Human3D

Learning 3D Human Foundation Models

Prof. Dr. Siyu Tang

Department of Computer Science

ETH Zürich

Recent breakthroughs in large language and vision models

ChatGPT (OpenAI)

SegmentAnything (Meta)

Veo (Google Deepmind)

What is still missing?

AI’s ability to perceive the world in 3D

Our world is dynamic, and inhabited and shaped by humans

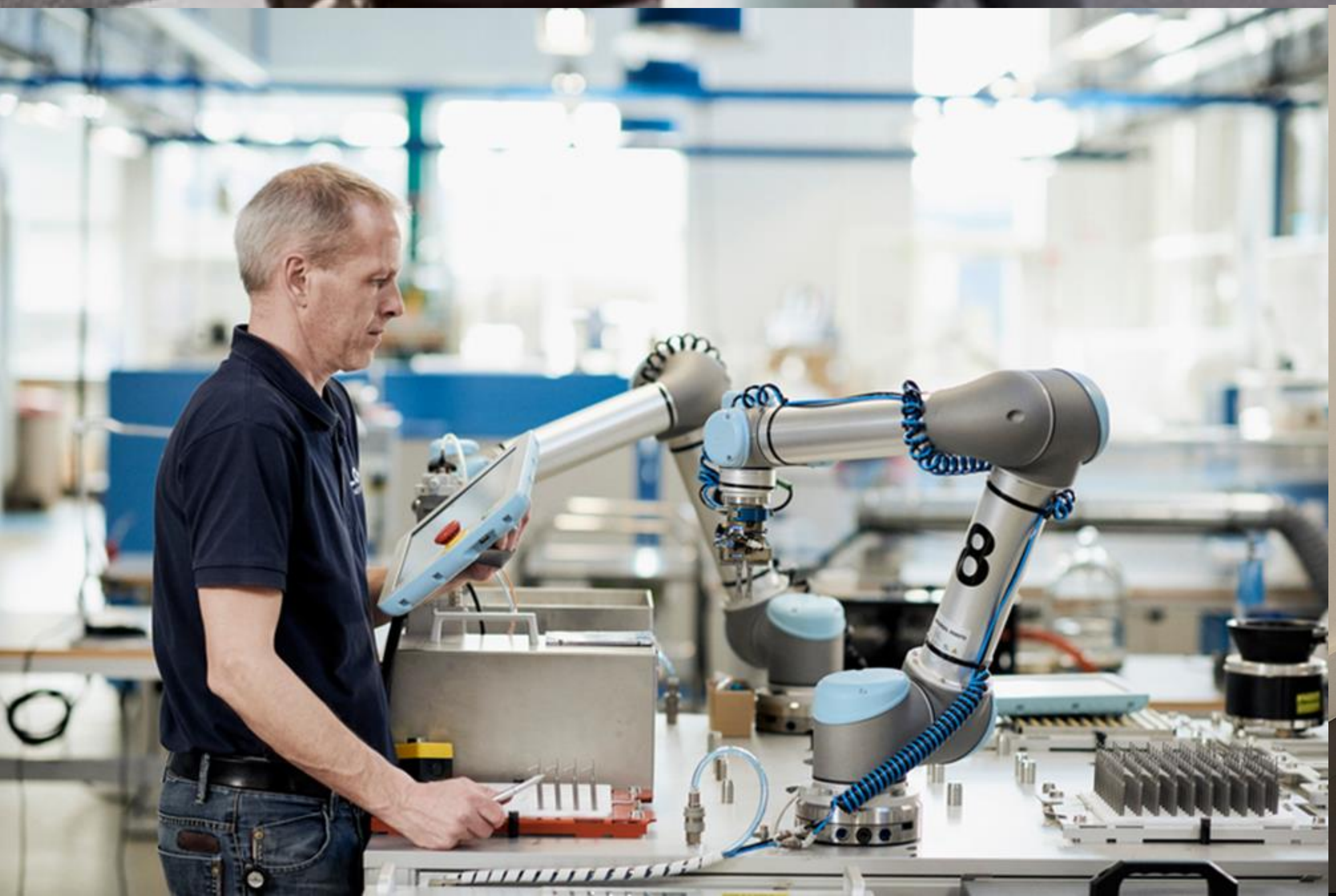
Our goal:

*To enable AI systems to **see** and **understand** the dynamic world, and, most importantly, interact intelligently and safely with **humans**.*

Video source: Pexels

Autonomous Agents



To develop AI systems capable of seeing and engaging with humans

- the key challenge -

Lack of human-centric training data

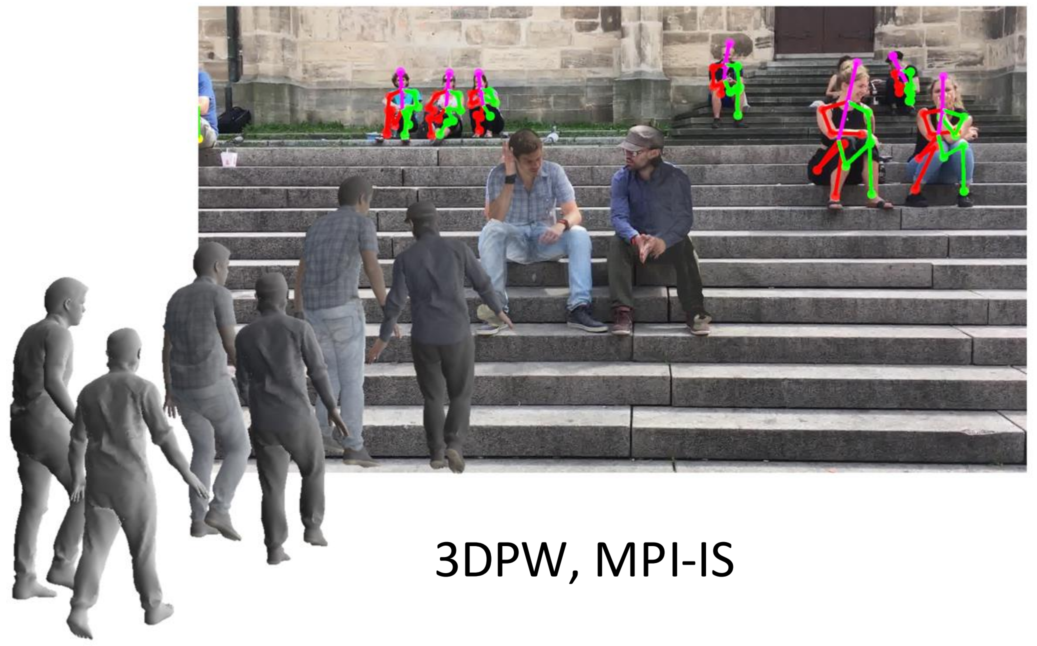
ChatGPT

Open AI

GPT-3: 300+ billion tokens
WebText, Wikipedia,
Common crawl, books

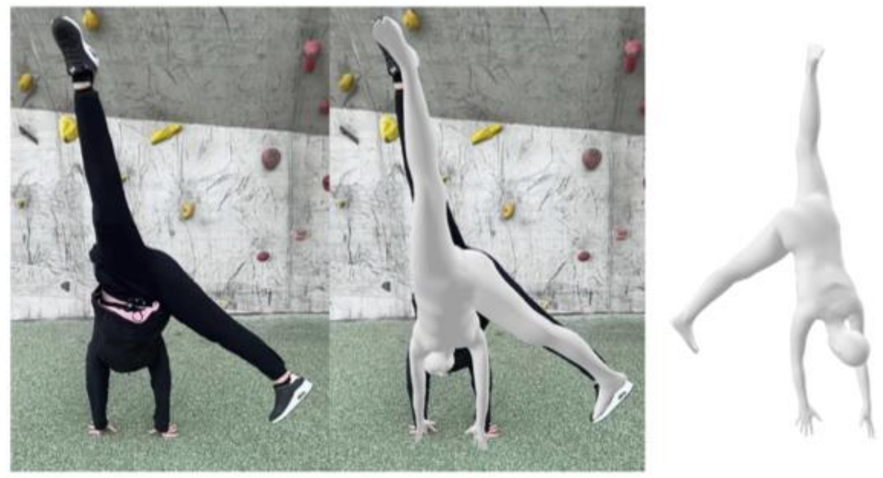
Movie Gen, Meta

around 100 million videos,
billions of images



3DPW, MPI-IS

7 actors in 18 clothing styles
60 sequences

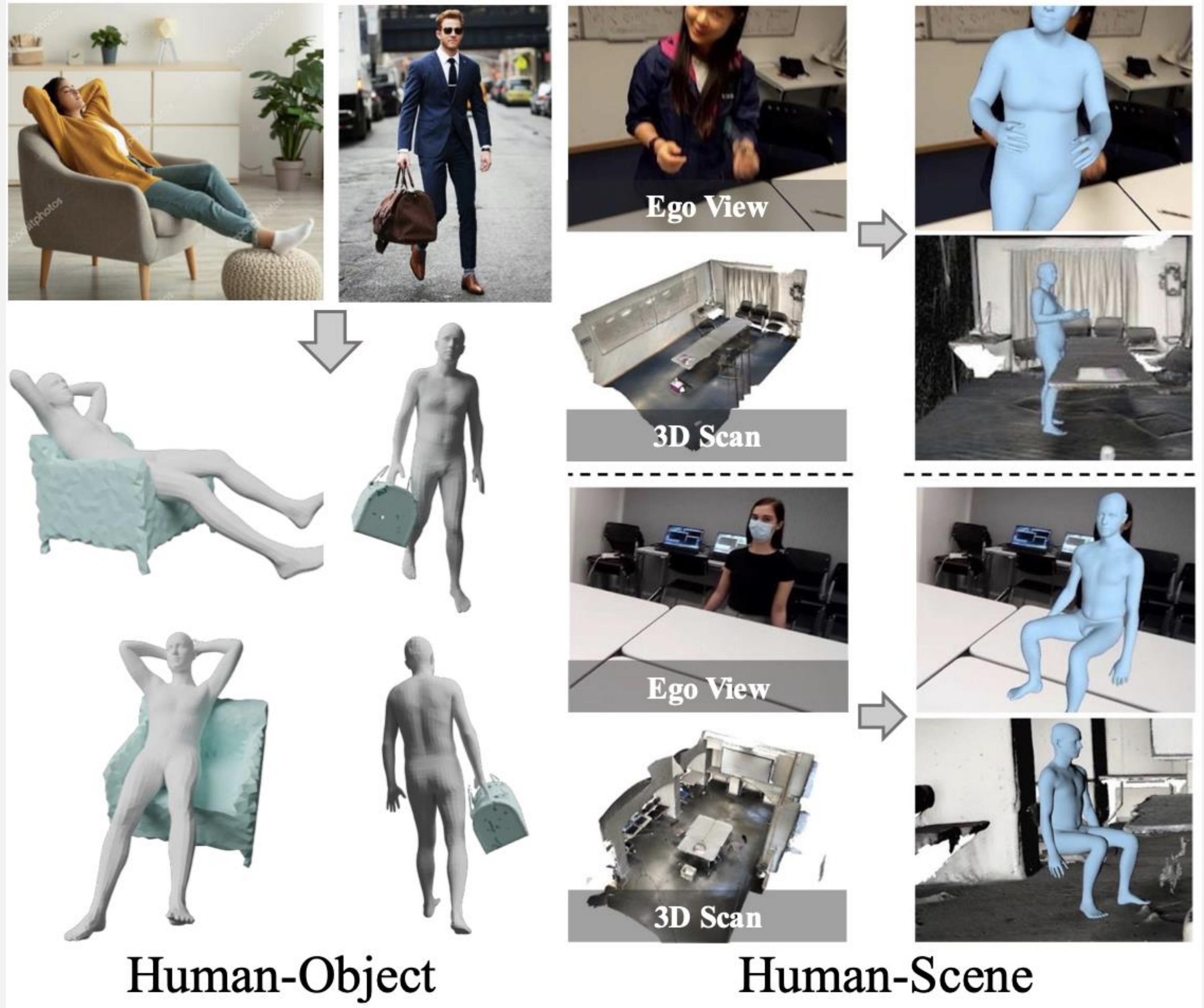


EMDB, ETH Zürich

10 actors, 81 scenes
58 minutes

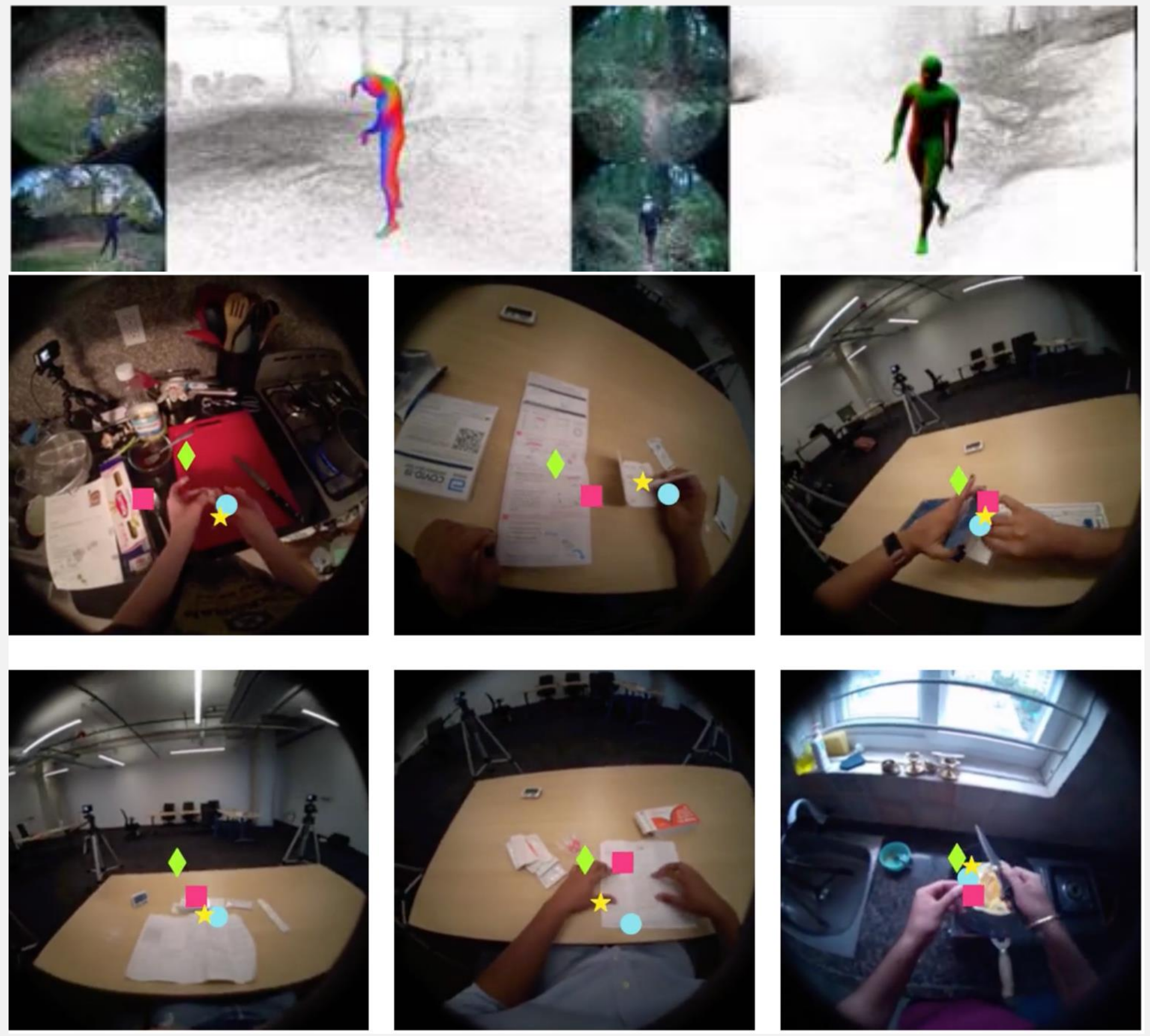
Our roadmap

1. Perceiving and reconstructing *real* humans



2. Synthesizing *virtual* humans

3. Embodied *digital* humans



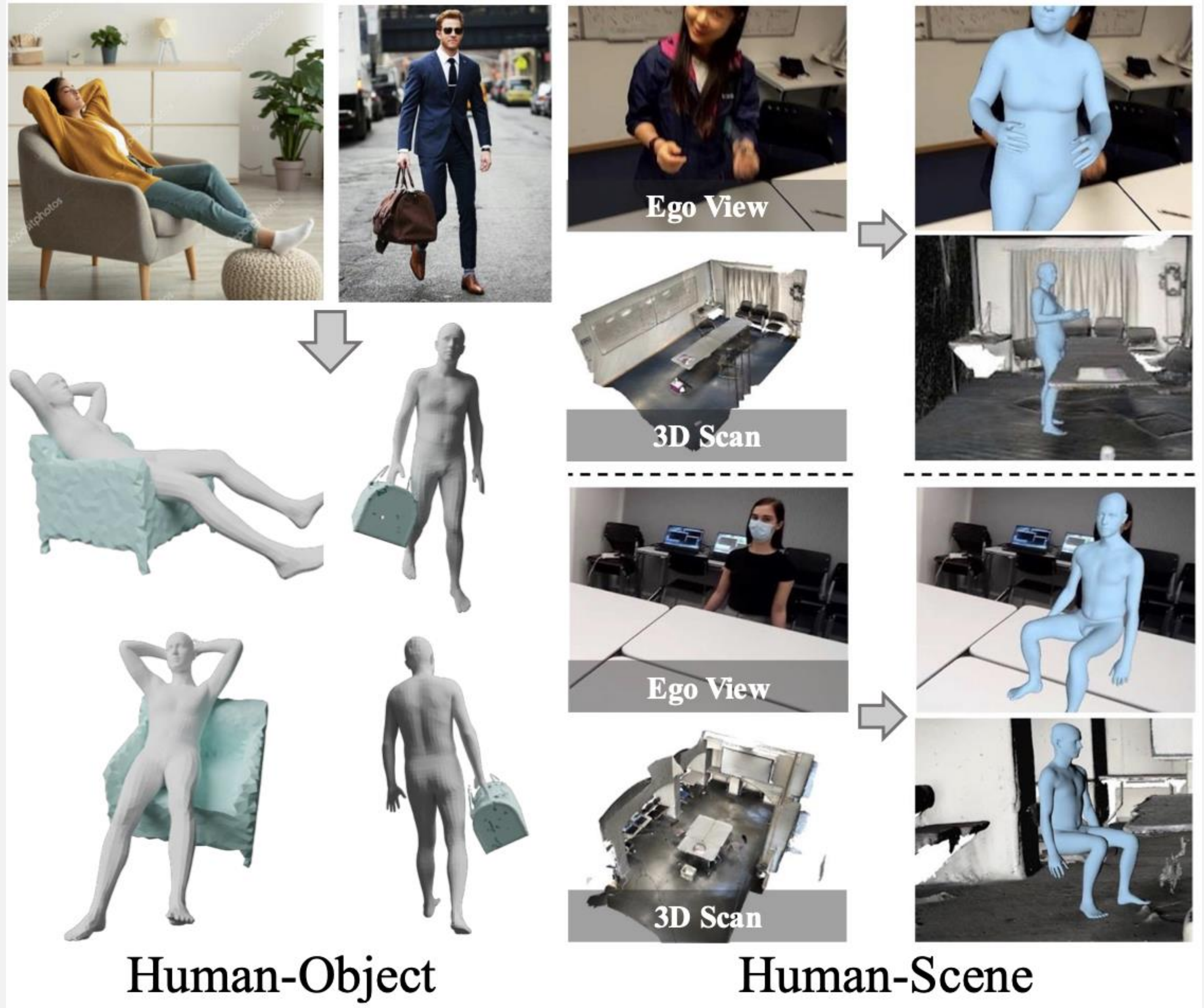
- Rich semantics
- Diverse motion and appearance
- Very limited 3D ground truth data

- Rich and accurate 3D ground-truth annotations
- Human behavior synthesis is a really hard problem

- Multi-modality data
- hand-object interaction
- Human motion capture with very limited observations

