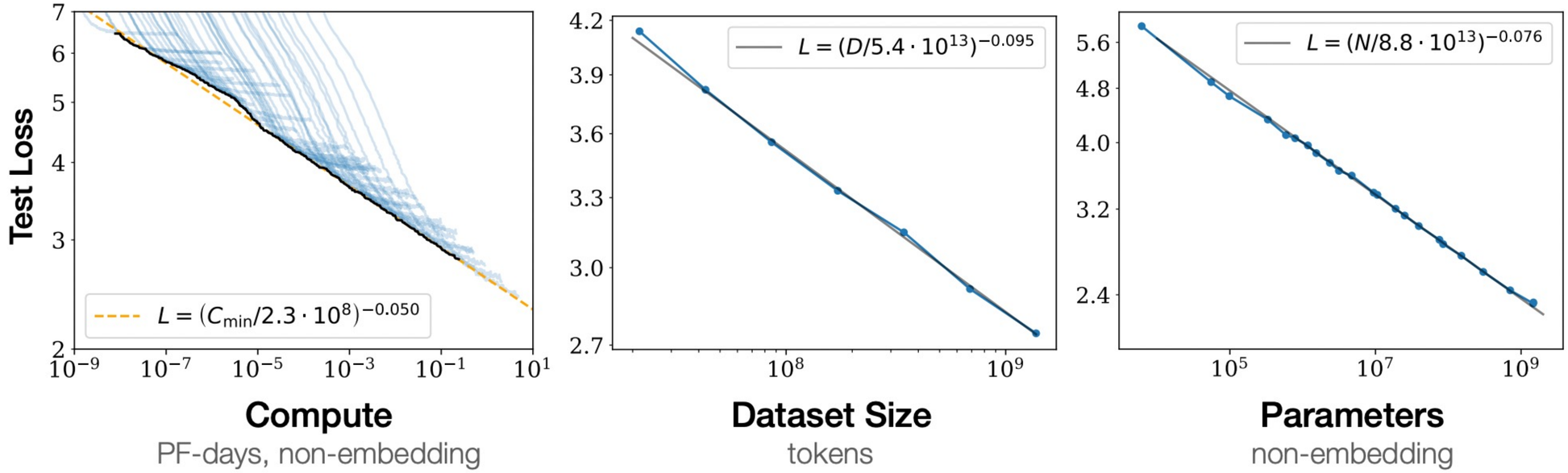
Toward Egocentric Multimodal Multitask Pretraining

Gen Li
ETH Zurich

Scaling and Foundation Models



- Language models, third-person-view video models, 3D reconstruction models…
- Key: large-scale high-quality training data

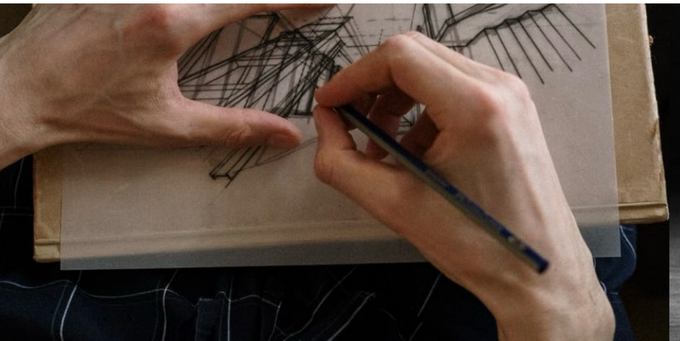


Humans are inherently multimodal:

