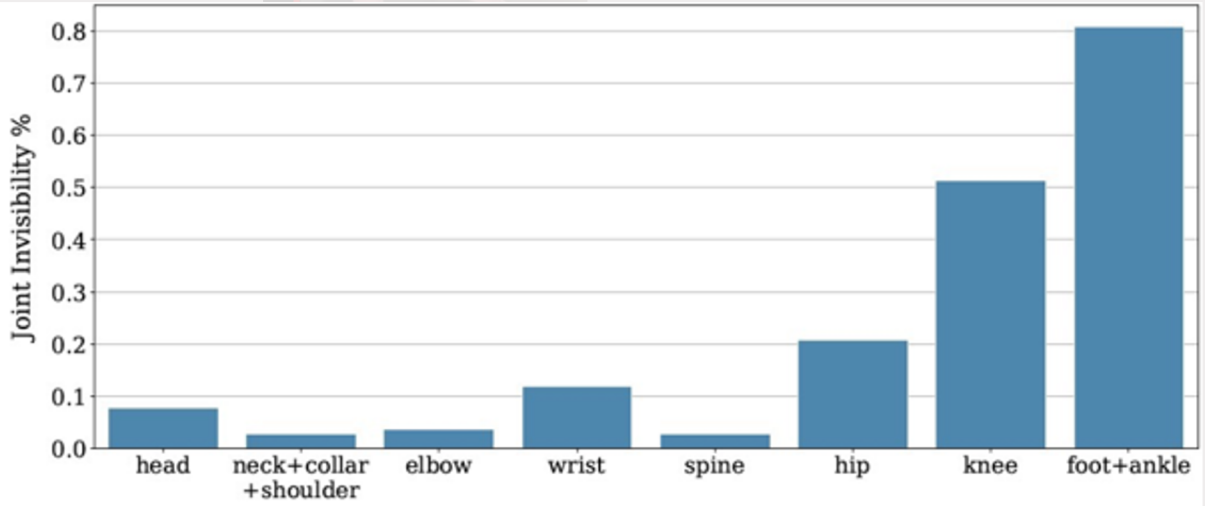






Input Image

SPIN
(Kolotouros et al.)

EFT
(Joo et al.)





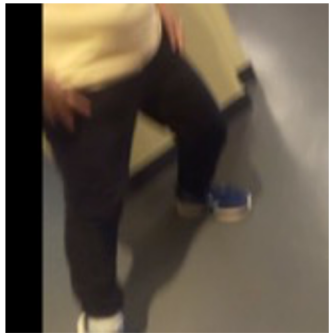
Error on each body joint group

How often a joint group is truncated from the image

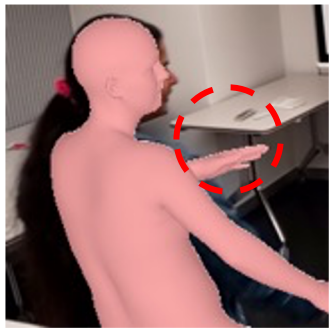
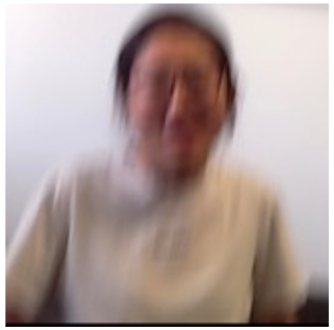
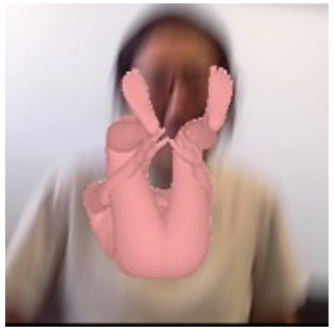
Input Image