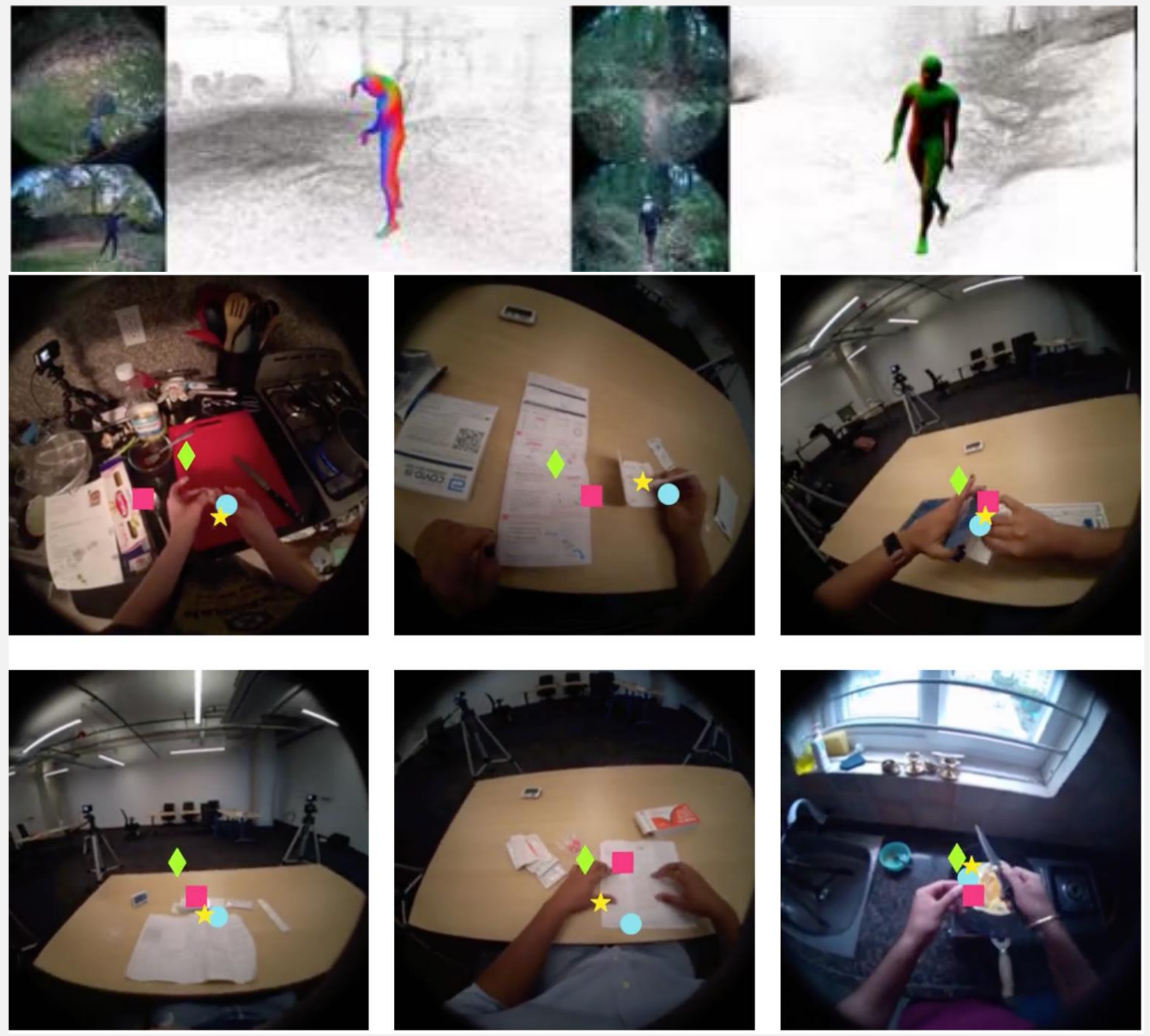
Our roadmap

1. Perceiving and reconstructing *real* humans



2. Synthesizing *virtual* humans

3. Embodied *digital* humans



- Rich semantics
- Diverse motion and appearance
- Very limited 3D ground truth data

- Rich and accurate 3D ground-truth annotations
- Human behavior synthesis is a really hard problem

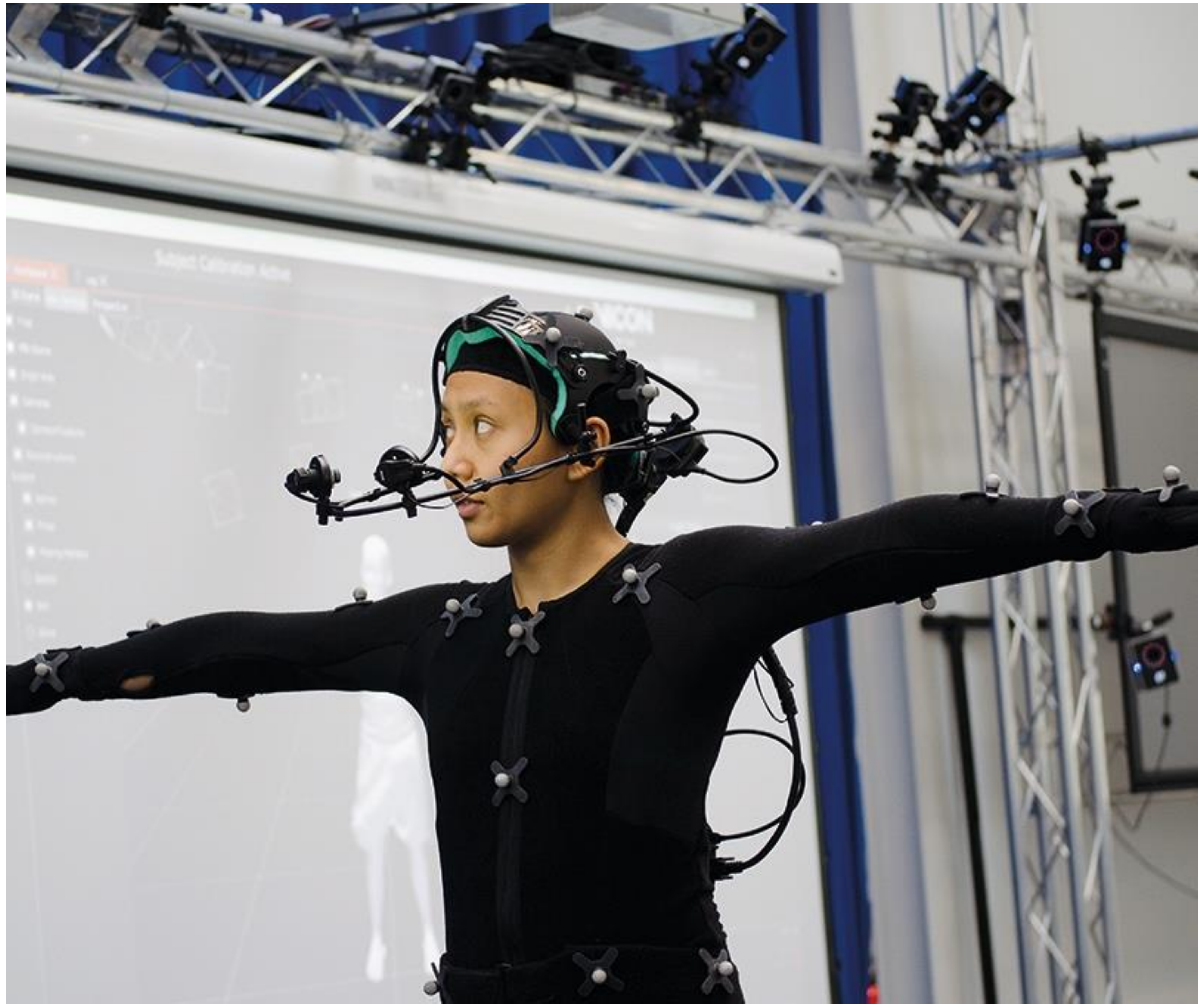
- Multi-modality data
- Hand-object interaction
- Human motion capture with very limited observations

4. Unified human foundation models that are 3D-grounded and multimodally capable.

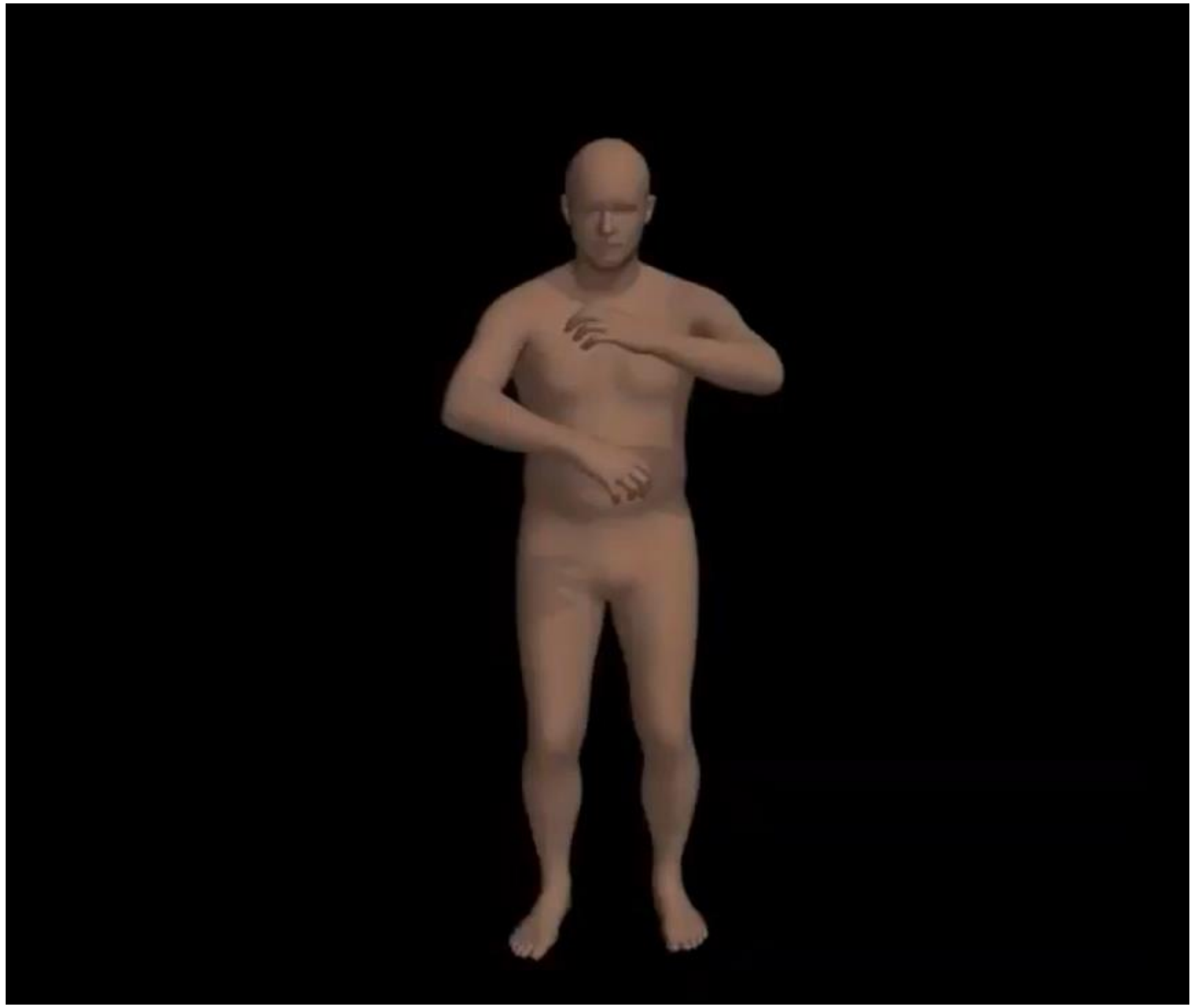
ERC Starting Grant 2025

Part 1. 3D human body and motion estimation

Human motion capture



Motion Capture System



AMASS @ICCV 2019

Monocular*, *markerless* human motion capture *in the wild

Human motion estimation from a RGBD camera

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

Monocular, ***markerless*** human motion capture ***in the wild***

How to reconstruct natural and smooth human motions in 3D scenes with a monocular camera?

Key insight: learning motion priors from high quality mocap datasets

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

LEMO

Learning Motion Priors for 4D Human Body Capture in 3D Scenes

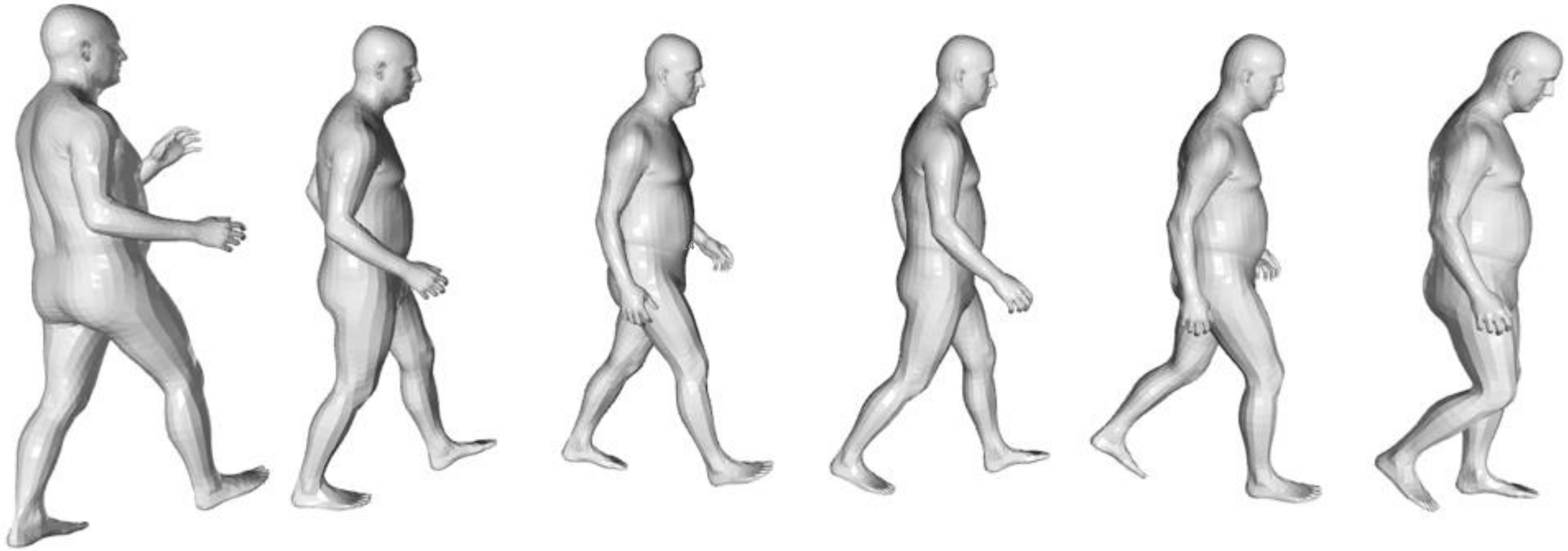
Siwei Zhang[1] Yan Zhang[1] Federica Bogo[2] Marc Pollefeys[1,2] Siyu Tang[1]

[1]ETH Zurich [2]Microsoft

ICCV 2021 Oral (acceptance rate: 3.3%)