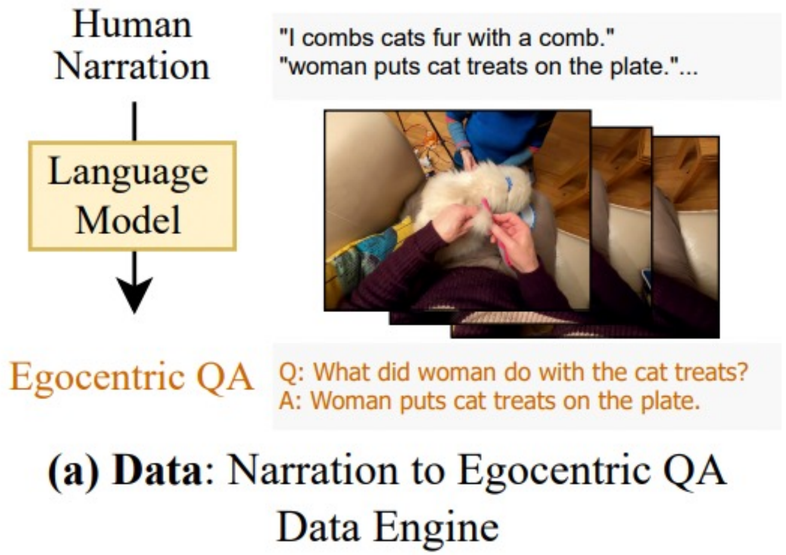
both in how we express ourselves and how we perceive the world.

Our expression: body pose, facial expression, speech (audio), gaze

Our perception: egocentric vision, hearing, touch, proprioception



Egocentric multimodal models



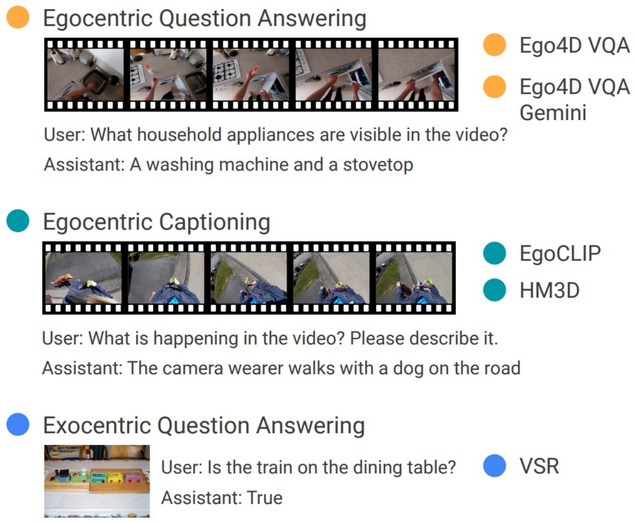


Figure 1. **The Overview of EgoLife Project.** The EgoLife project features six participants living together for a week to prepare an Earth Day celebration. Each participant wears Meta Aria glasses [1], recording approximately 8 hours of egocentric video and signals daily. In addition, 15 cameras and 2 mmWave devices provide synchronized third-person perspective data (detailed in Figure 2). These comprehensive annotations enable the development of state-of-the-art multimodal egocentric AI assistants and introduce novel tasks to advance long-term

MM-Ego [Ye et al. 2024]

AlanaVLM [Suglia et al. 2024]

EgoLife [Yang et al. 2025]

Limited exploration of the full spectrum of human-centric multimodal information

Overview

1. Egocentric multimodal data scaling up:

EgoGen: An Egocentric Synthetic Data Generator. CVPR 2024

2. Toward large-scale egocentric pre-training:

EgoM2P: Egocentric Multimodal Multitask Pretraining. ICCV 2025

EgoGen:

An Egocentric Synthetic Data Generator

CVPR 2024 oral

Gen Li Kaifeng Zhao Siwei Zhang Xiaozhong Lyu Mihai Dusmanu Yan Zhang Marc Pollefeys Siyu Tang

ETH zürich VLG Microsoft

EgoGen: An Egocentric Synthetic Data Generator

Human-centric Synthetic Data

BEDLAM [Black et al. CVPR 2023]