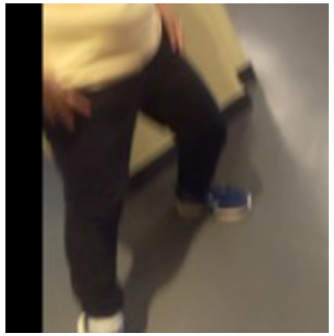
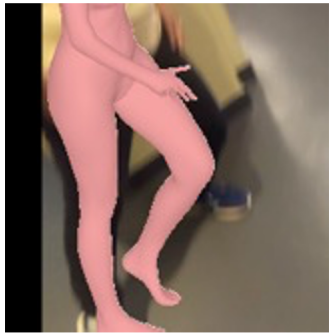
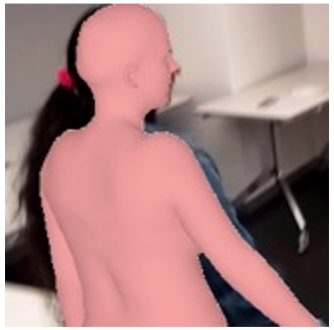
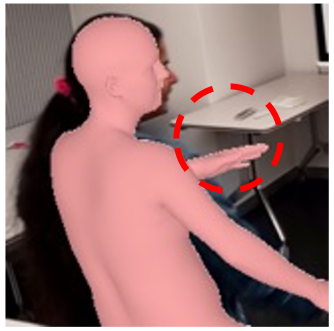
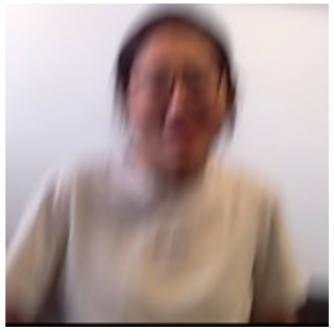
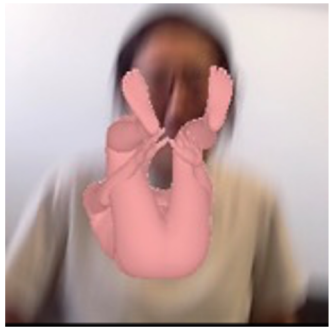
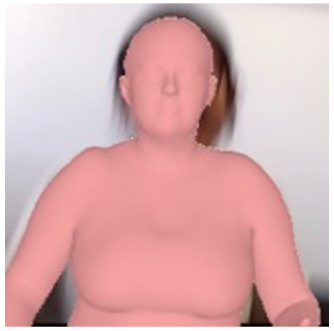
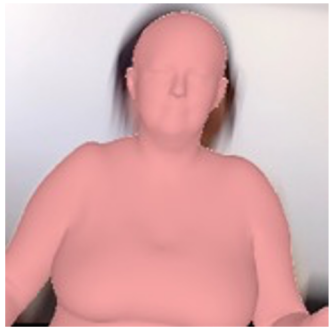
SPIN
(Kolotouros et al.)

EFT
(Joo et al.)






Input Image	SPIN (Kolotouros et al.)	SPIN fine-tuned on our training set	EFT (Joo et al.)	EFT fine-tuned on our training set
				
				
				

Pose and Shape Estimation Errors on EgoBody Test Set

Method	MPJPE ↓	PA-MPJPE ↓	V2V ↓	PA-V2V ↓
SPIN	189.9	-49%	210.5	-42%
SPIN-ft (Ours)	96.2		122.3	
METRO	161.5	-35%	187.5	-44%
METRO-ft (Ours)	105.4		**105.9**	
EFT	123.3	-25%	143.3	-20%
EFT-ft (Ours)	**92.5**		115.1	

Pose and Shape Estimation Errors on EgoBody Test Set

Method	MPJPE ↓	PA-MPJPE ↓	V2V ↓	PA-V2V ↓
SPIN	189.9	-49%	210.5	-42%
SPIN-ft (Ours)	96.2		122.3	
METRO	161.5	-35%	187.5	-44%
METRO-ft (Ours)	105.4		**105.9**	
EFT	123.3	-25%	143.3	-20%
EFT-ft (Ours)	**92.5**		115.1	