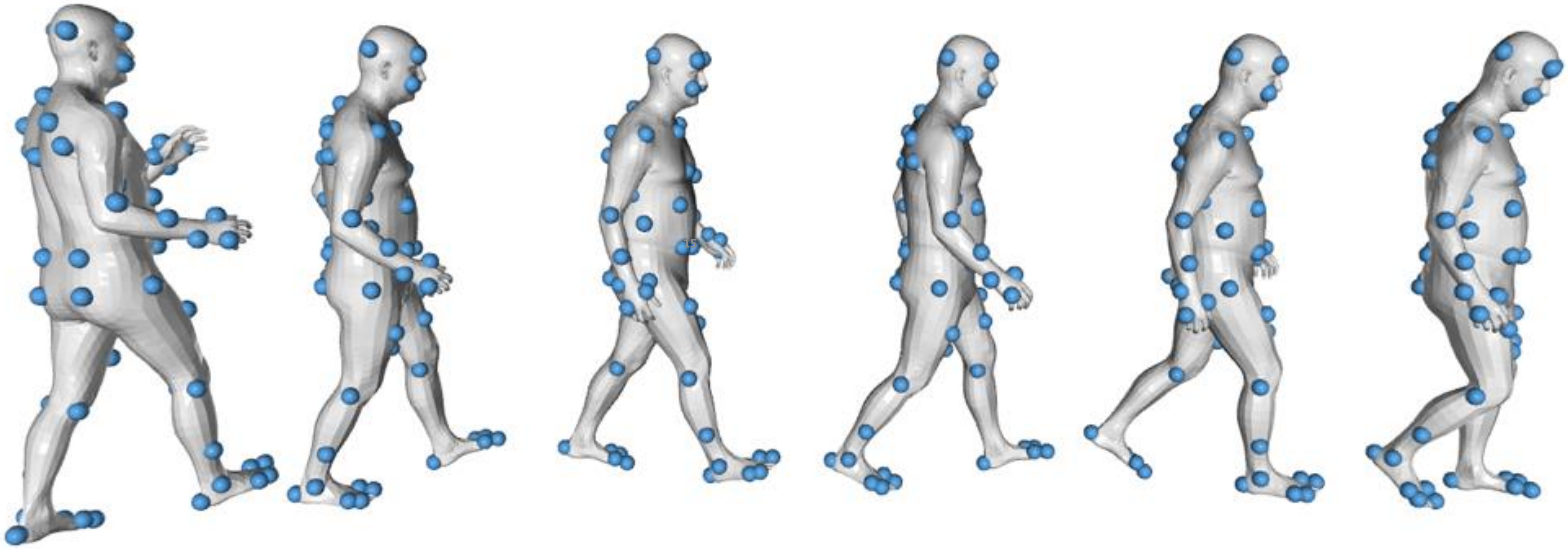
LEMO: Learning Motion Priors

Marker-based motion priors



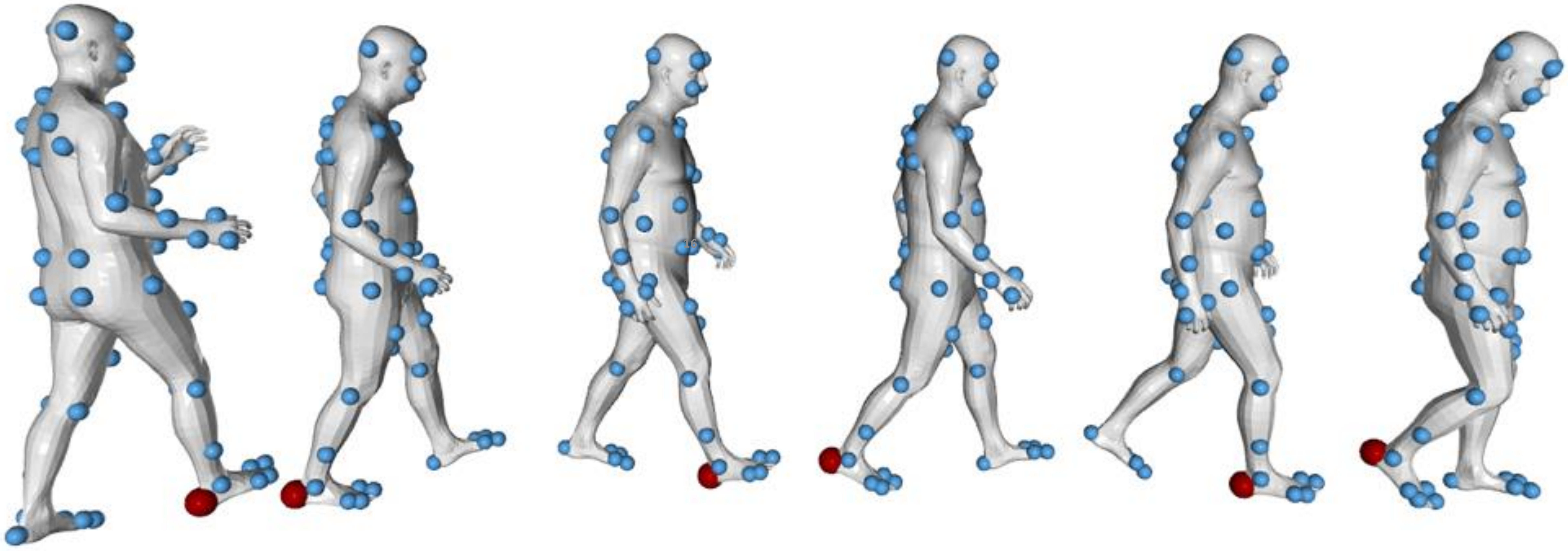
LEMO: Learning Motion Priors

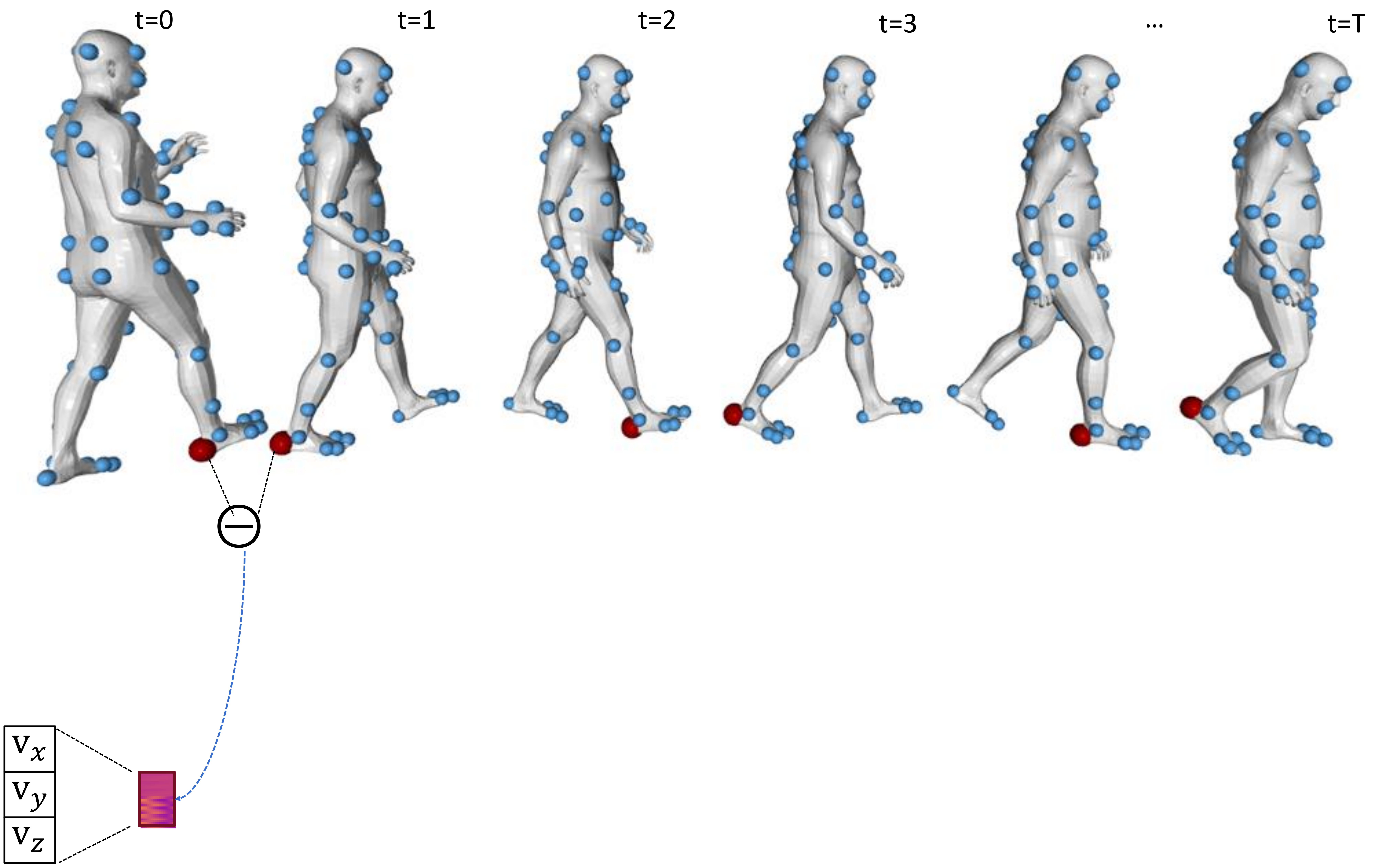
Marker-based motion priors



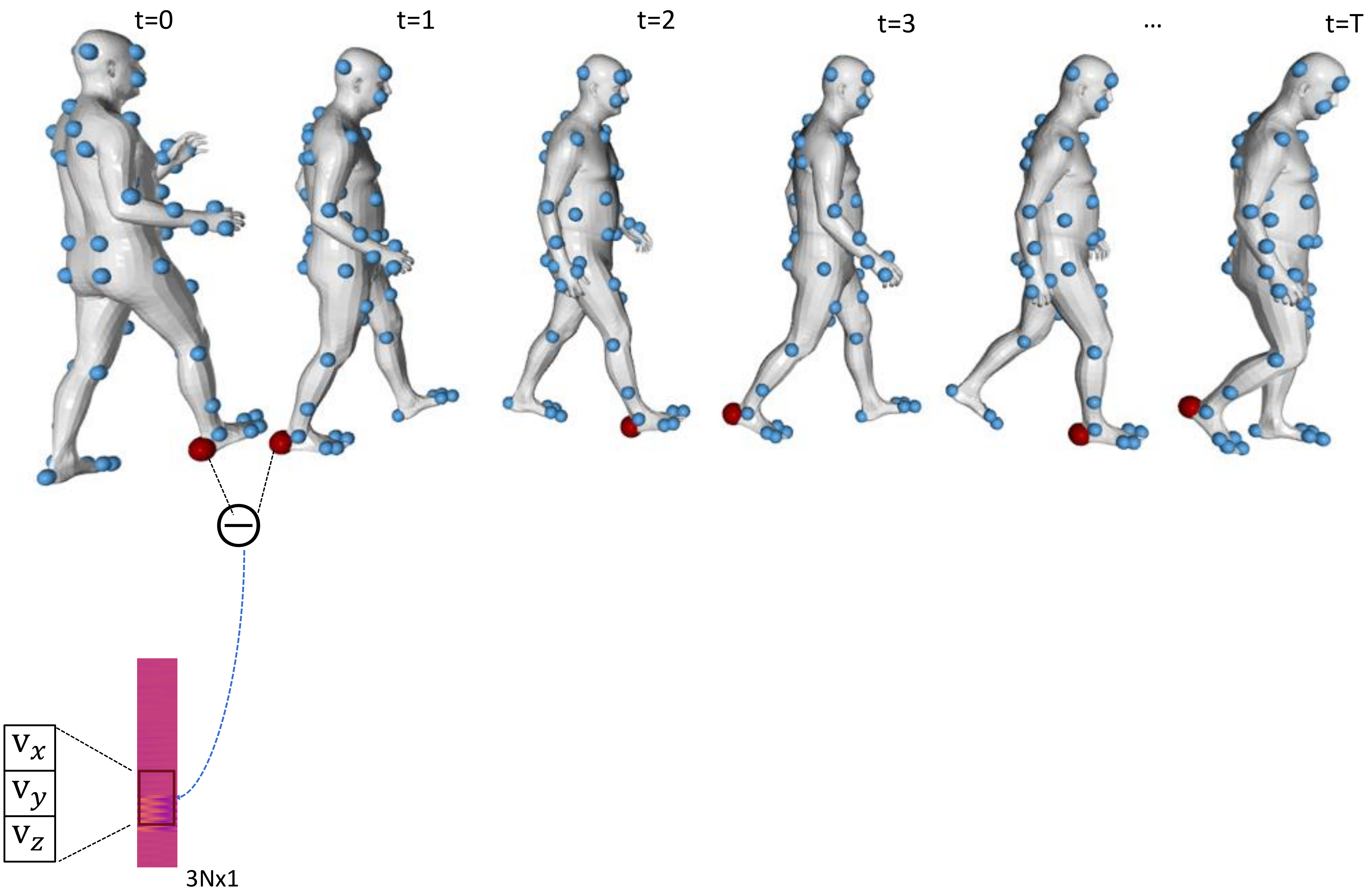
LEMO: Learning Motion Priors

Marker-based motion priors

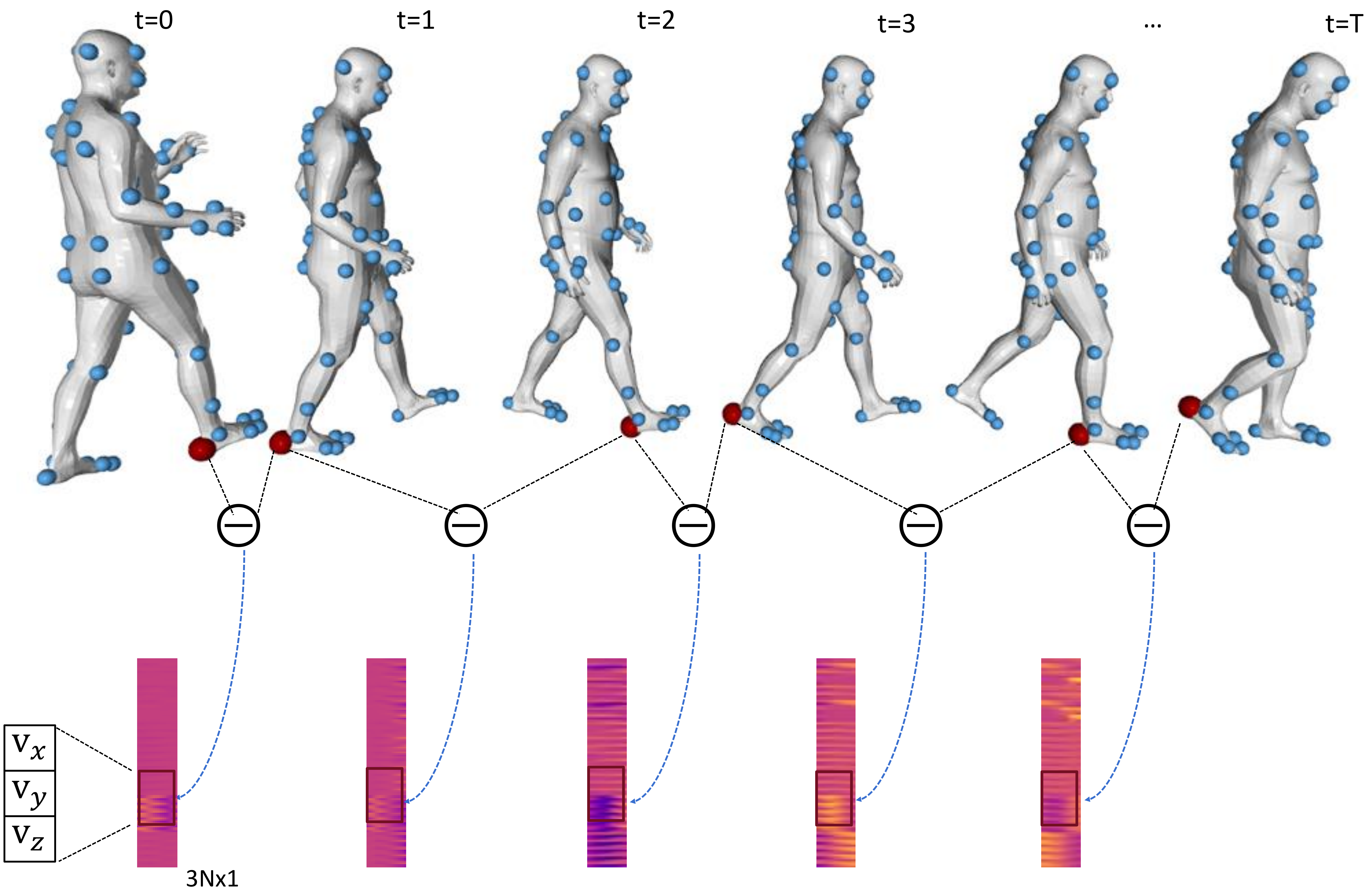




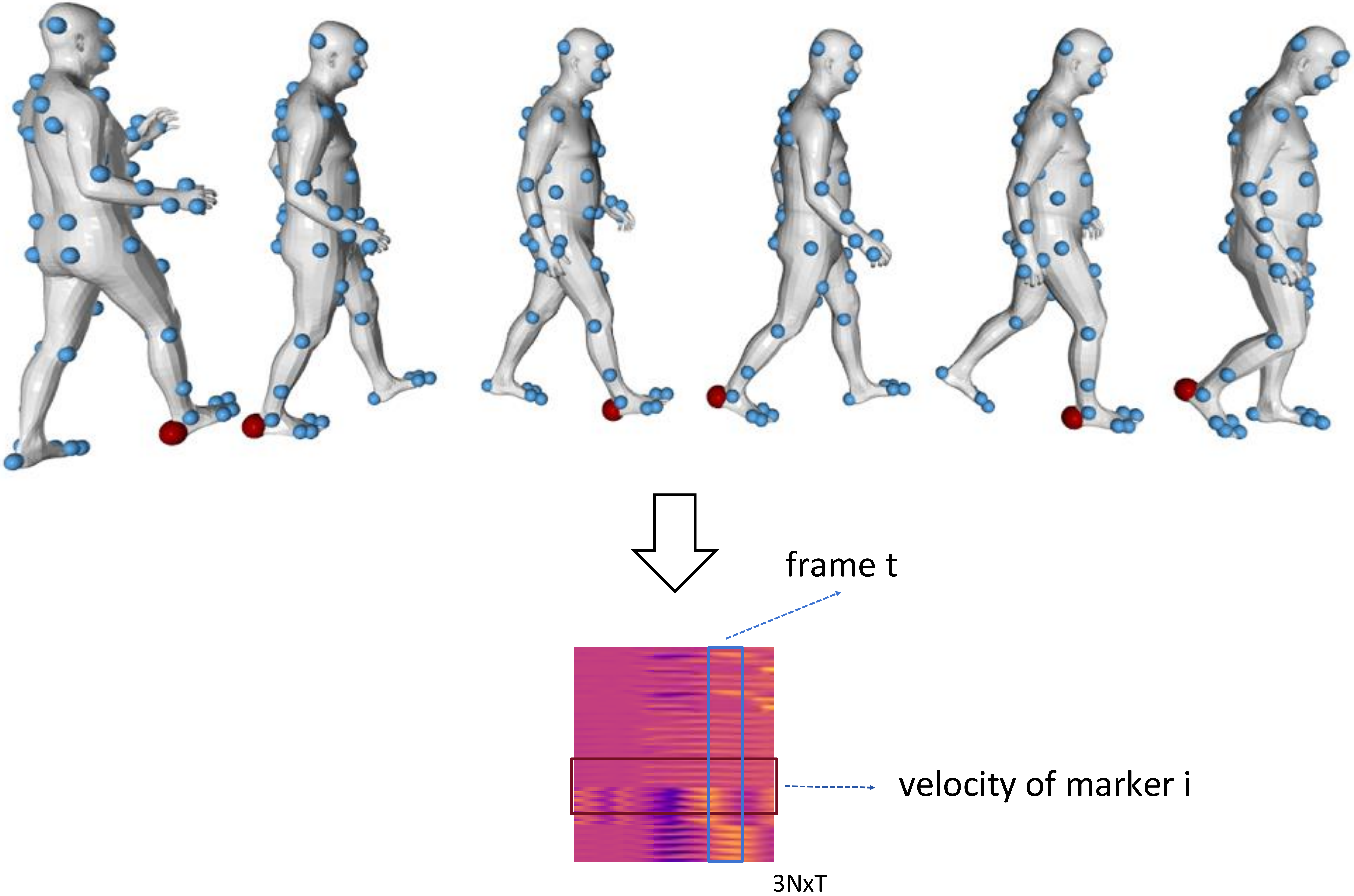
t=0
t=1
t=2
t=3
...
t=T
v_x
v_y
v_z



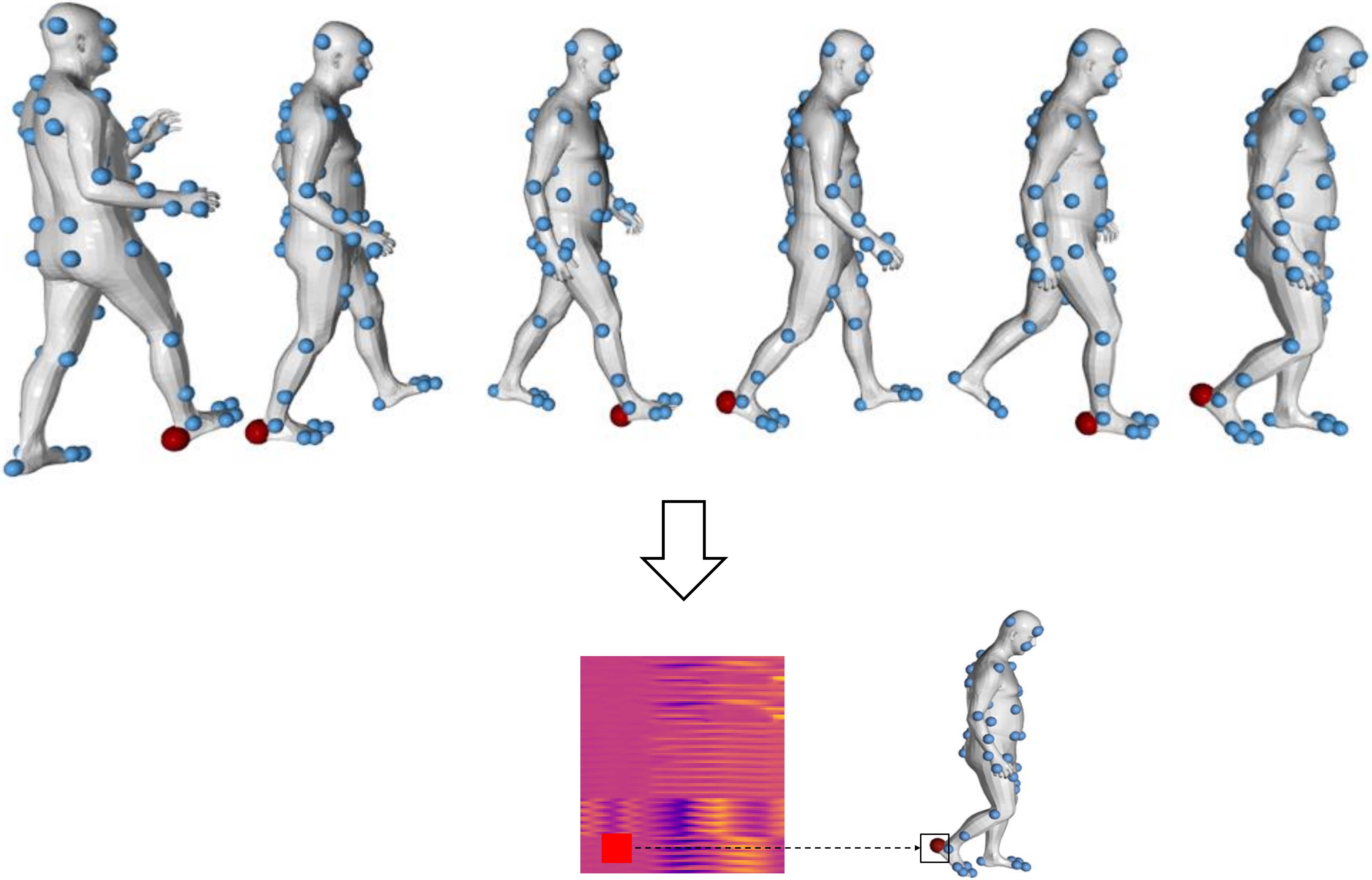
t=0
t=1
t=2
t=3
...
t=T
v_x
v_y
v_z
3Nx1