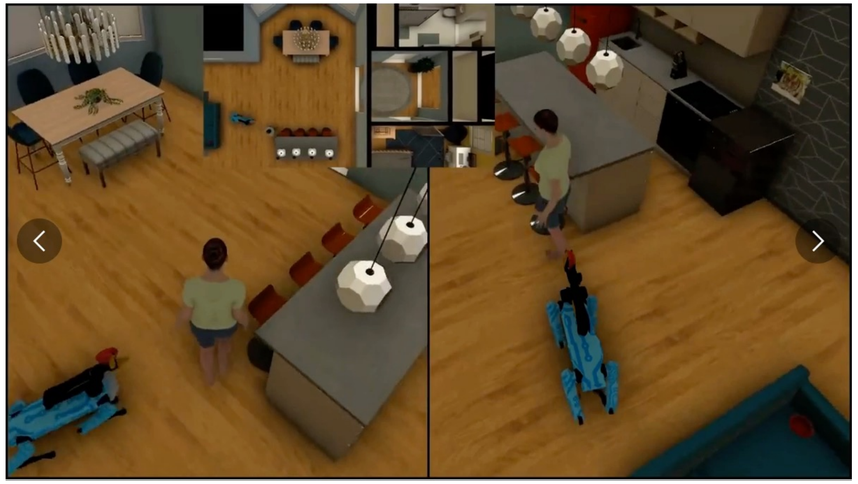
SynBody [Yang et al. ICCV 2023]

UnrealEgo [Akada et al. ECCV 2022]

Random motion sampling from MoCap datasets

Lack of human-scene interactions

Virtual Humans in Simulations



Habitat 3.0 [Puig et al. ICLR 2024]

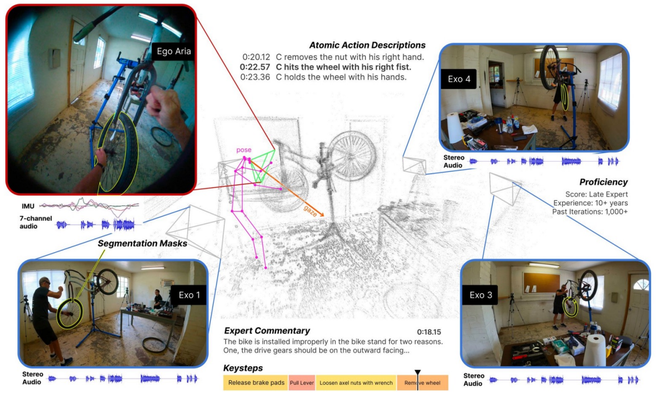
uHumans2 [Rosinol et al. IJRR 2021]

Carla [Dosovitskiy et al. CoRL 2017]

Deterministic, unnatural human motions

Lack of realistic and diverse human appearances

Egocentric HMD Datasets



Ego-Exo4D [Grauman et al. CVPR 2024] **HoloAssist [Wang et al. ICCV 2023]** **Aria Digital Twin [Pan et al. ICCV 2023]**