Pose accuracy

Shape accuracy

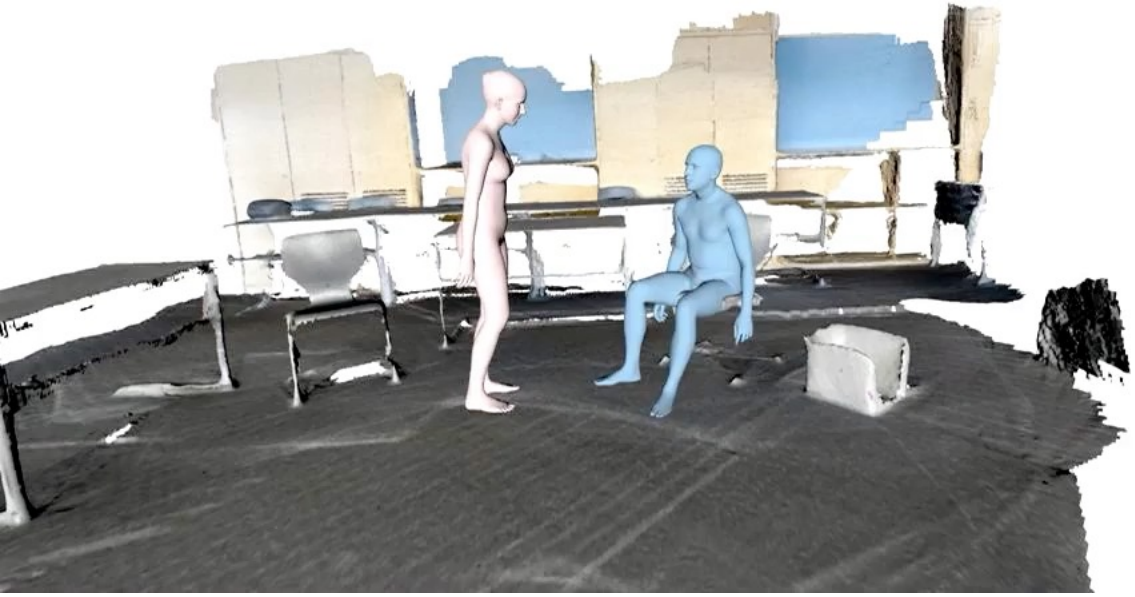
Cross-dataset Evaluation on You2Me

Finetuned on EgoBody training set

Pose and Shape Estimation Errors on the You2Me Dataset

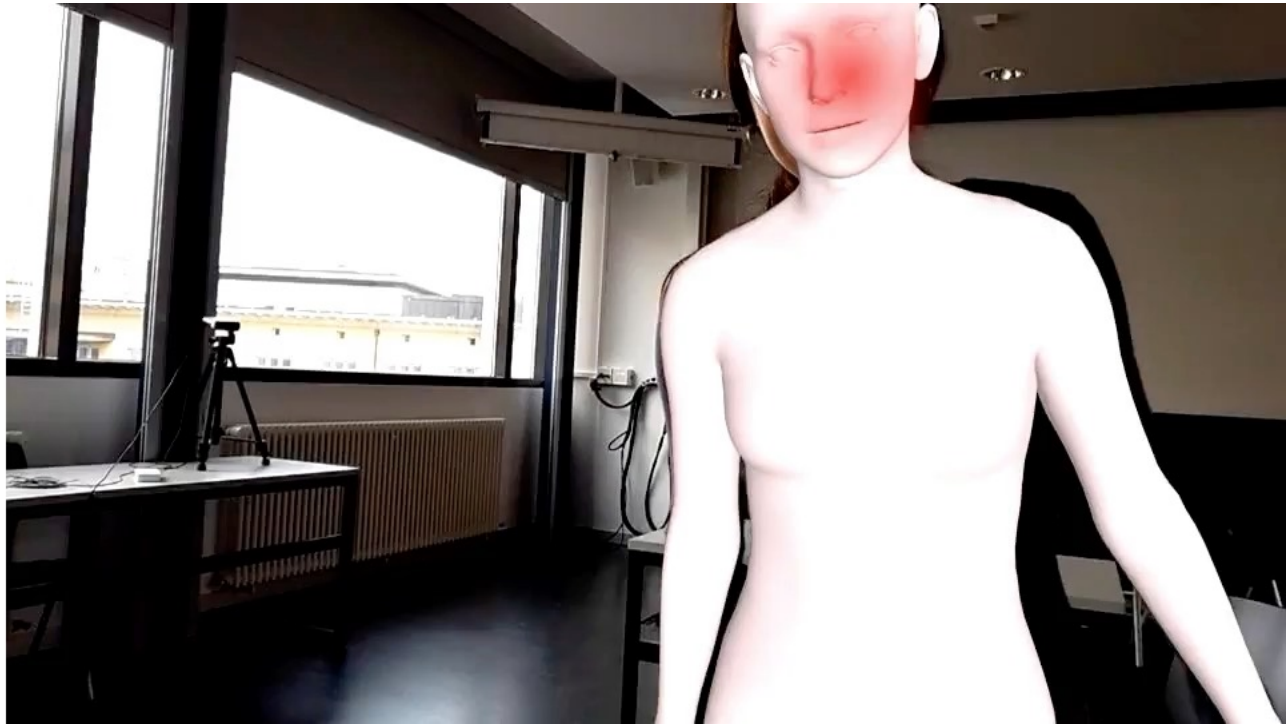
Method	PA-MPJPE ↓	
SPIN	155.0	