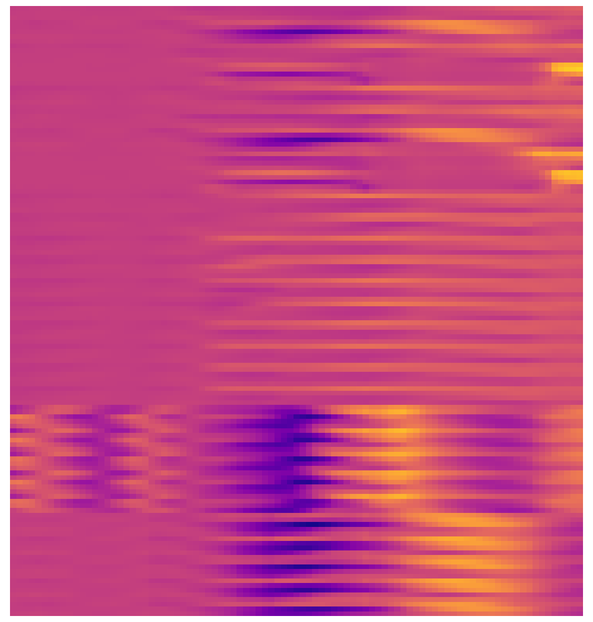


t=0
t=1
t=2
t=3
...
t=T
v_x
v_y
v_z
3Nx1



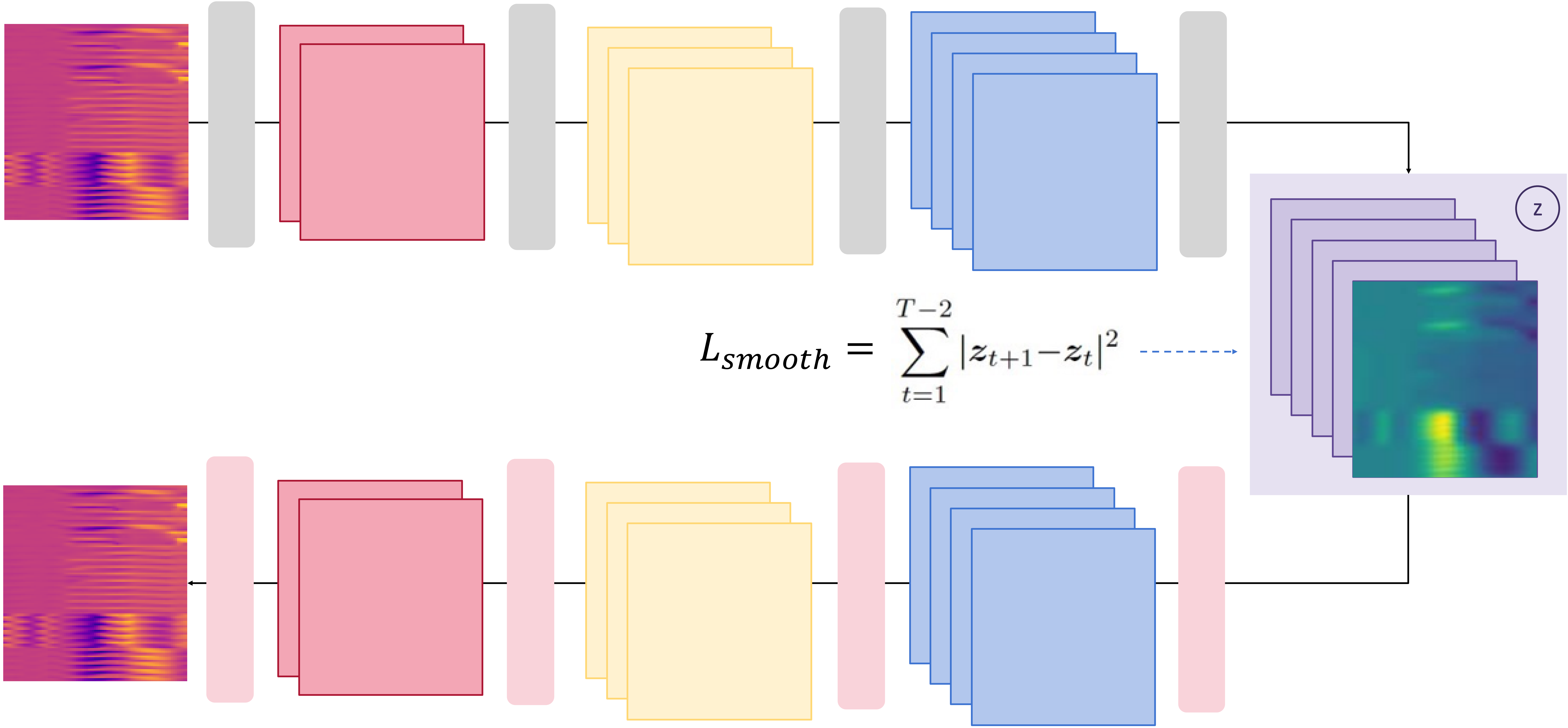
frame t
velocity of marker i
3NxT



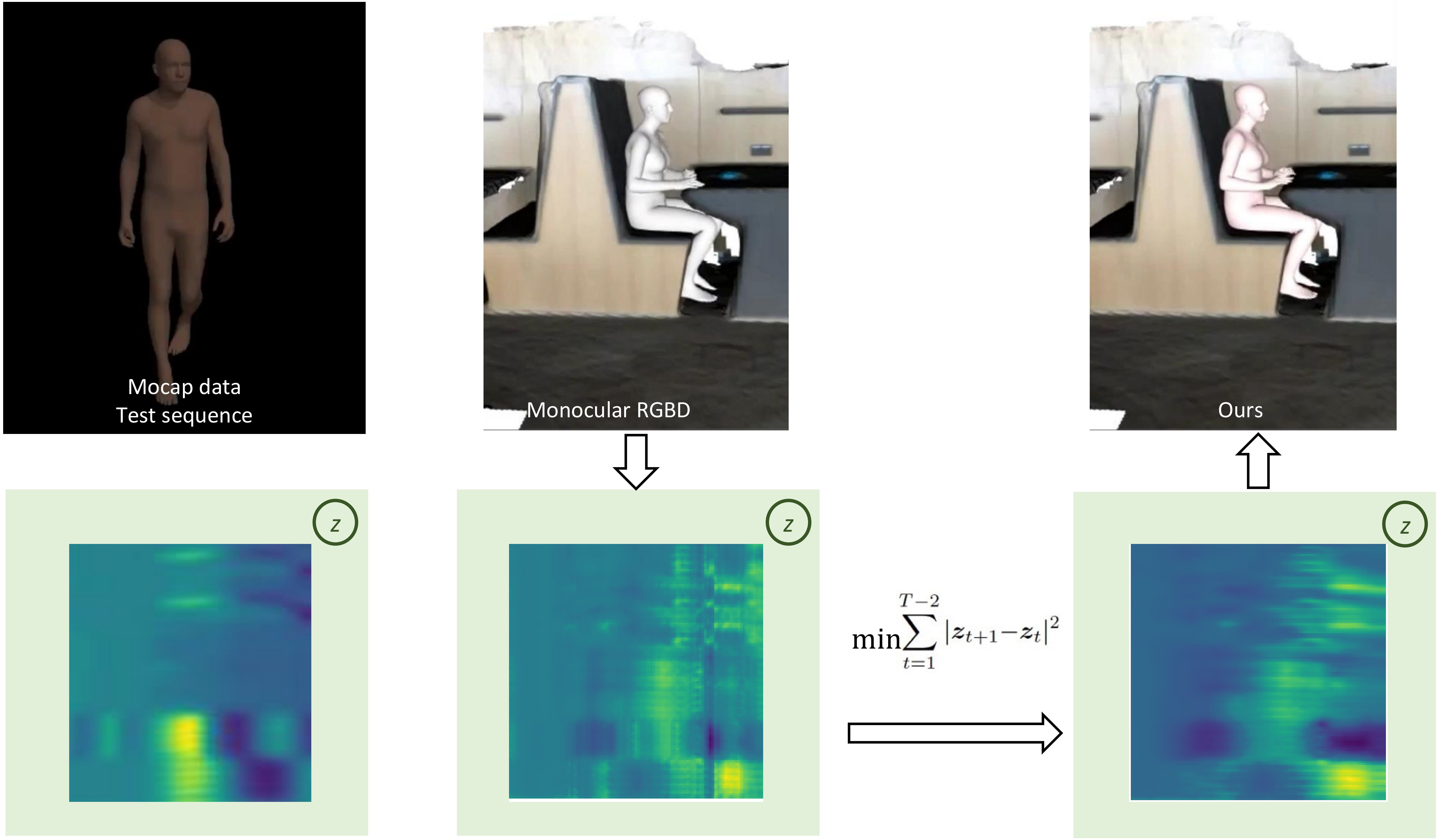


LEMO: Learning Motion Priors

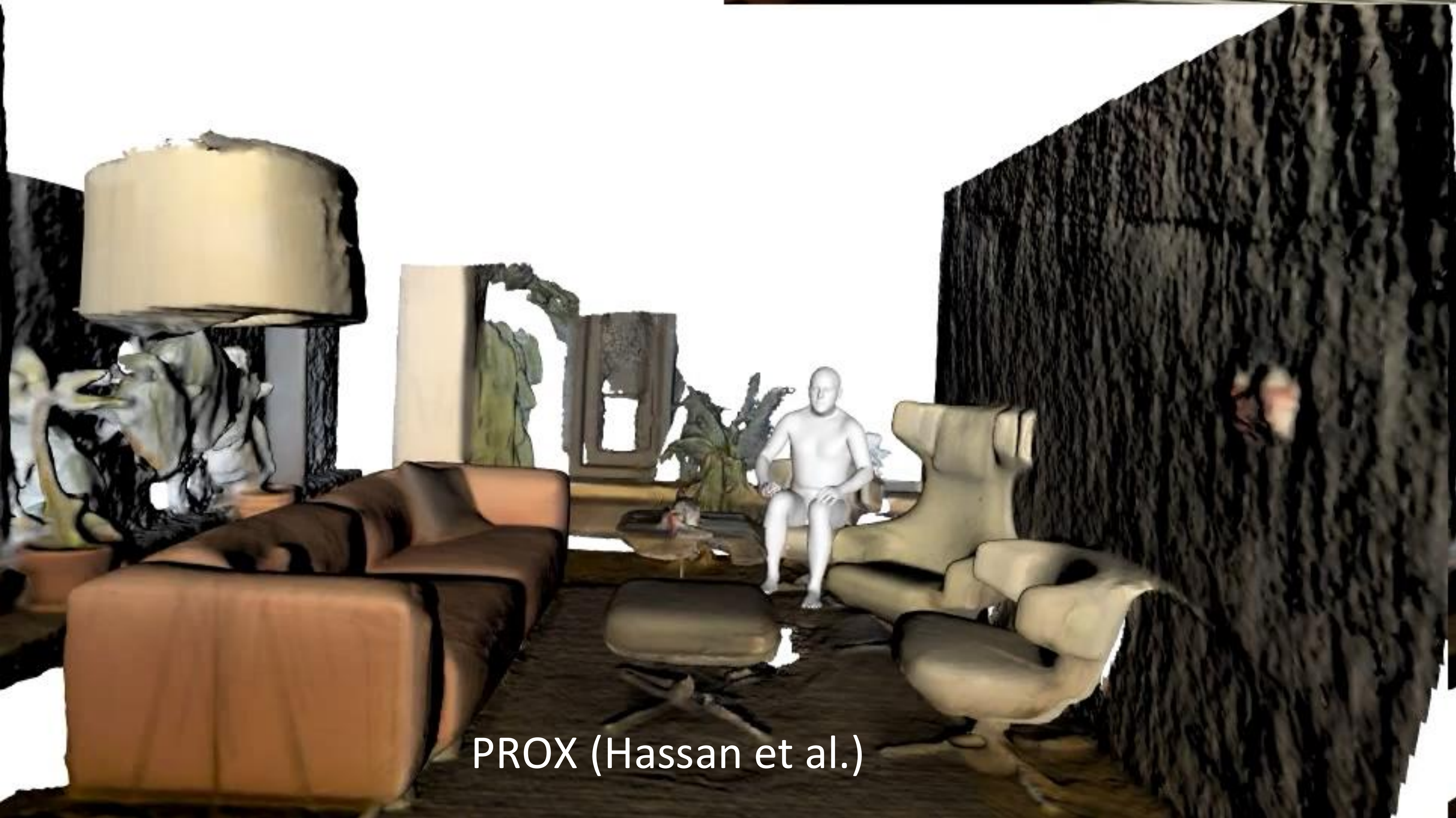
Marker-based motion priors



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

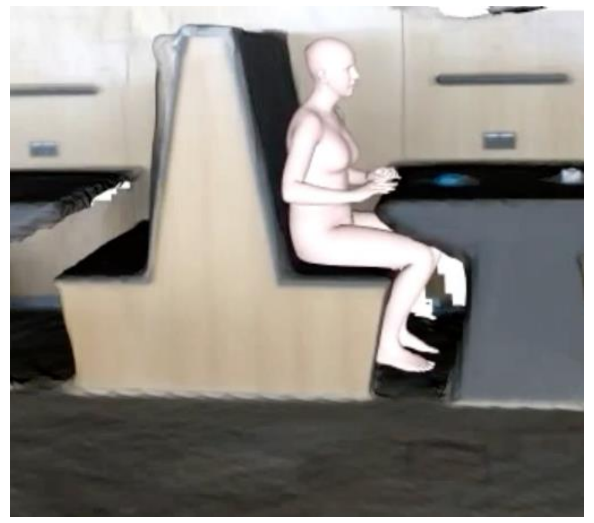

Mocap data
Test sequence
Monocular RGBD
Ours
z
z
z
$\min \sum_{t=1}^{T-2} |z_{t+1} - z_t|^2$

LEMO Results



Our journey on 3D human body and motion estimation

Deep motion prior learning



LEMO
ICCV 2021 (**oral**)
acceptance rate **3.3%**

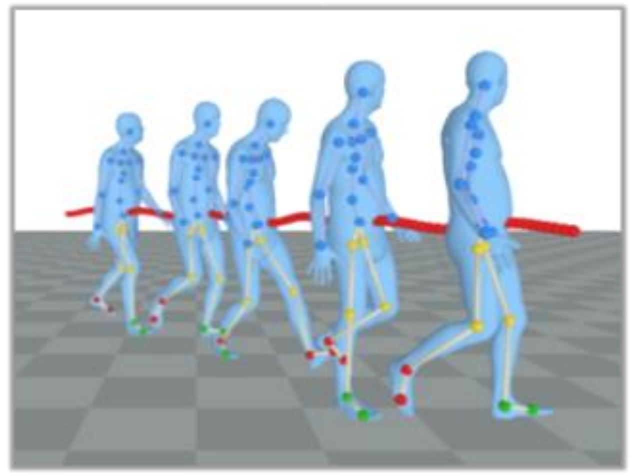
Ego-centric 3D human dataset

EgoBody
ECCV 2022

Ego-centric 3D human body estimation

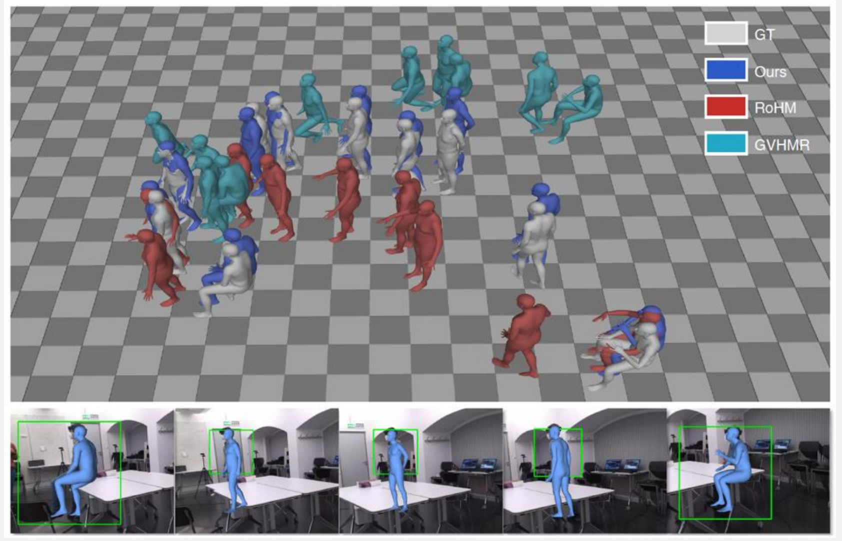
EgoHMR
ICCV 2023 (**oral**)
acceptance rate **2.3%**

Diffusion-based robust human motion estimation

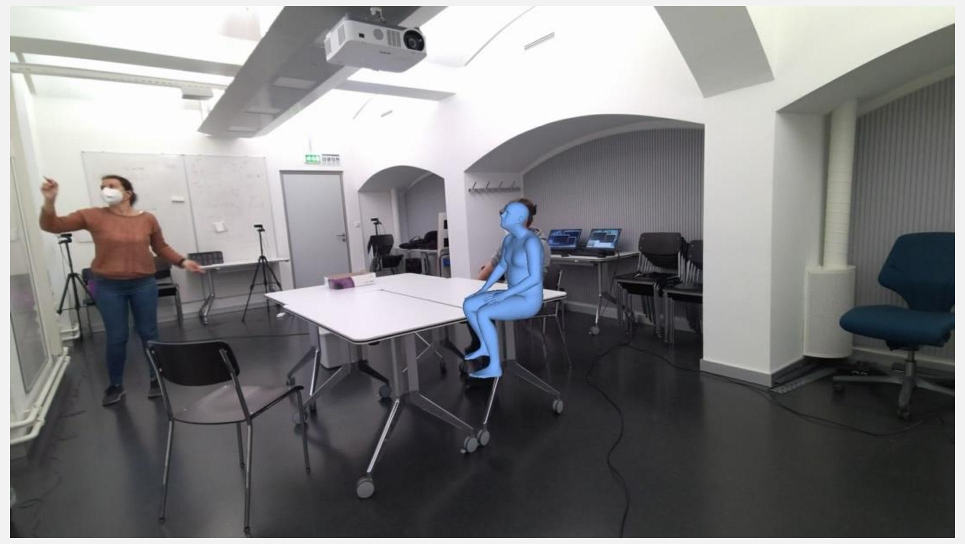


RoHM
CVPR 2024 (**oral**)
acceptance rate **0.78%**

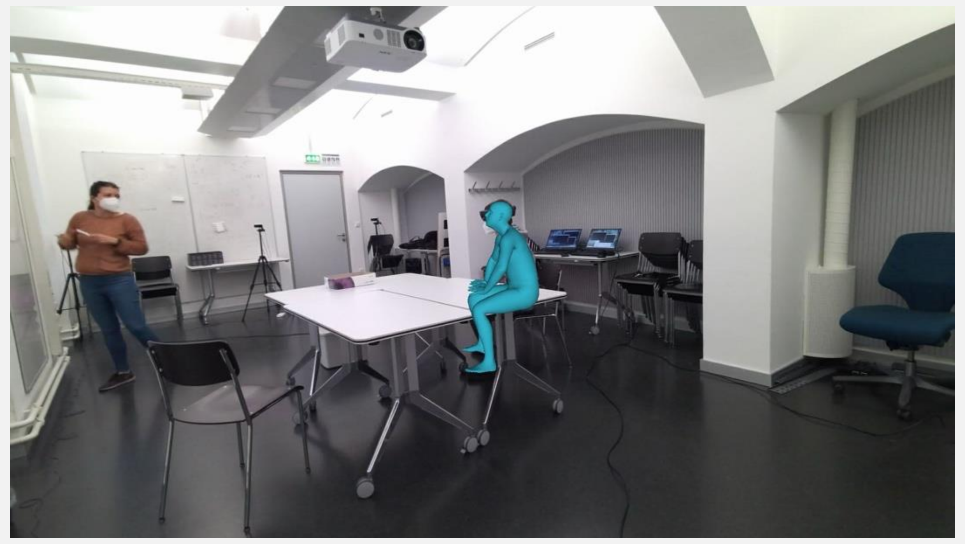
Efficient feedforward human motion estimation