Expensive to create rich and accurate ground truth annotations

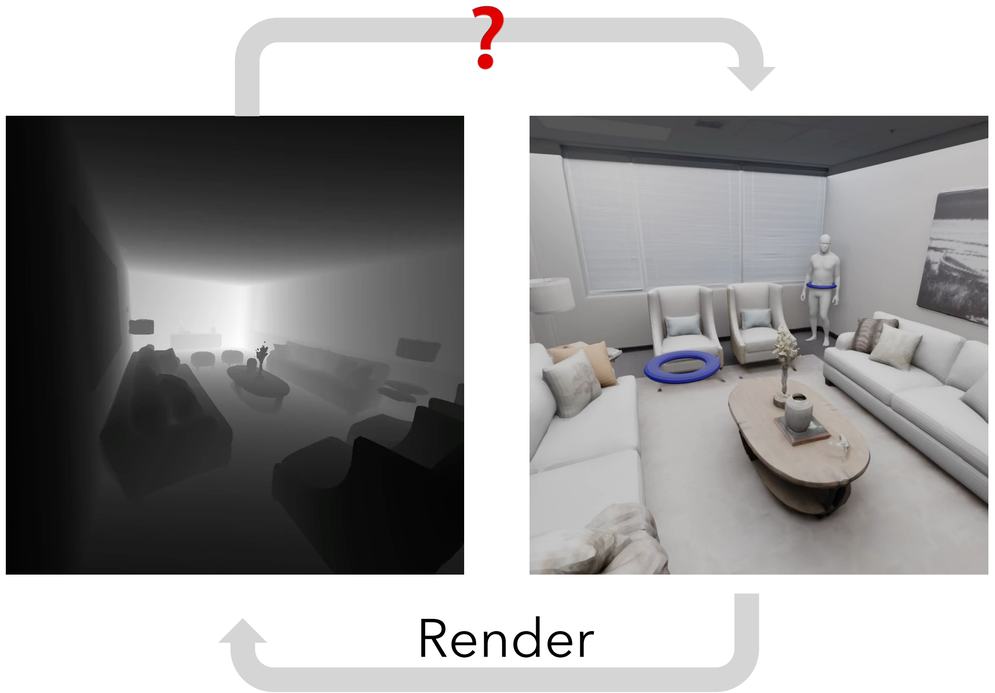
Privacy concerns

Challenges

- Motion diversity & realism
  - ➡️ Generative motion model
- Appearance diversity