SPIN-ft (Ours)	89.4	-42%
METRO	117.7	
METRO (Ours)	90.1	-23%
EFT	96.0	
EFT-ft (Ours)	88.7	-8%

More results: EgoBody Dataset



Third-Person View

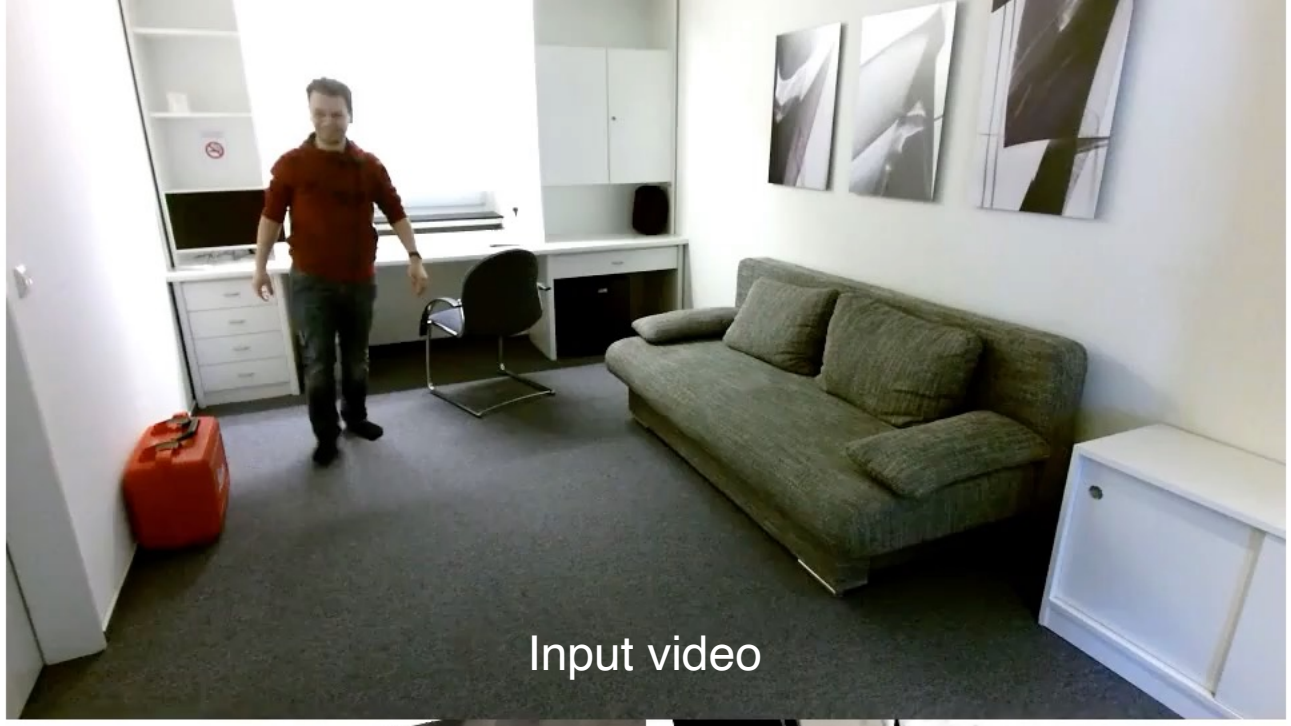
+ ground truth body annotation of
the camera wearer & the second person
+ scene reconstruction