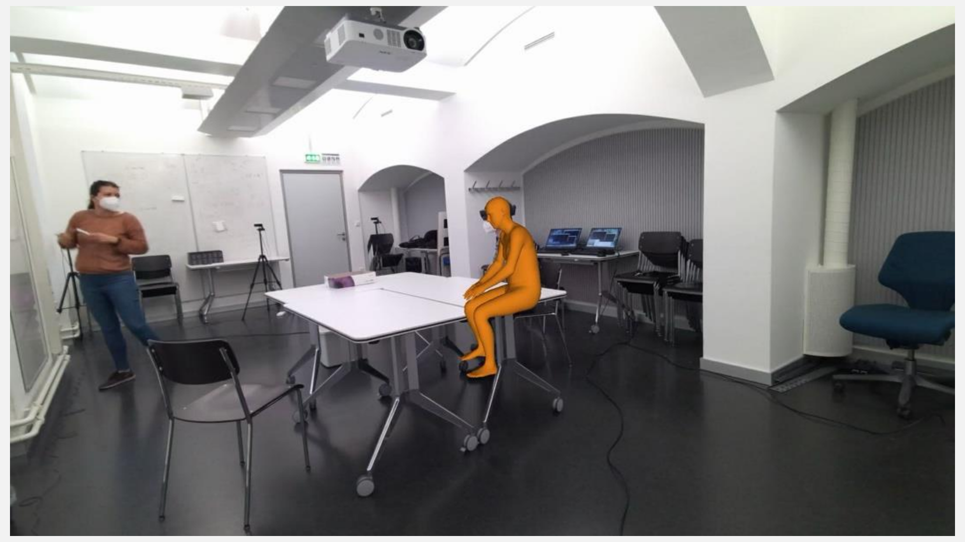
MoRo
3DV 2026



Ours



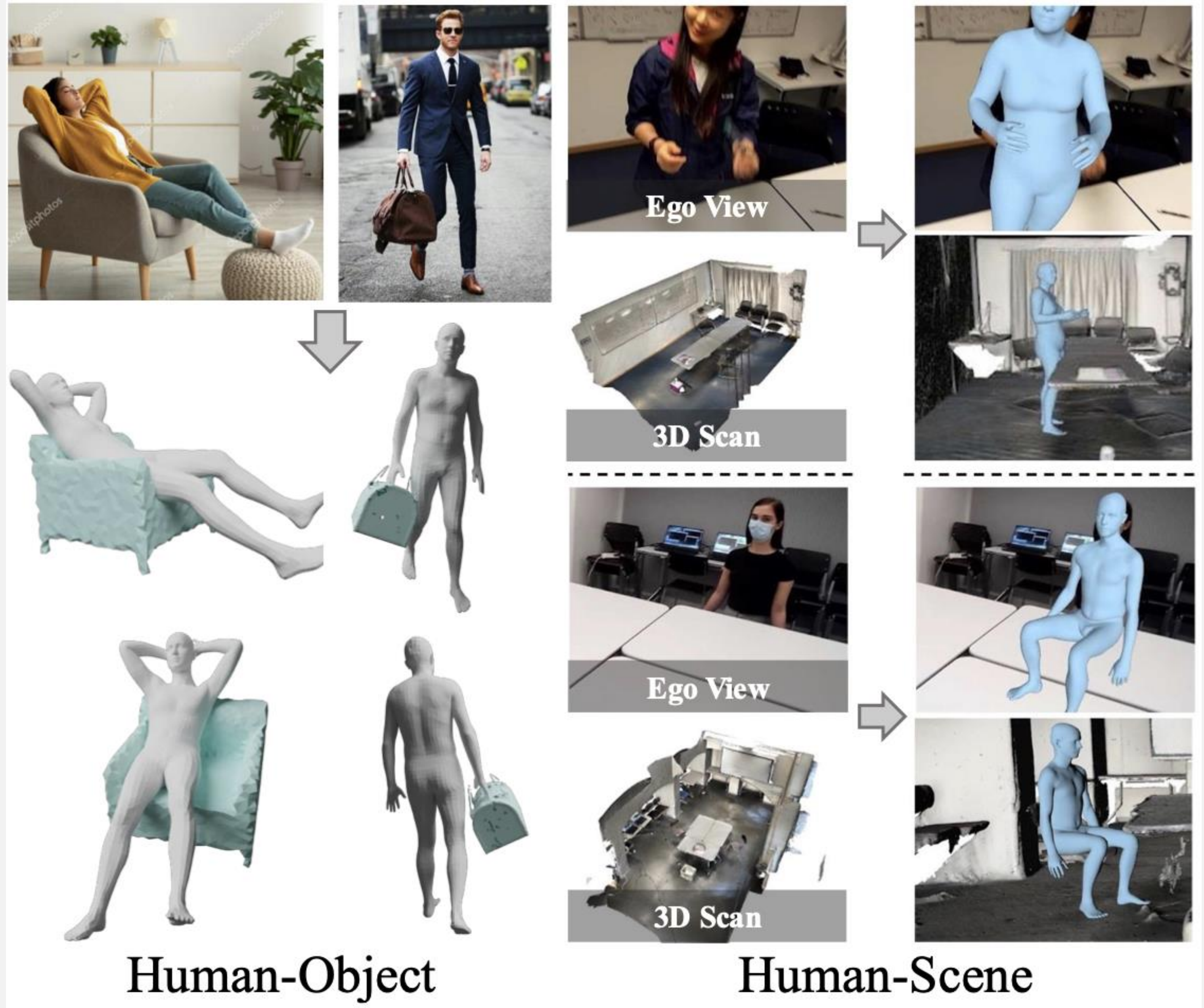
GVHMR, SIGGRAPH Asia 2024



WHAM, CVPR 2024

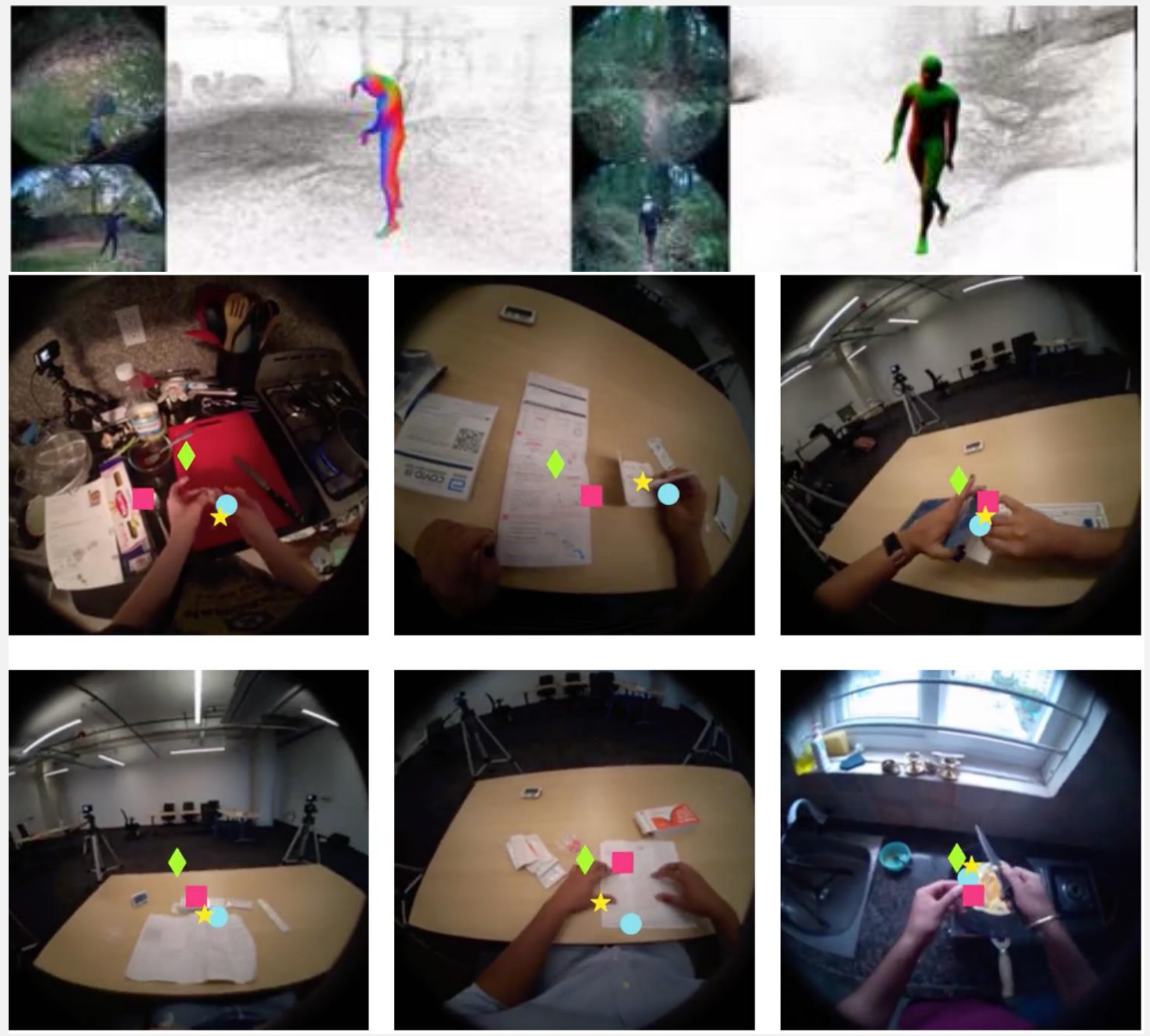
Research overview

1. Perceiving and reconstructing *real* humans



2. Synthesizing *virtual* humans

3. Embodied *digital* humans



4. Unified human foundation models that are 3D-grounded and multimodally capable.

- Rich semantics
- Diverse motion and appearance
- Very limited 3D ground truth data

- Rich and accurate 3D ground-truth annotations
- Human behavior synthesis is a really hard problem

- Multi-modality data
- hand-object interaction
- Human motion capture with very limited observations

Part 2: synthesizing virtual humans

Synthesizing Virtual Humans

Computer games? Character animation?

Our goal:

Diverse human behaviors

Text-aligned motion representations

Easy to control to compose long-term complex human activities

Two key ideas

1. Generative motion primitives

2. RL-based motion control

MOJO: We are More than Our Joints: Predicting how 3D Bodies Move

Yan Zhang, Michael Black, Siyu Tang. **CVPR 2021**

The Wanderings of Odysseus in 3D Scenes

Yan Zhang, Siyu Tang. **CVPR 2022**

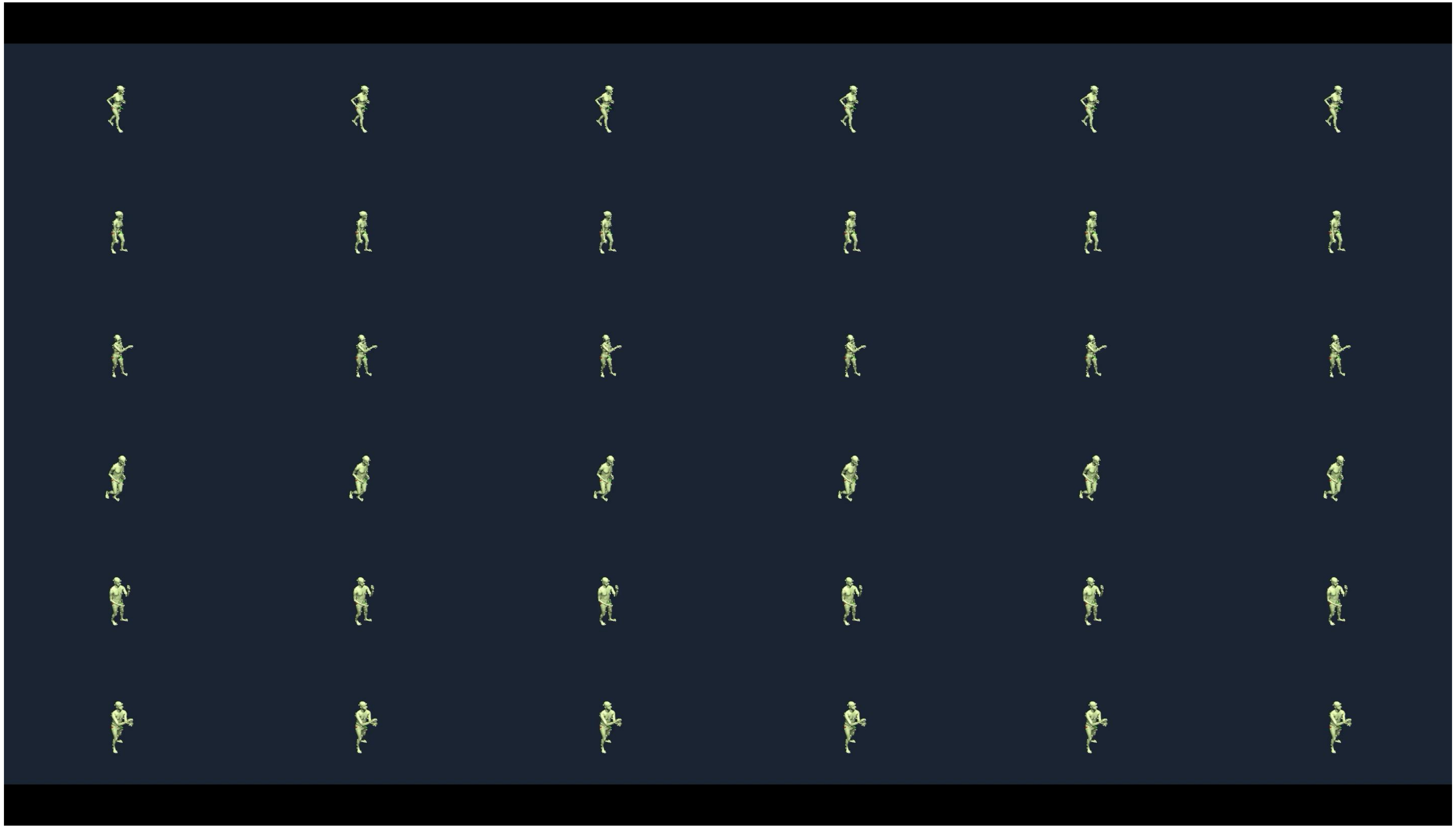
DIMOS: Synthesizing Diverse Human Motion in 3D Indoor Scenes

Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, Siyu Tang. **ICCV 2023**

DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control

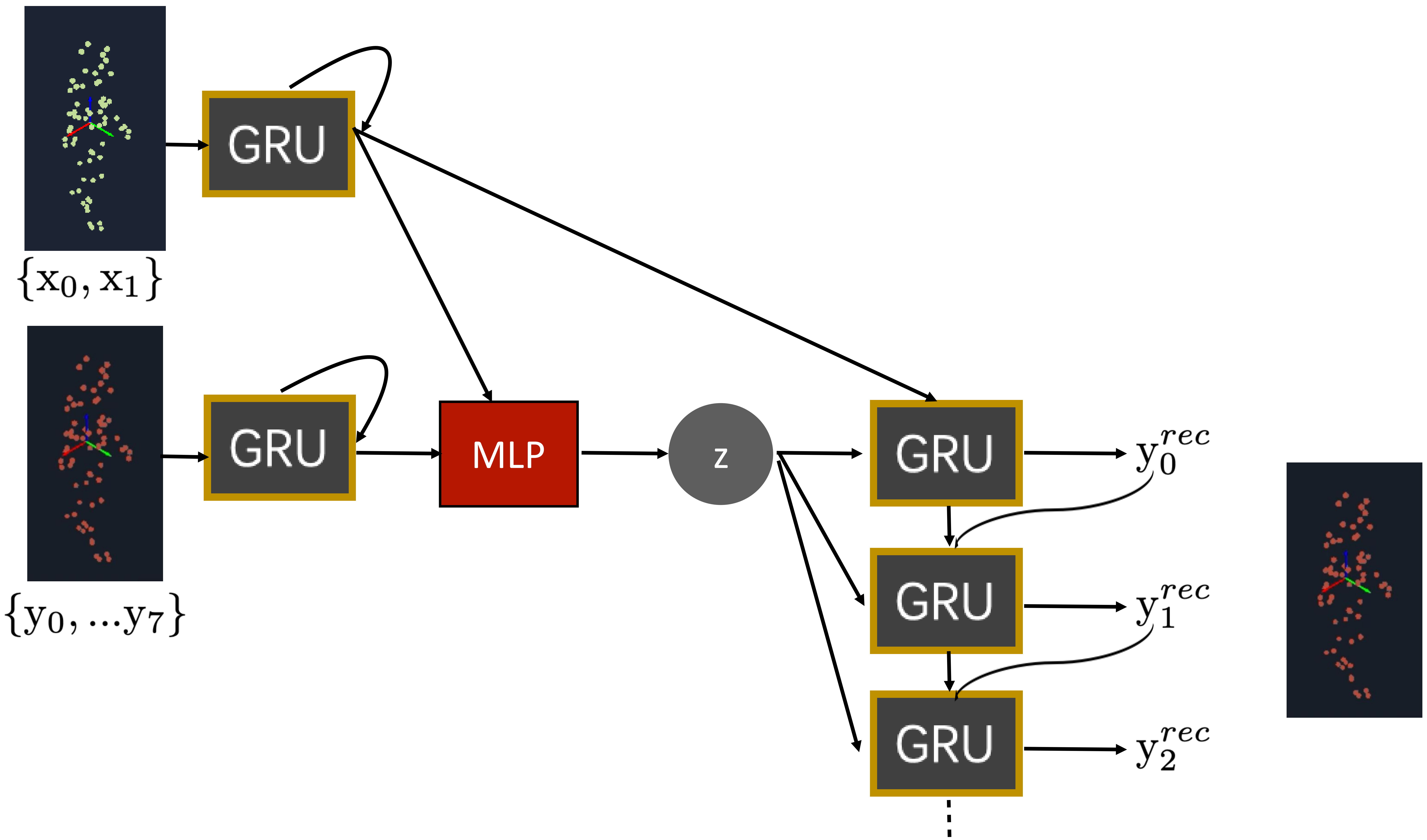
Kaifeng Zhao, Gen Li, Siyu Tang. **ICLR 2025, Spotlight**