  - ➡️ Diverse body shape, texture, …
  - ➡️ Automated clothing simulation

Challenges

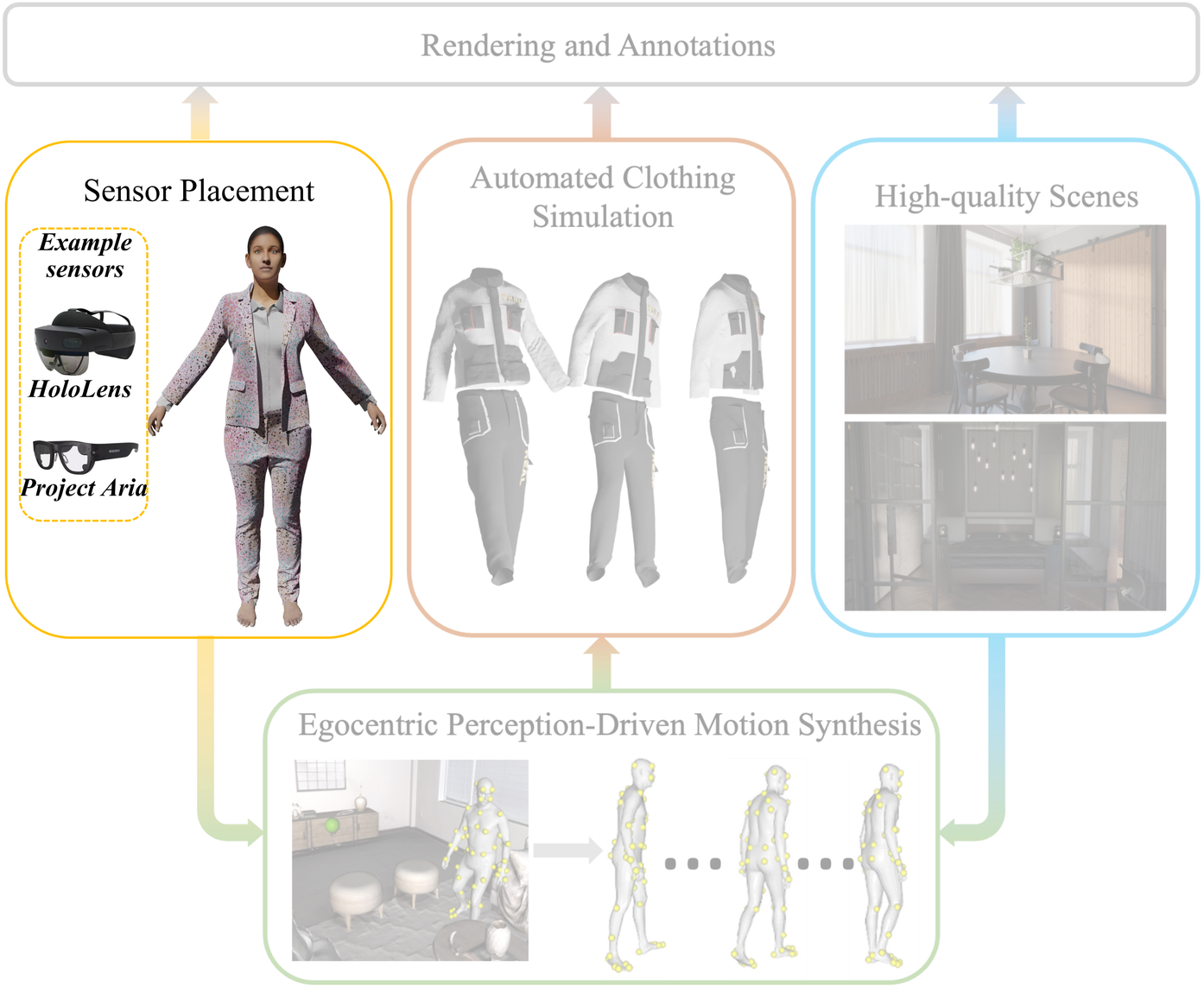
- Motion diversity & realism
  - ➡️ Generative motion model
- Interdependence