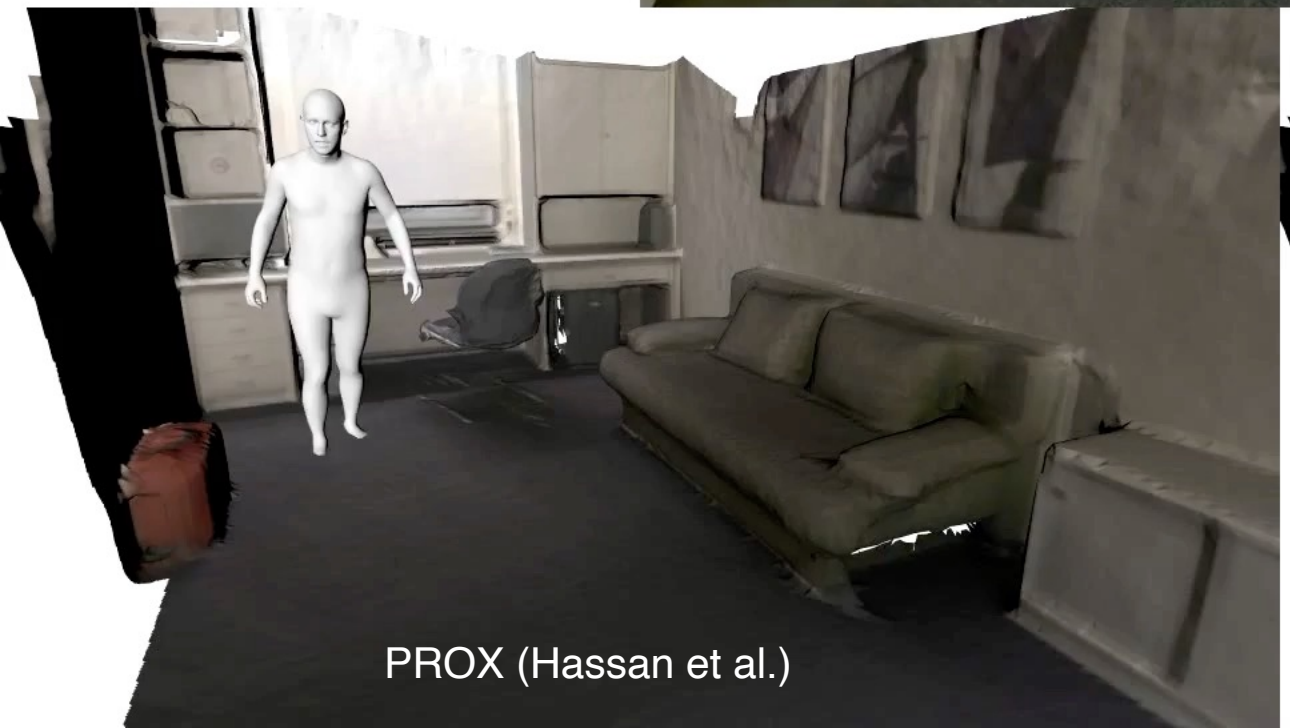


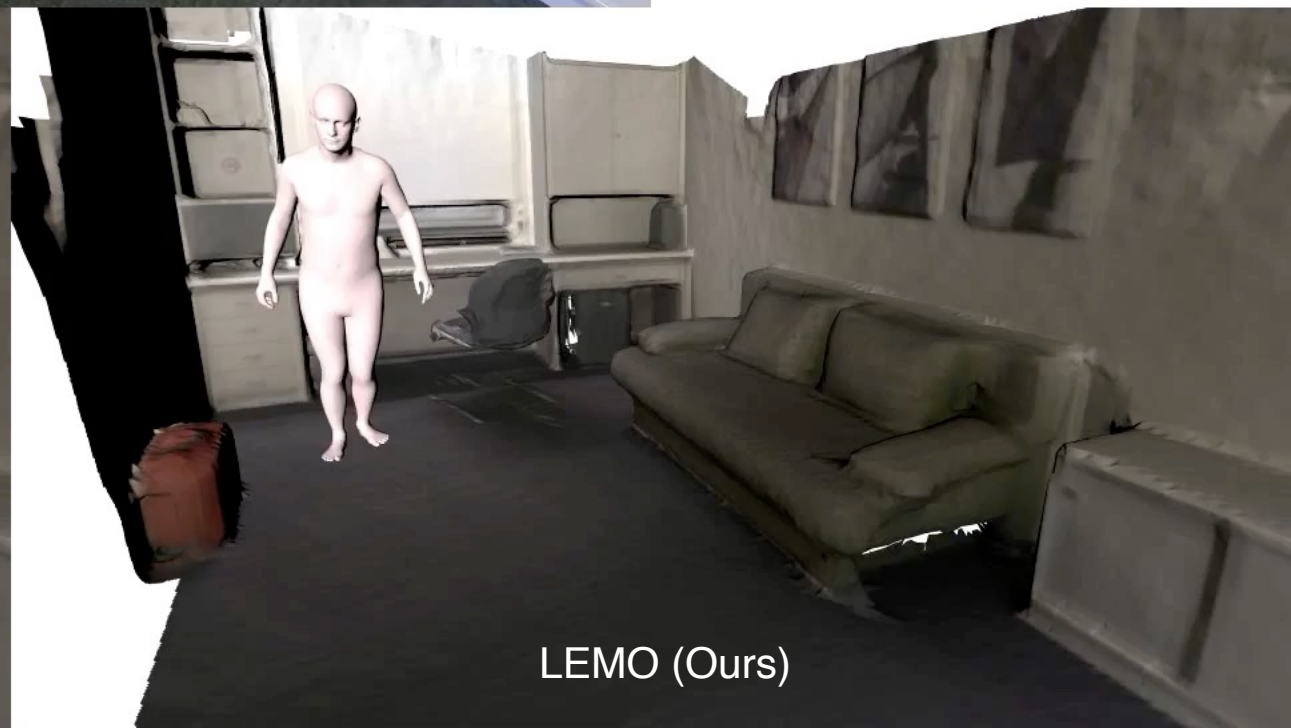
Egocentric View

+ ground truth body annotation
+ eye gaze / attention

More results: LEMO







Egocentric Interaction Capture for Mixed Reality

Project page (LEMO):
https://sanweiliti.github.io/LEMO/LEMO.html

Project page (EgoBody):
https://sanweiliti.github.io/egobody/egobody.html

ETH zürich VLG Computer Vision and Learning Group Microsoft

siwei.zhang@inf.ethz.ch

Contact-aware Motion Infilling Prior