Generative Motion Primitives

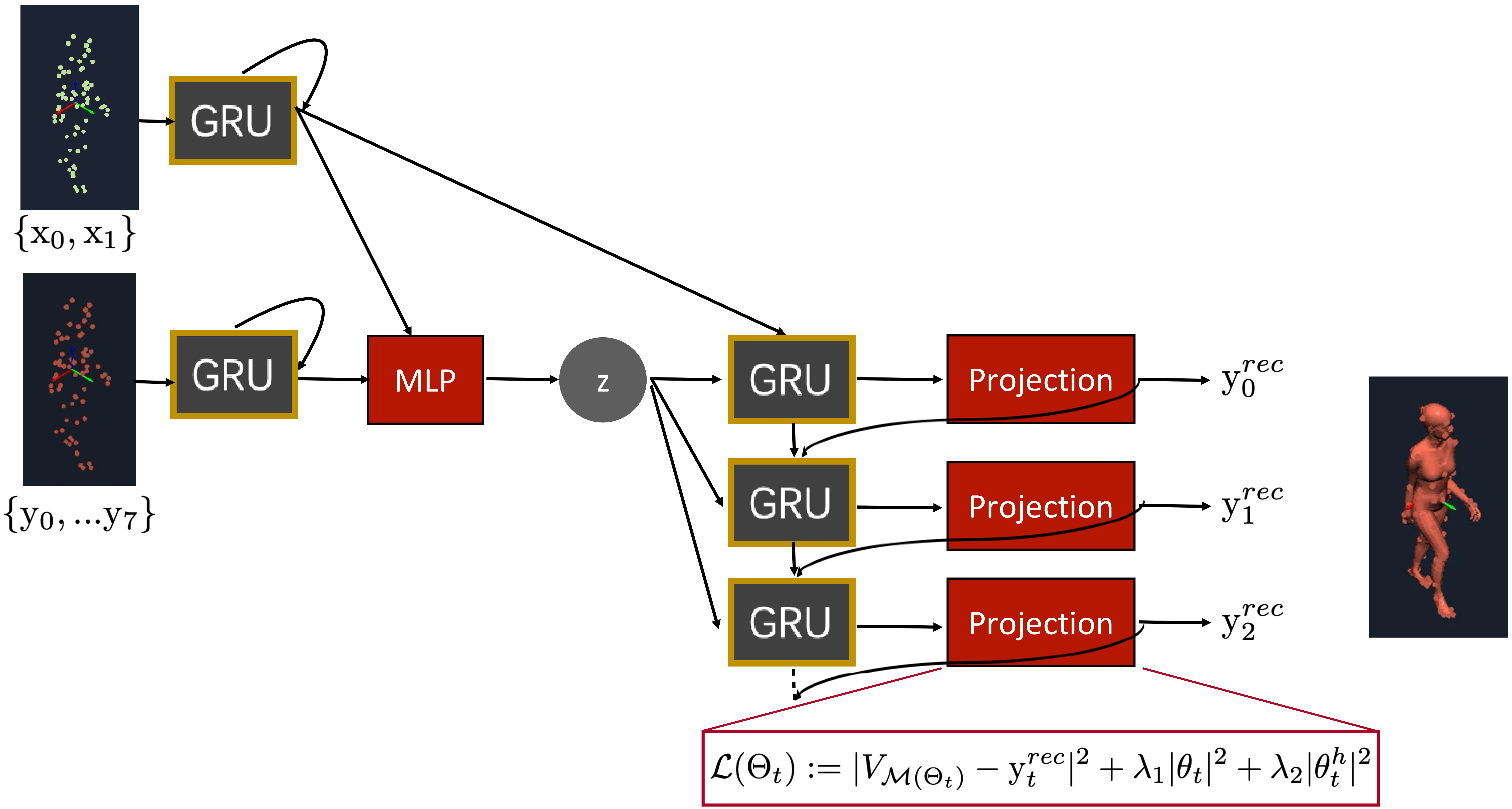


< 1 second

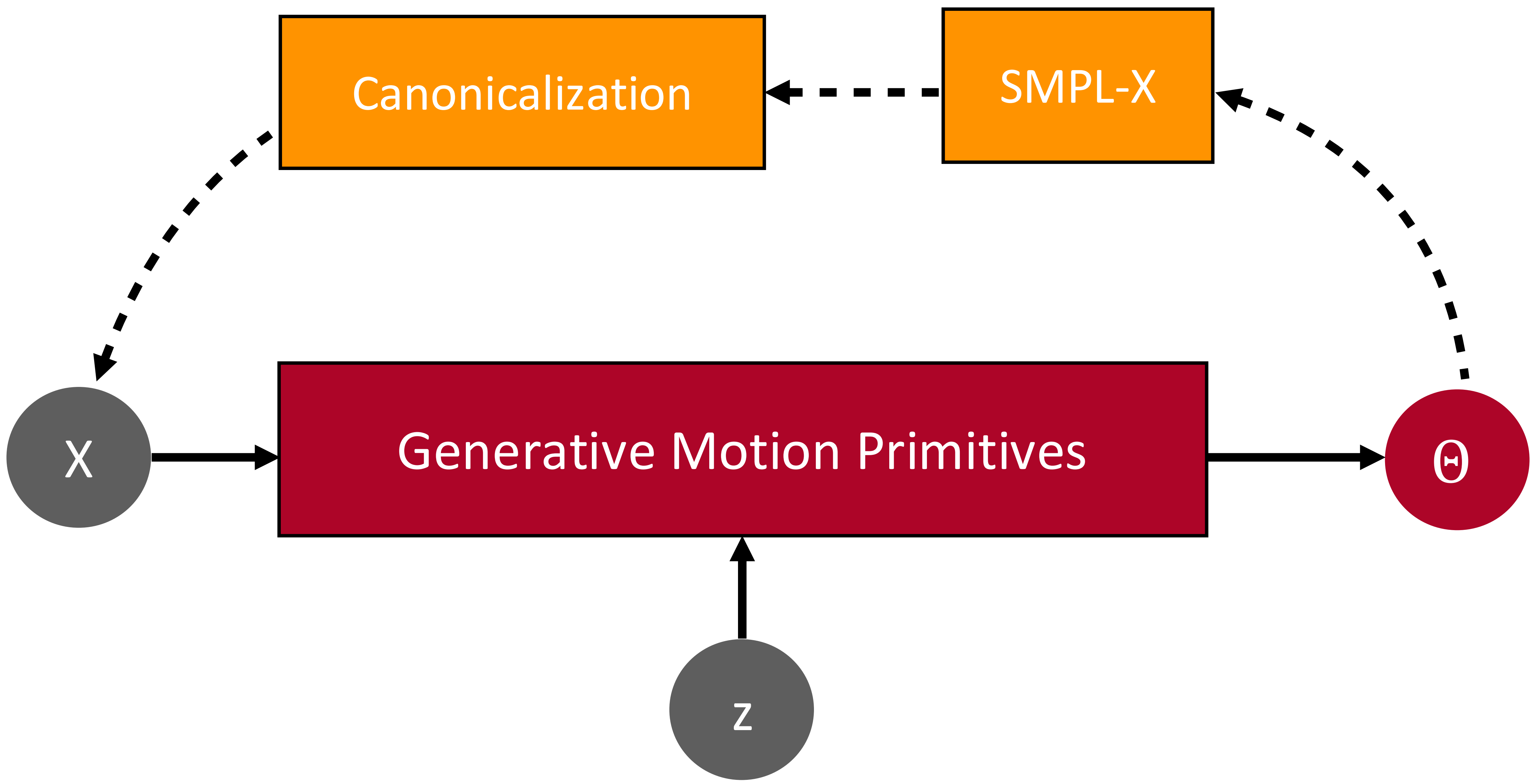
Generative Motion Primitives



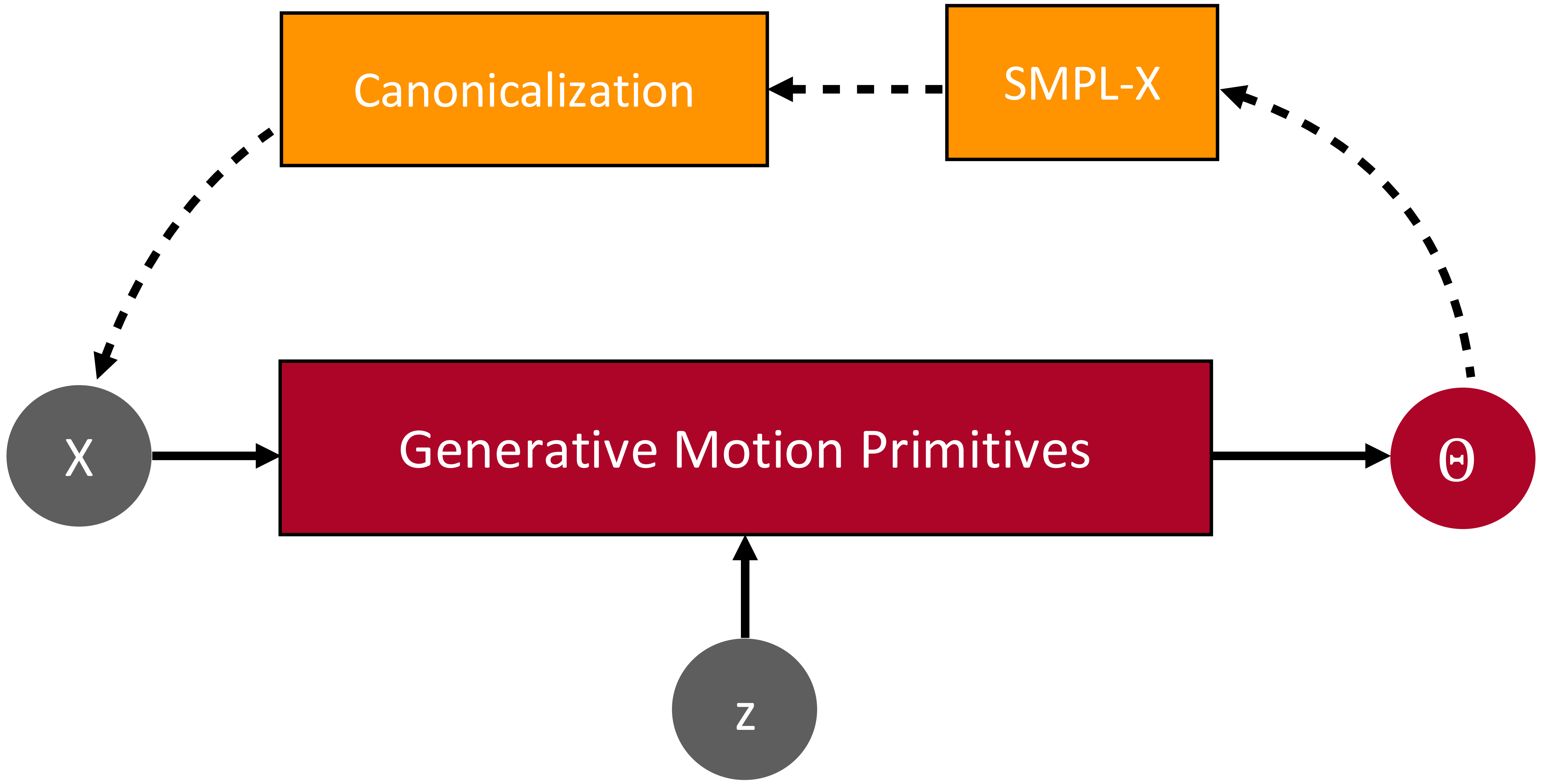
Generative Motion Primitives



Generative Motion Primitives



Generative Motion Primitives

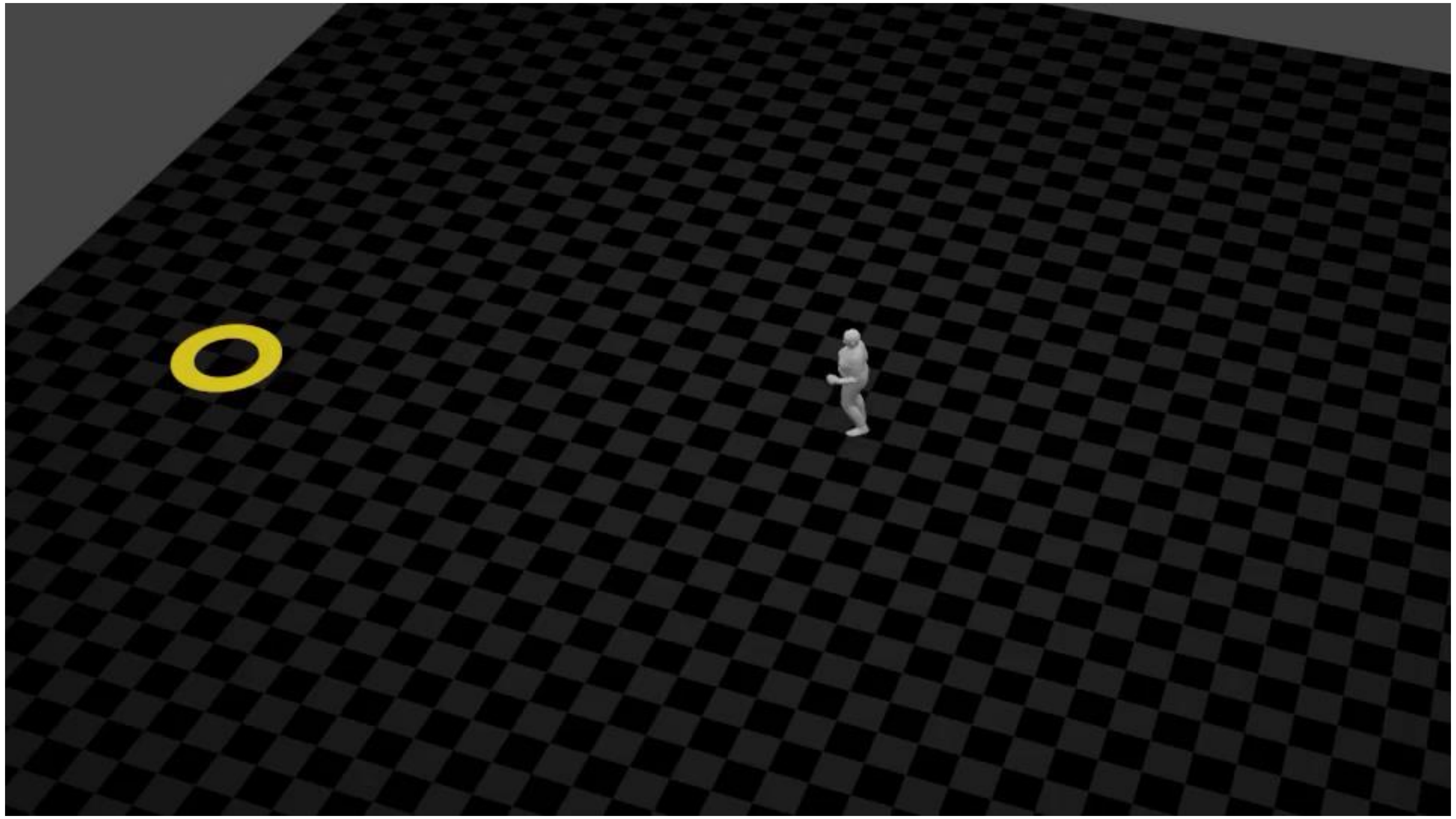


Two key ideas

Generative motion primitives

RL-based motion control

Controllable Motion Generation



Our goal: human motion is guided by the waypoint on the ground.

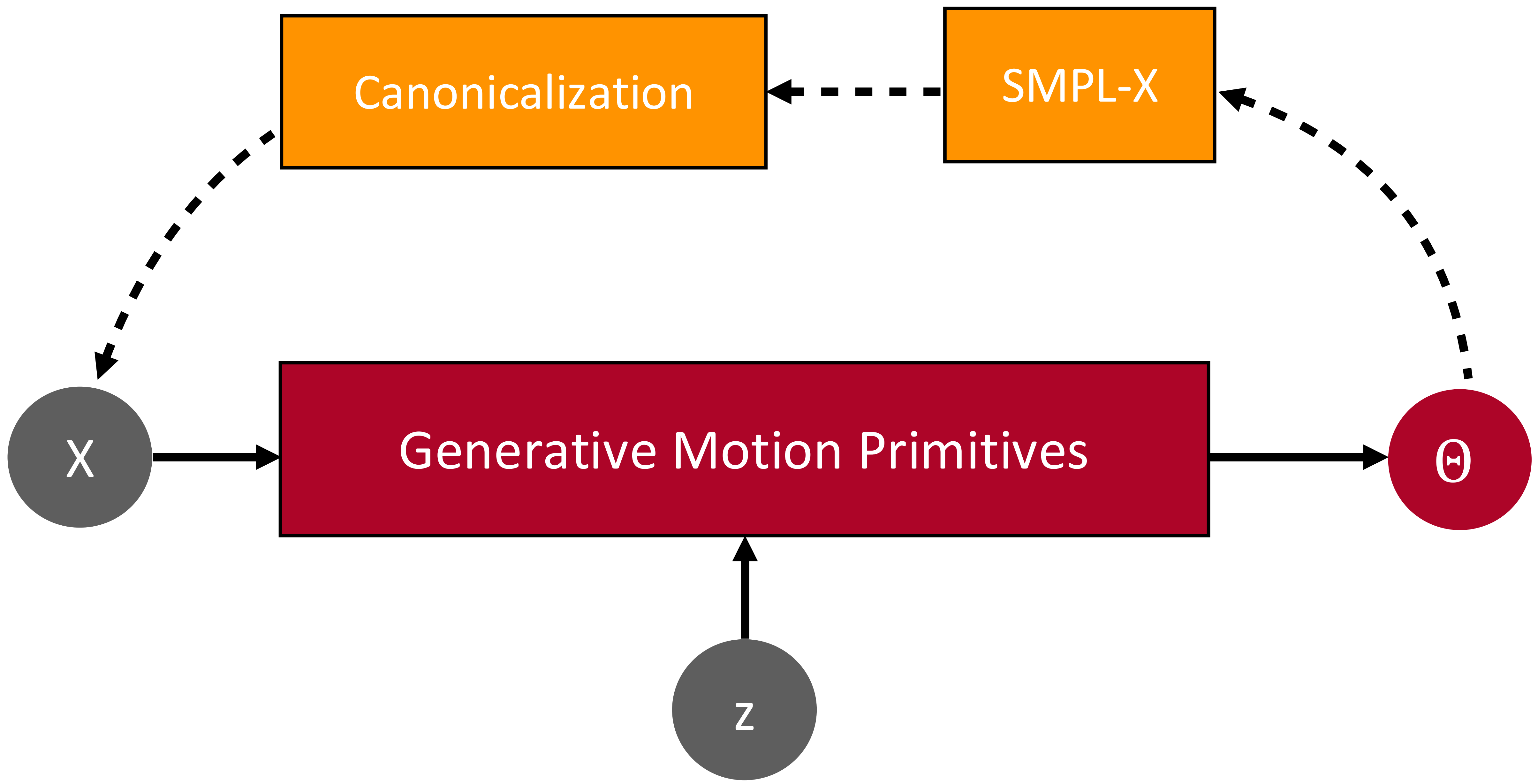
Without control, the 3D humans move randomly and ignores the waypoints.

RL-based motion control

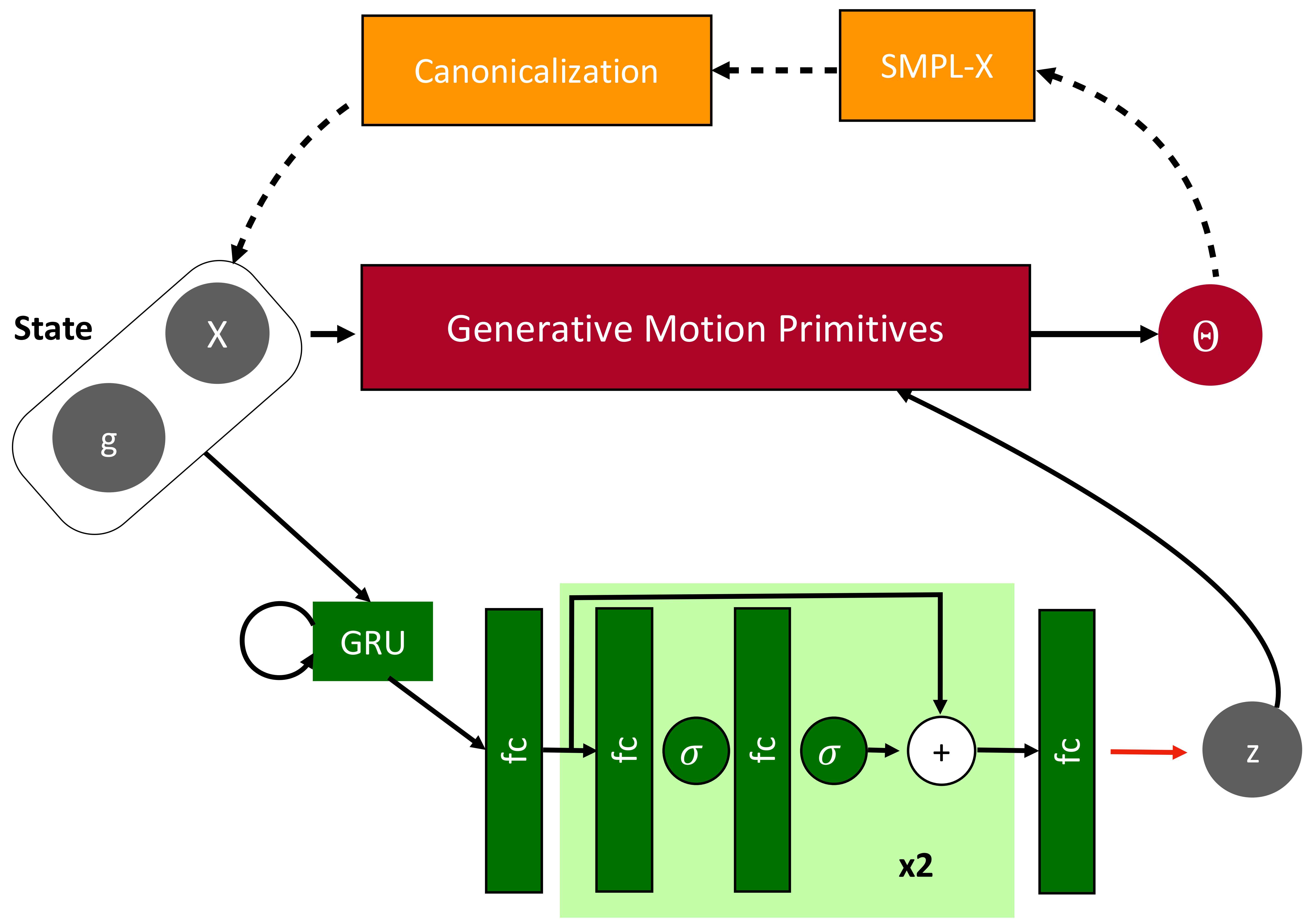
- **State**: the initial body and its distance to the goal.
- **Model**: generative motion primitives models.
- **Action**: latent variable of the generative model.
- **Policy**: a neural network mapping from the **State** to the **Action.**

Character controllers using motion VAEs . ACM Transactions on Graphics, 2020. HungYu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne