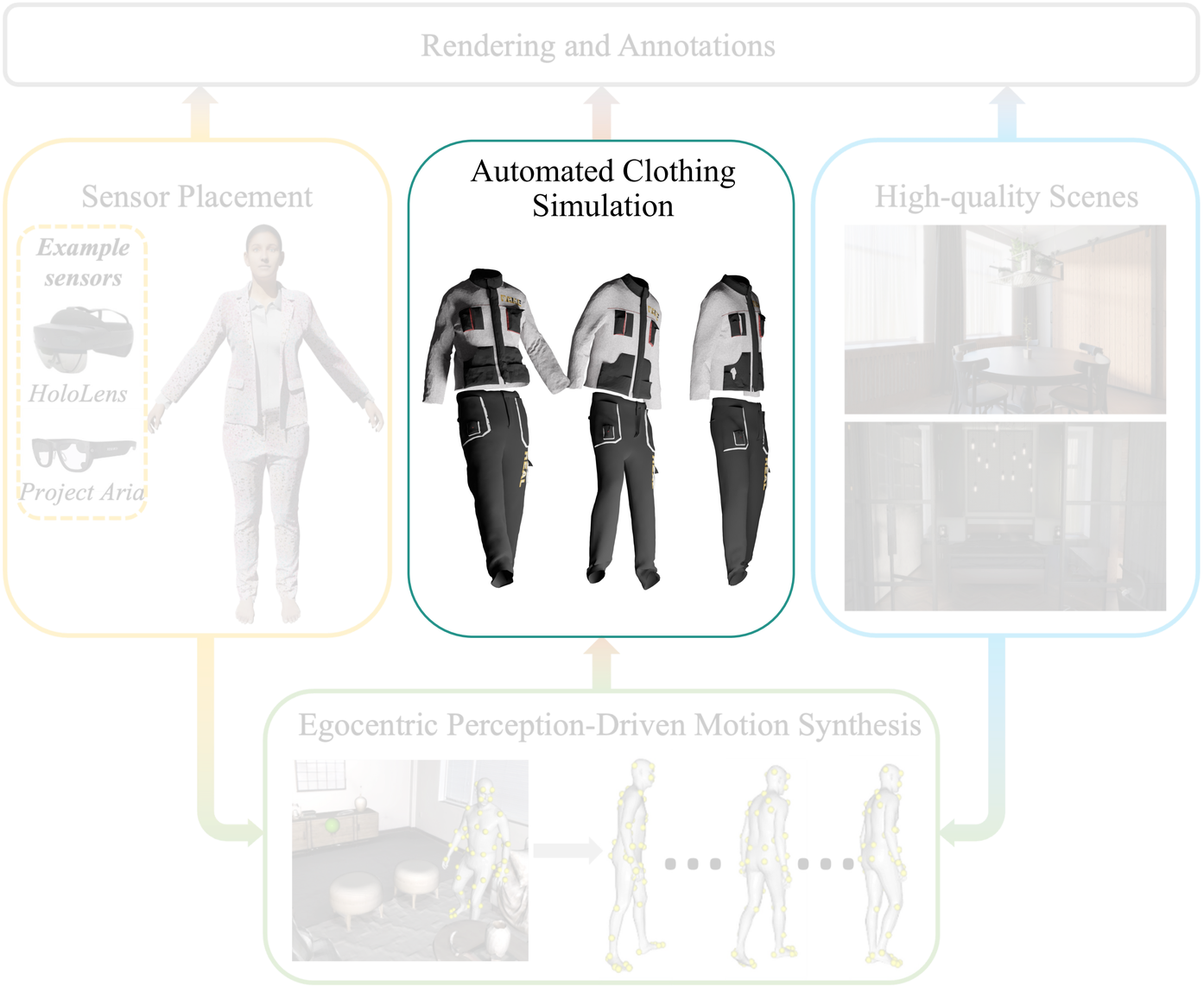
- Appearance diversity
  - ➡️ Diverse body shape, texture, …
  - ➡️ Automated clothing simulation

➡️ Novel embodied human motion synthesis

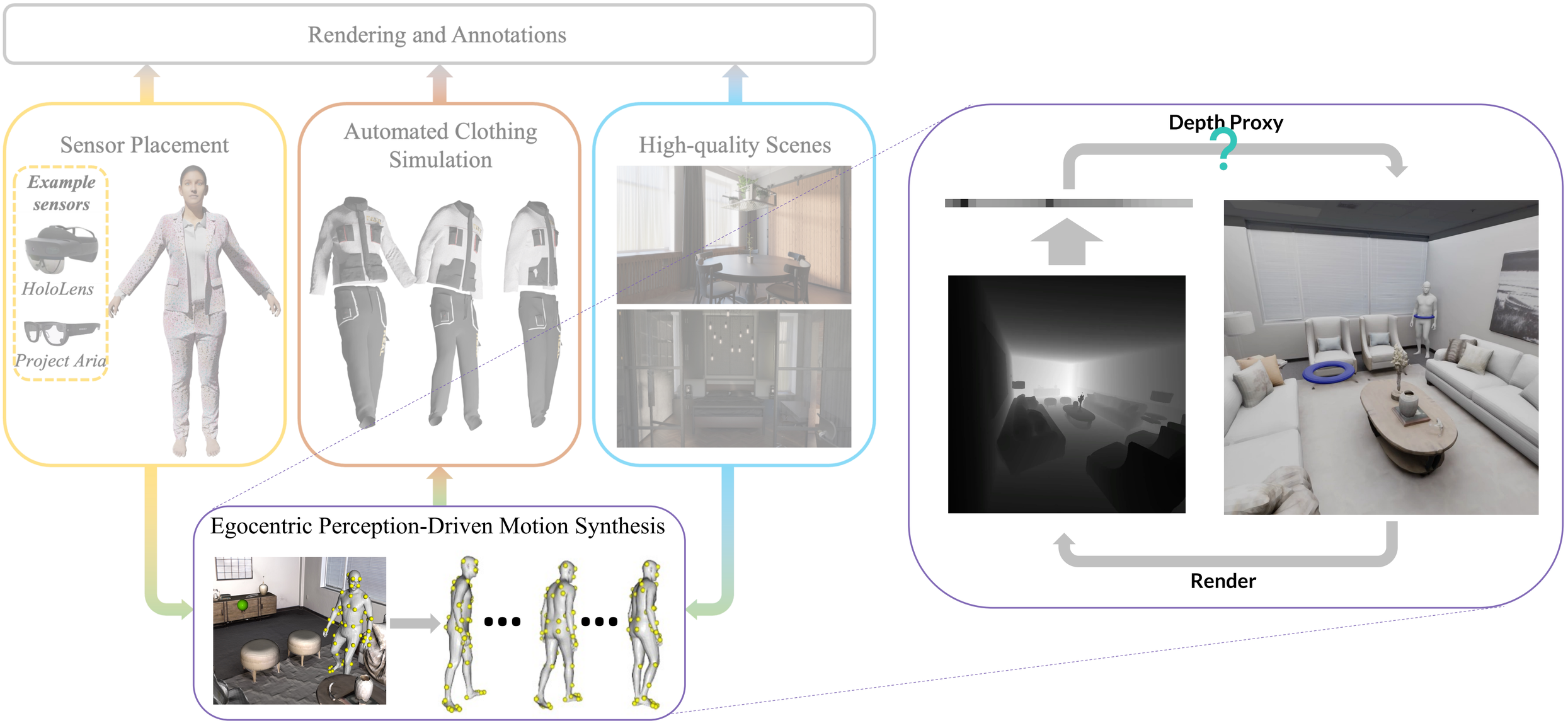
Overview of EgoGen



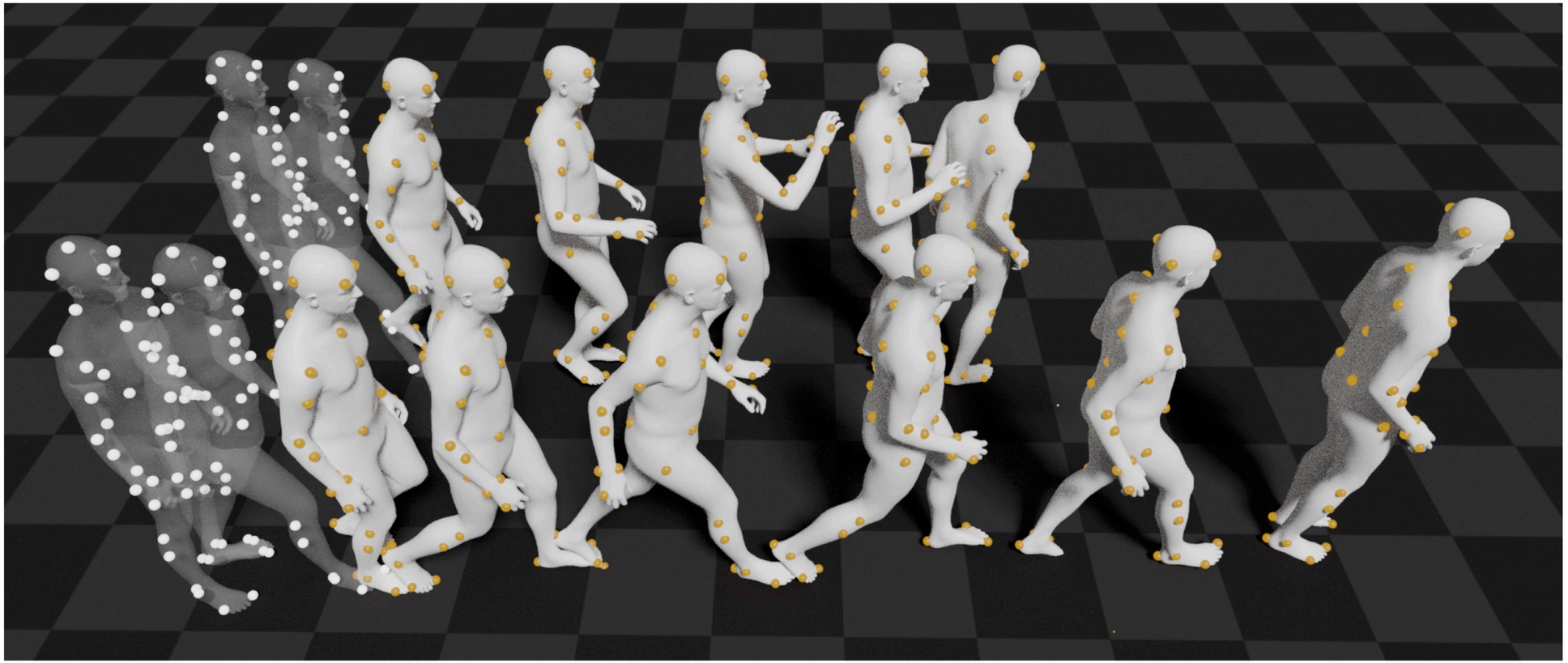
Overview of EgoGen



Overview of EgoGen