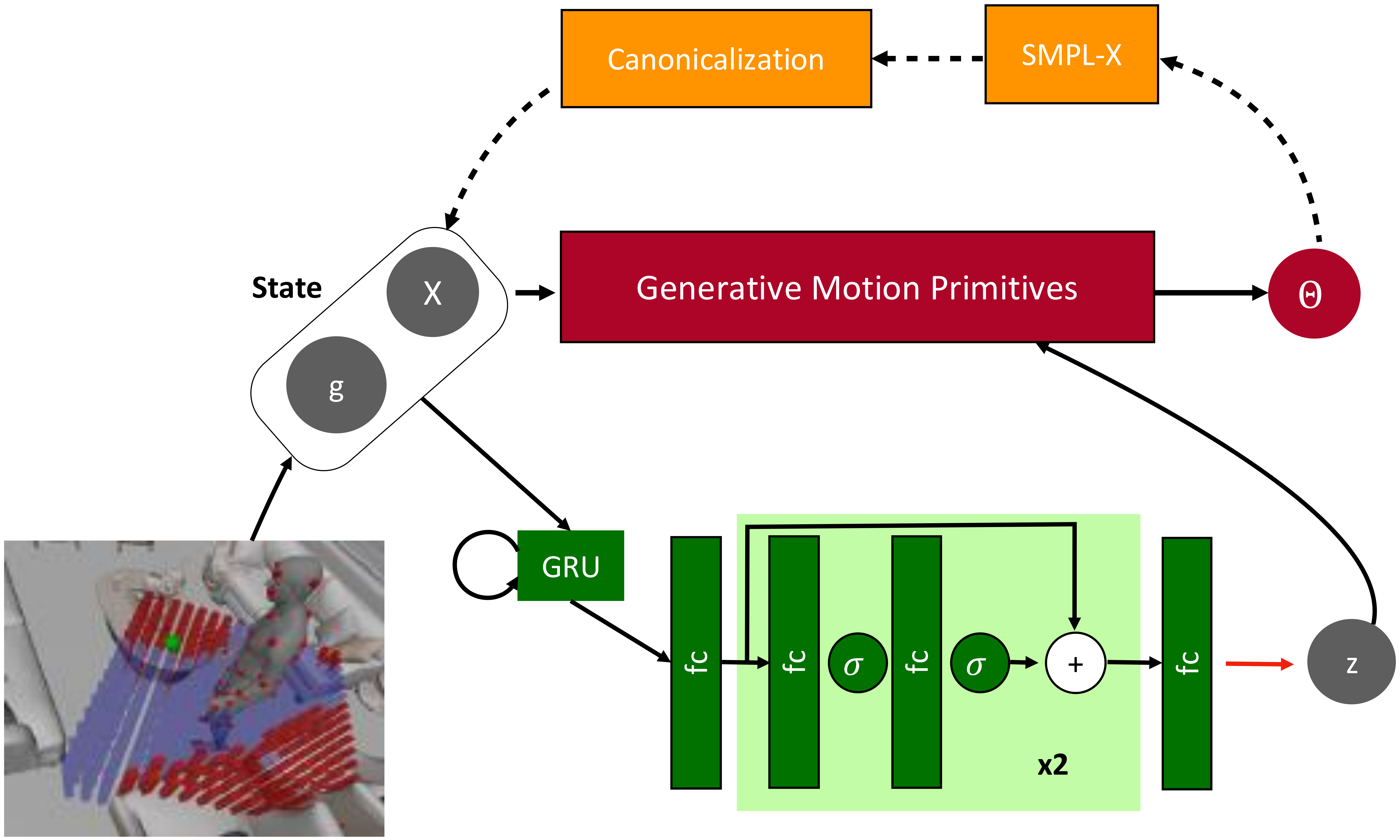
Controllable Motion Generation



Controllable Motion Generation



Scene-Aware Motion Control



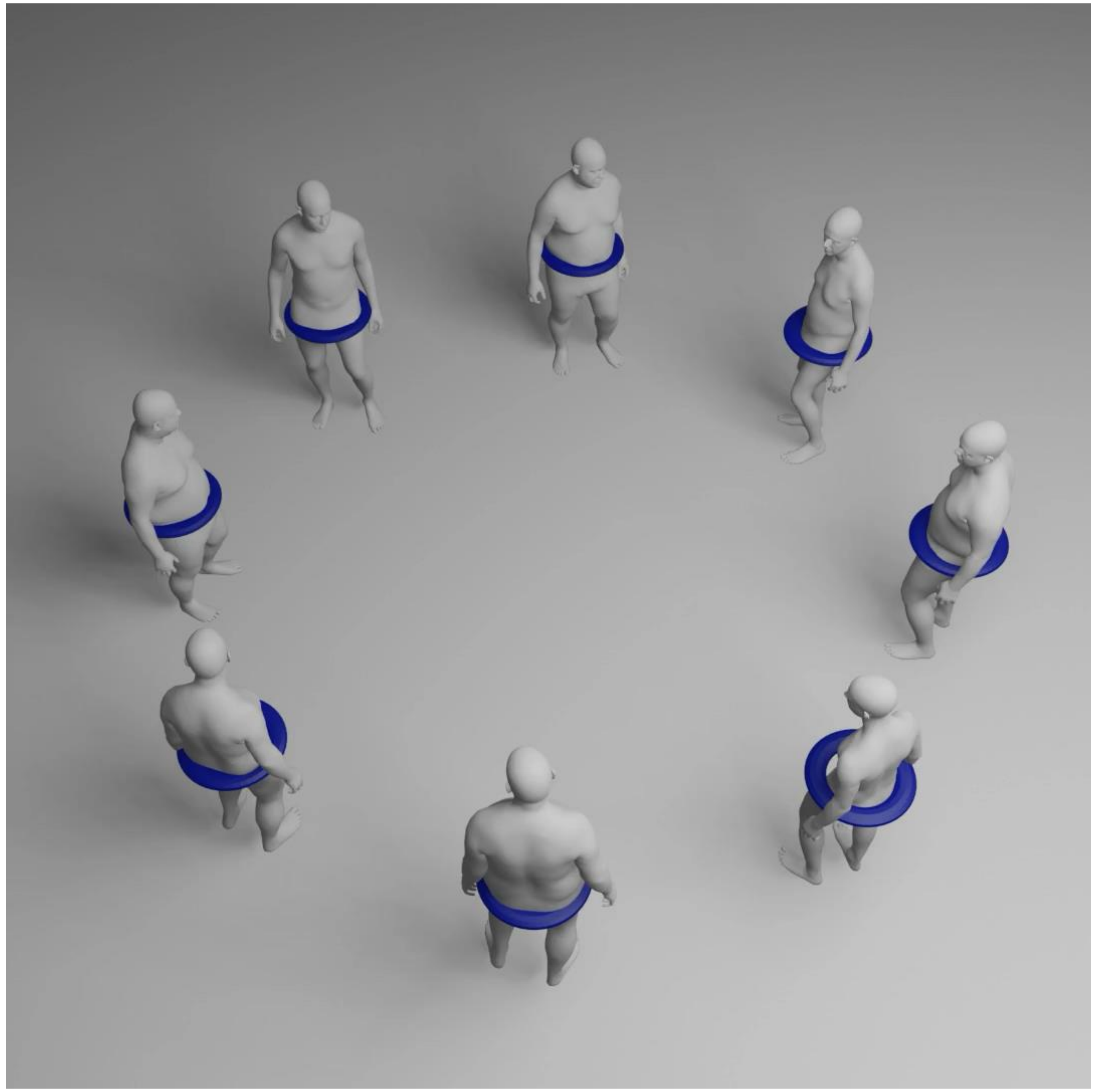
Synthesizing human-scene interaction

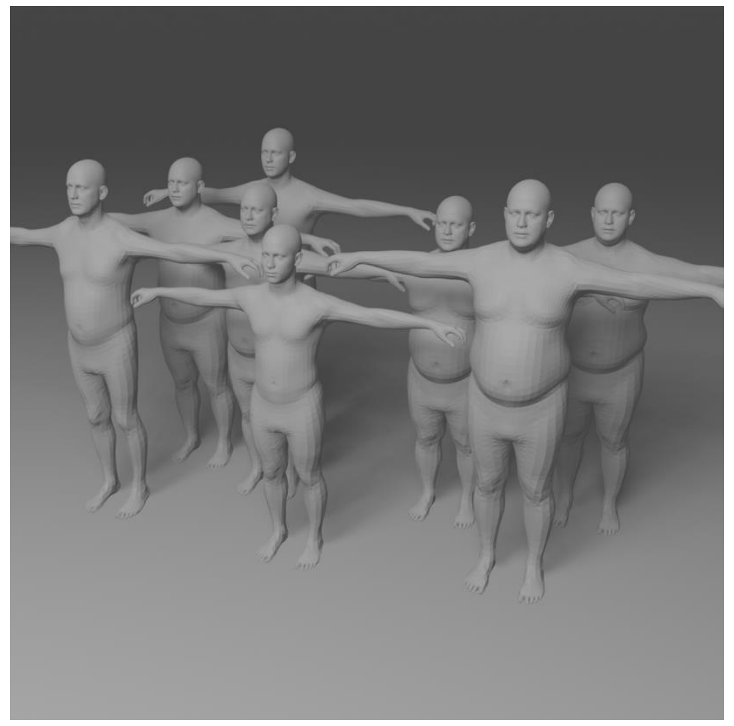
DIMOS: Synthesizing Diverse Human Motion in 3D Indoor Scenes

K. Zhao, Y. Zhang, S. Wang, T. Beeler, S. Tang. **ICCV 2023**

Synthesizing multi-human goal-reaching behaviors

PhysicsVAE [Won et al. SIGGRAPH 2022]





Ours (**EgoGen**)

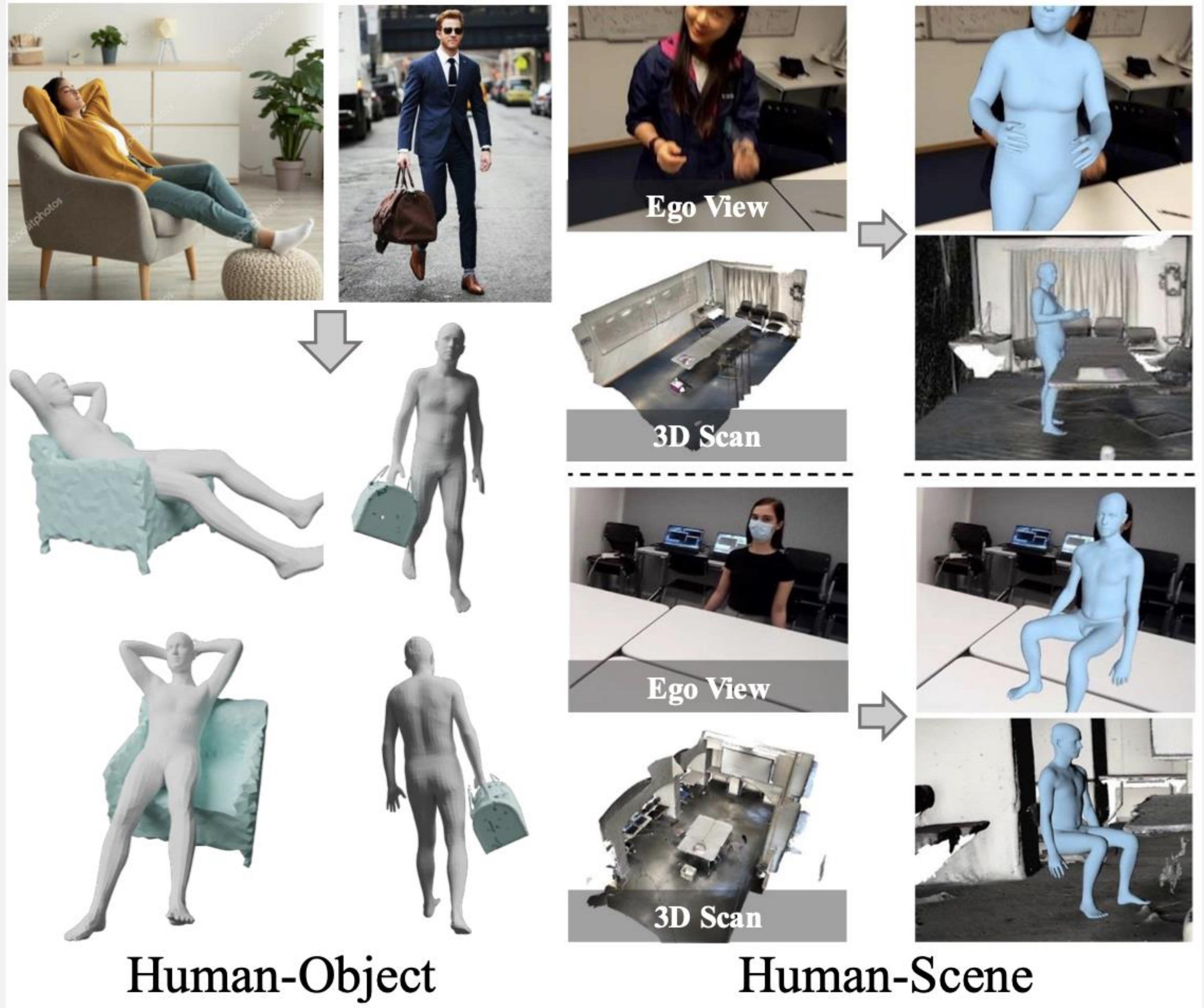
EgoGen: An Egocentric Synthetic Data Generator

G. Li, K. Zhang S. Zhang X. Lyu, M. Dusmanu, Y. Zhang, M. Pollefeys, S. Tang.

CVPR 2024 Oral, top 90 out of 11532 submissions

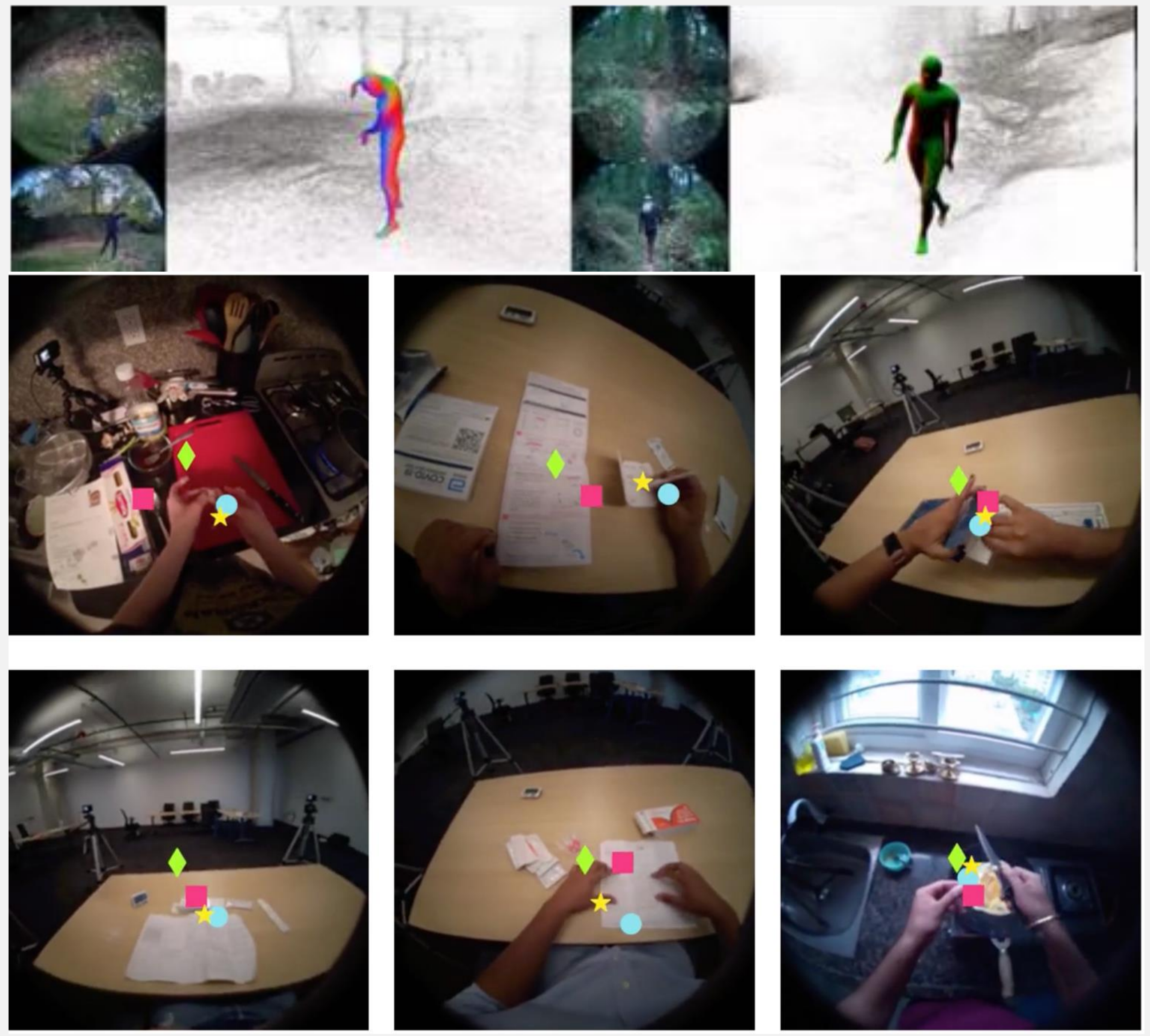
Research overview

1. Perceiving and reconstructing *real* humans

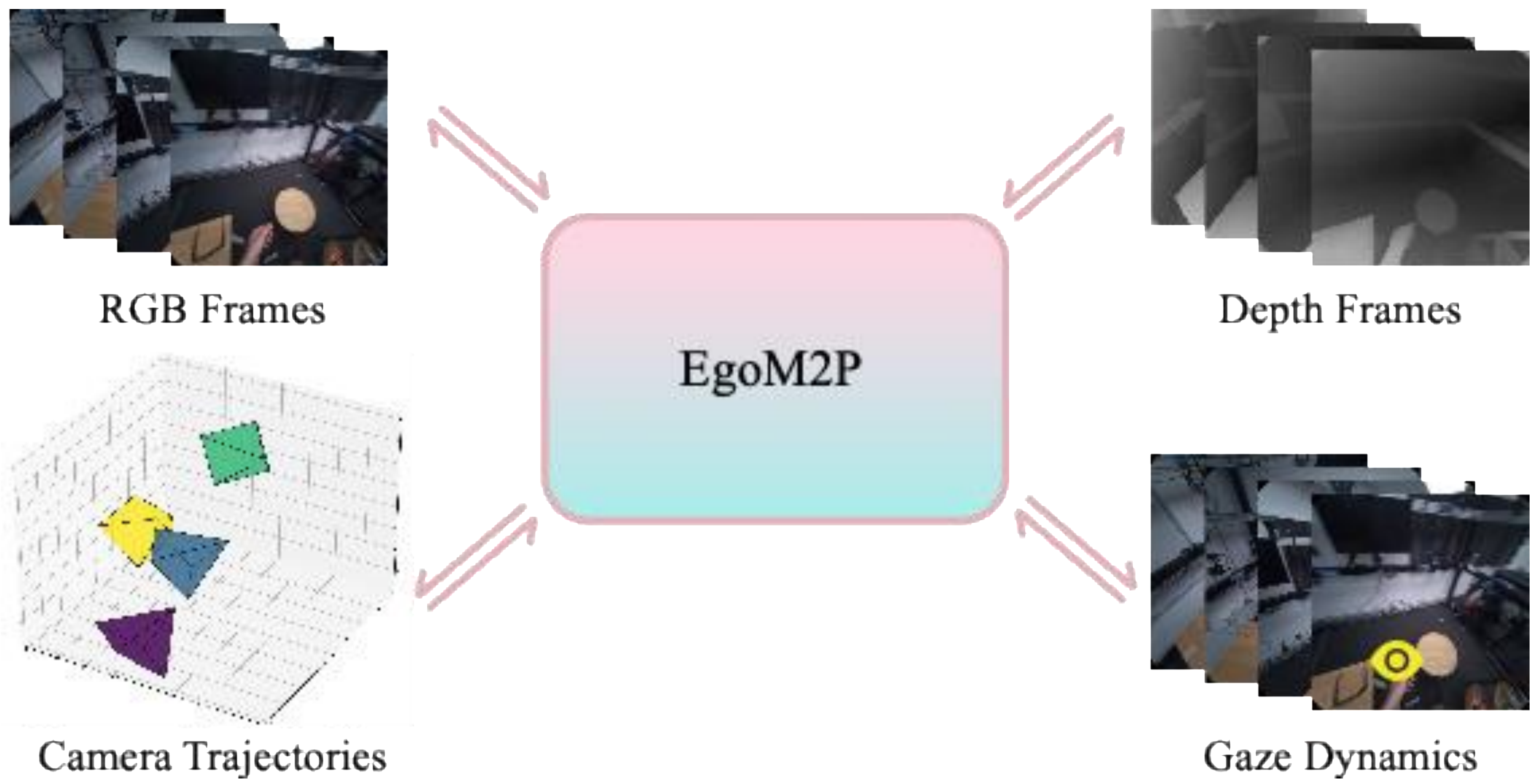


2. Synthesizing *virtual* humans

3. Embodied *digital* humans



4. Unified human foundation models that are 3D-grounded and multimodally capable.



EgoM2P: Egocentric Multimodal Multitask Pretraining

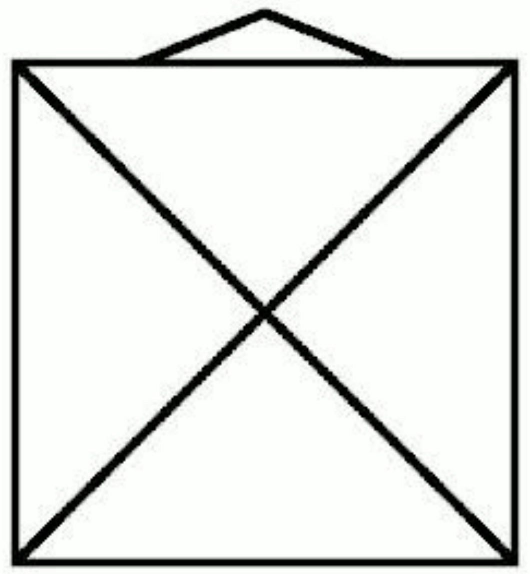
Gen Li, Yutong Chen*, Yiqian Wu*, Kaifeng Zhao*, Marc Pollefeys, Siyu Tang

CVPR 2025

ETH zürich VLG Microsoft Swiss AI

Human-centric multimodal modeling

- *Humans are inherently multimodal: both in how we express ourselves and how we perceive the world.*
    - Our expression: body pose, facial expression, speech (audio), gaze
    - Our perception: egocentric vision, hearing, touch, proprioception
- Egocentric captures contain rich multimodal data
    - RGB, depth, gaze, camera trajectory, …

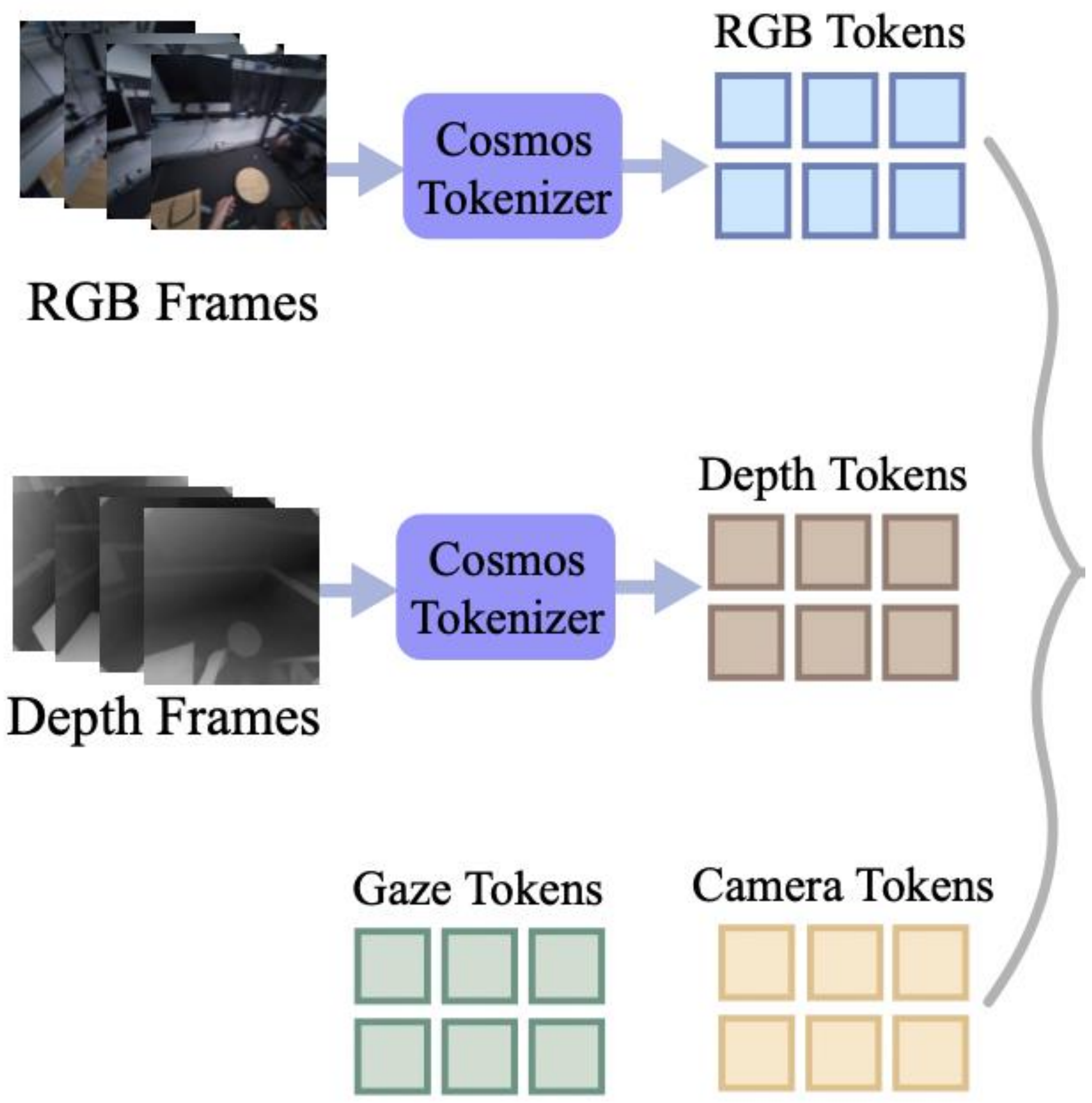


Challenges

- Besides the required know-how and substantial engineering efforts
- Heterogeneous modality annotations
  - Lack effective pseudo labelers
- Temporal consistency compared to multitask image foundation models
  - Fast-changing camera poses

Datasets \ Modalities	RGB	Depth	Gaze	Camera
EgoExo4D [29]	✓	✗	✓	✓
HoloAssist [103]	✓	✓*	✓	✓
HOT3D (Aria) [10]	✓	✓*	✓	✓
HOT3D (Quest) [10]	gray	✓*	✗	✓
ARCTIC [23]	✓	✓*	✗	✓
TACO [59]	✓	✓*	✗	✓
H2O [48]	✓	✓	✗	✓
EgoGen [51]	✓	✓	✗	✓

EgoM2P



400 billion multimodal training tokens

EgoM2P Training

EgoM2P Training

256 H100 GPUs, 16 hours

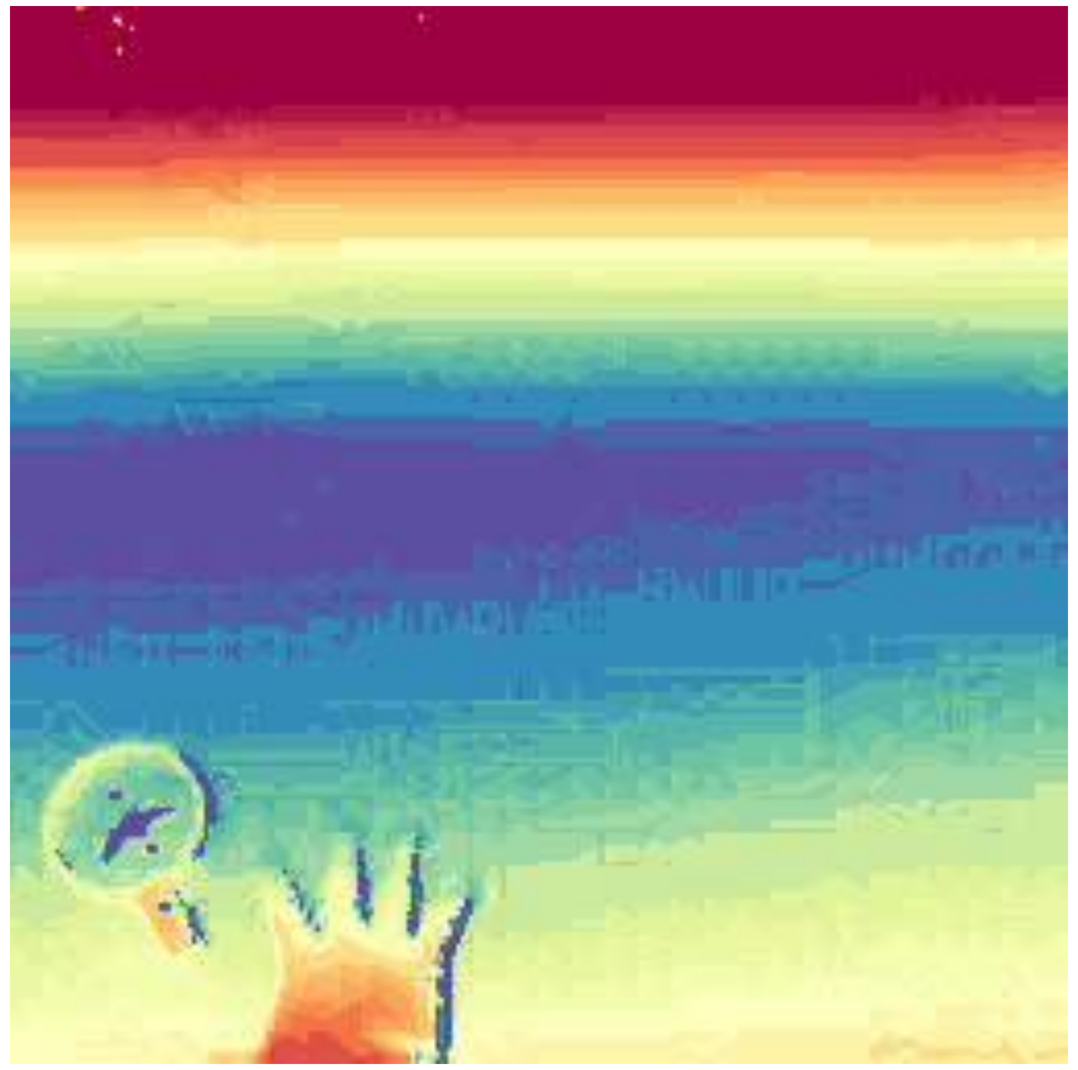
Egocentric gaze estimation

Ground Truth
Huang et al. 2018
Lai et al. 2022
Ours

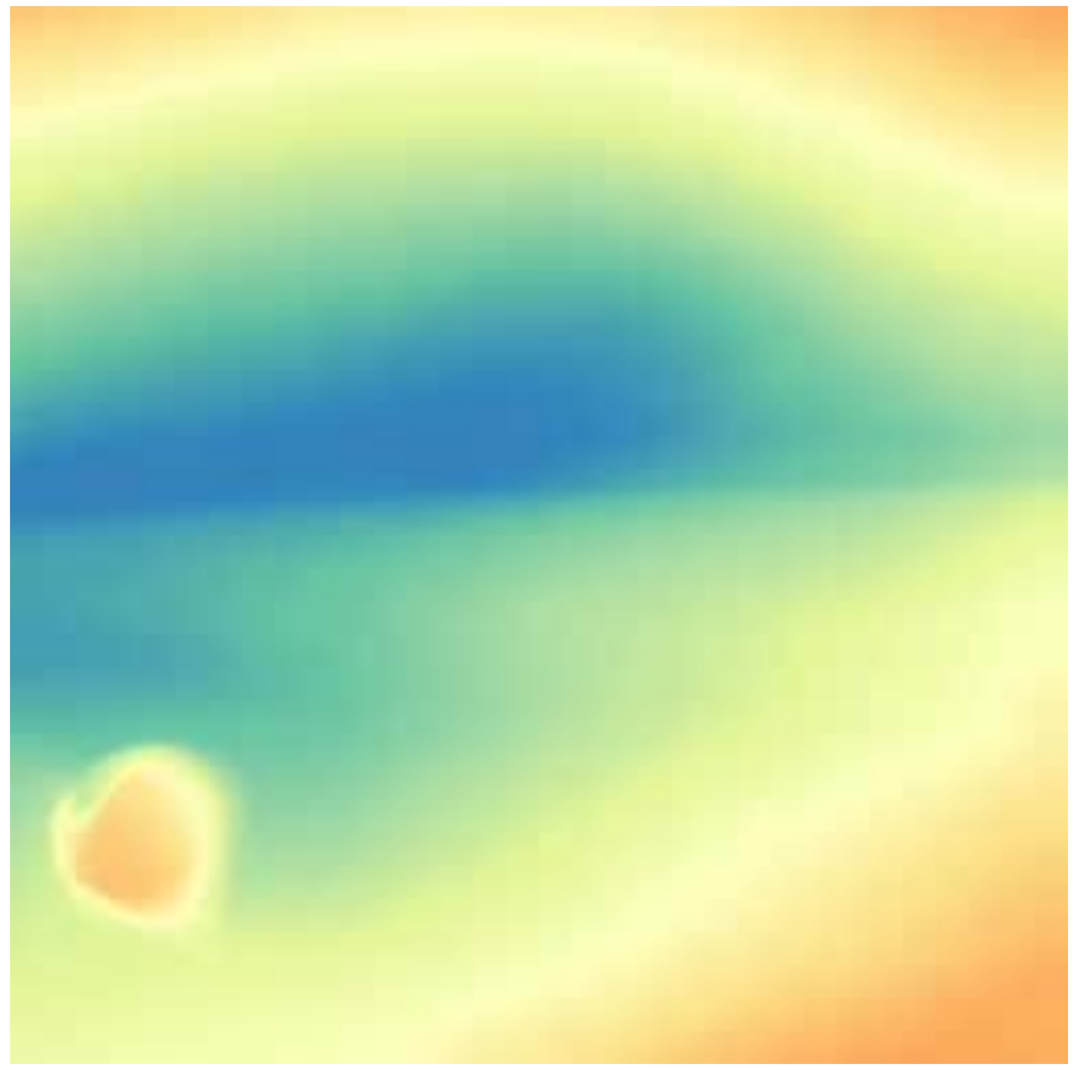
Egocentric depth estimation

- HOI4D (unseen dataset)

Input RGB



GT Depth



RollingDepth

Ours

Align3r

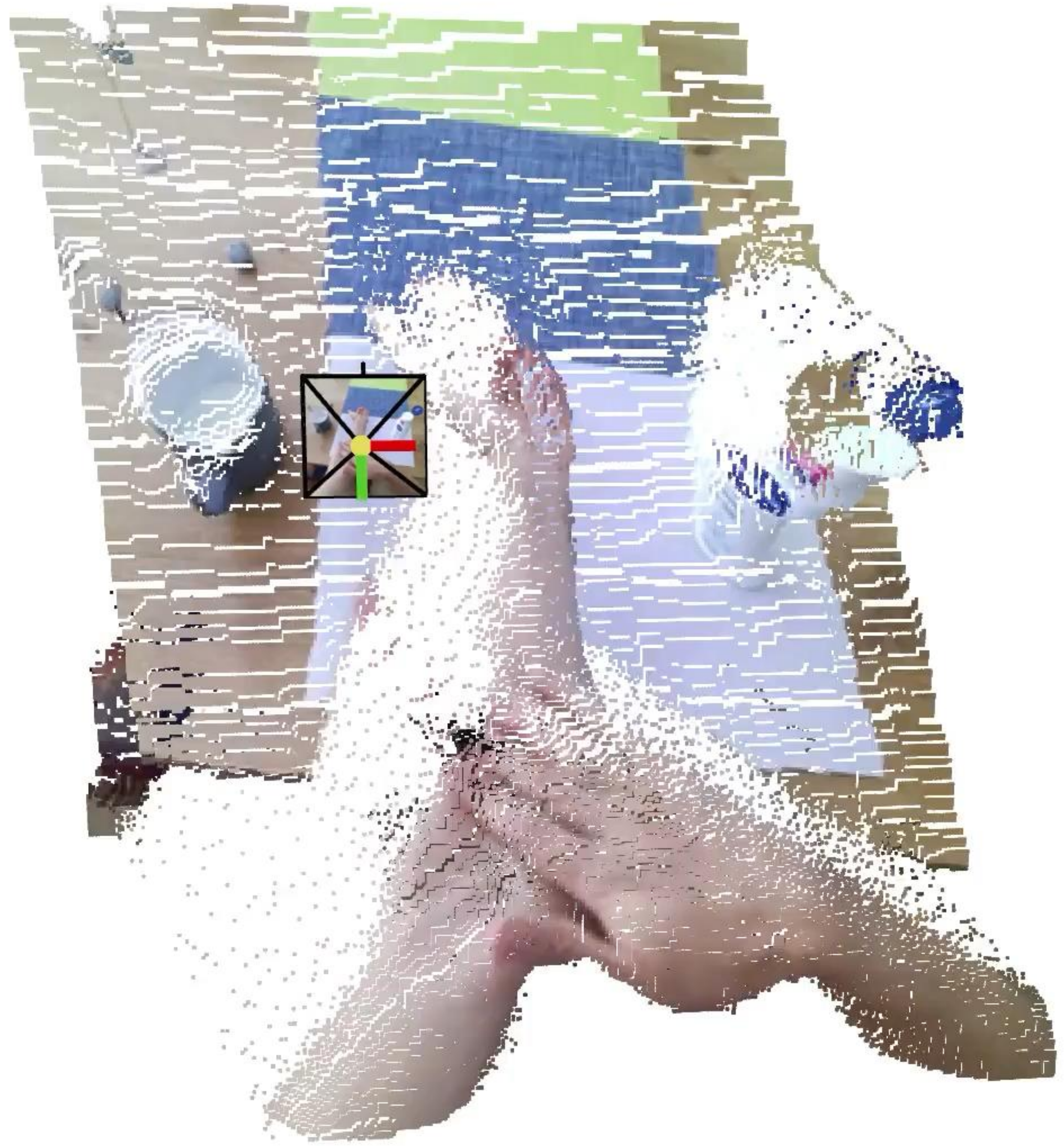
Egocentric depth estimation

Method	H2O [49] Abs Rel ↓	H2O [49] $\delta_{1.25}$ ↑	HOI4D [59] (*unseen*) Abs Rel ↓	HOI4D [59] (*unseen*) $\delta_{1.25}$ ↑	Time ↓
RollingDepth [44]	0.087	90.5	0.057	97.6	37s
Align3R [65]	0.074	91.8	**0.045**	**98.1**	90s
EgoM2P	**0.055**	**96.0**	0.061 **<u>0.041</u>**	98.0 **<u>99.0</u>**	**0.8s**

Table 3. **Evaluation on egocentric video depth estimation.** Compared to specialist SOTAs requiring geometry-based test-time optimization, the versatile *EgoM2P* achieves comparable performance while being significantly more efficient. With post-training described in Sec. 4.5, *EgoM2P* excels (see <u>underlined results</u>).

Egocentric 4D reconstruction

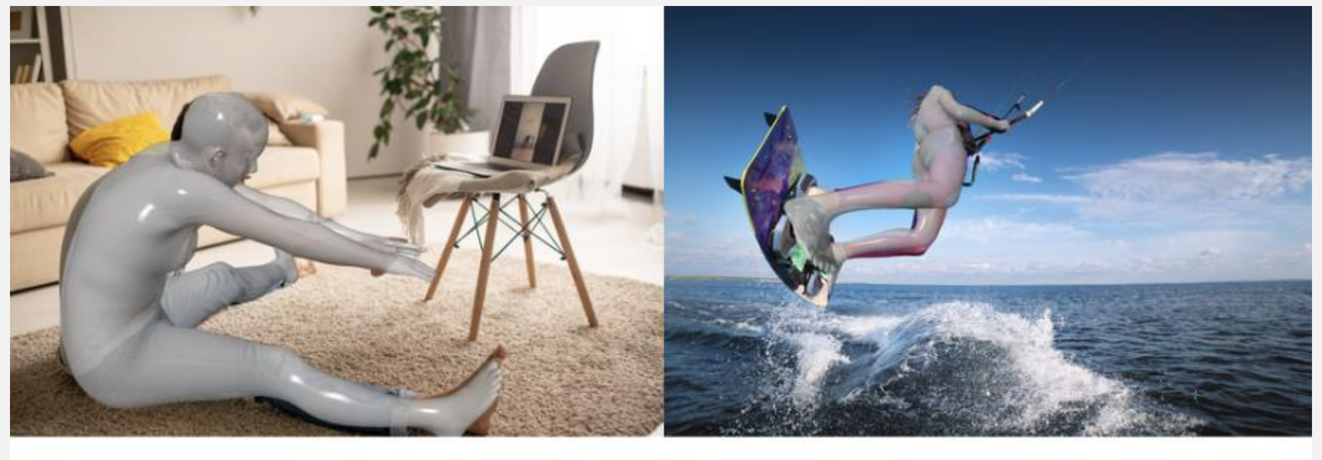
MegaSaM (71s)



Ours (<1s)

Summary

1. Perceiving and reconstructing *real* humans



2. Synthesizing *virtual* humans

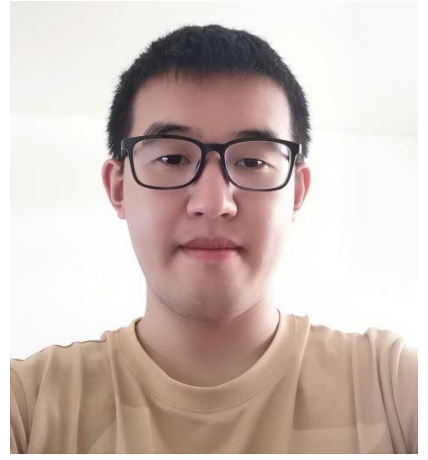
3. Embodied *digital* humans

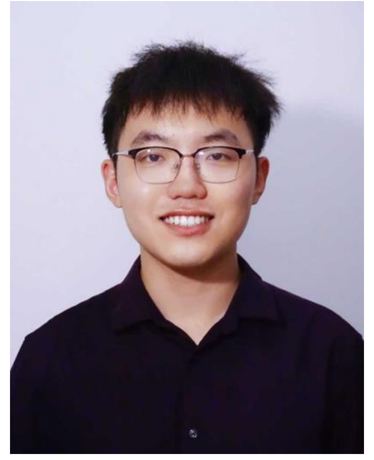
4. Unified human foundation models that are 3D-grounded and multimodally capable.

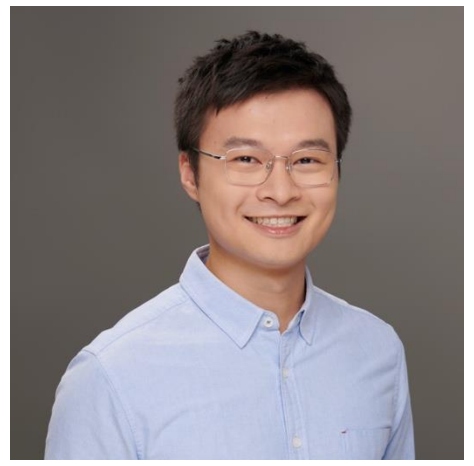
The source code and models of all the work are available: https://vlg.inf.ethz.ch/publications/

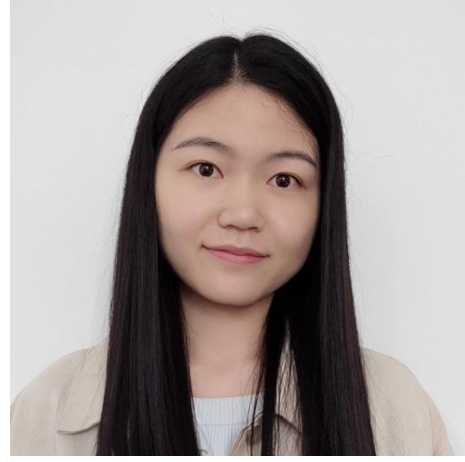
Thank you



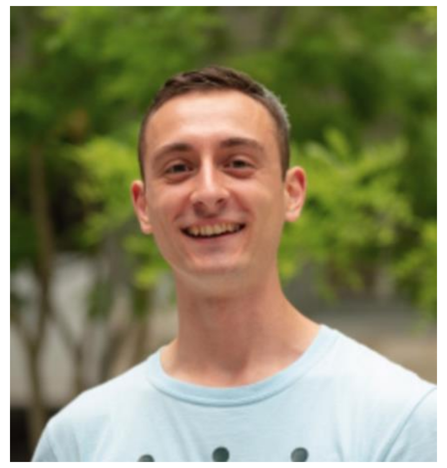






erc Starting Grants

ETH zürich Research Grant

ETHAR

Innosuisse

swiss AI

FIFA

Disney RESEARCH STUDIOS

Swiss National Science Foundation

Future Computing Laboratory

Google

Meta

Qualcomm

Microsoft