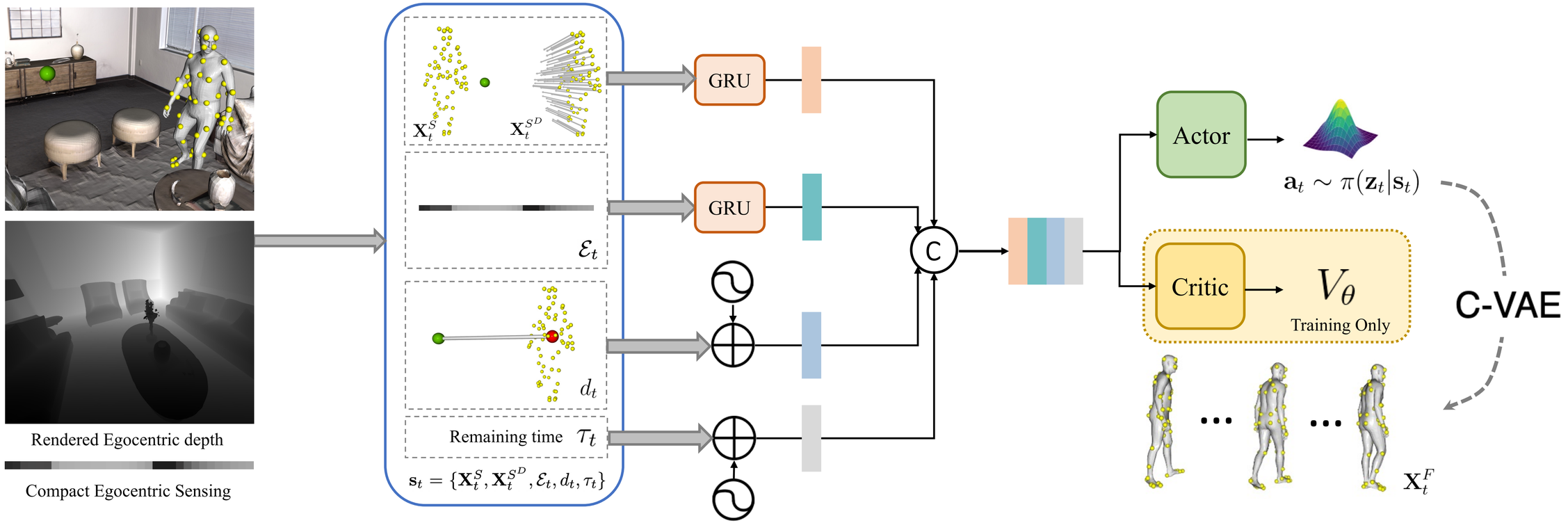
Ego-Sensing Driven Motion Model



World model for motion primitives learning

- C-VAE: history motion + Z => future motion
- Z: Natural motion manifold as our action space

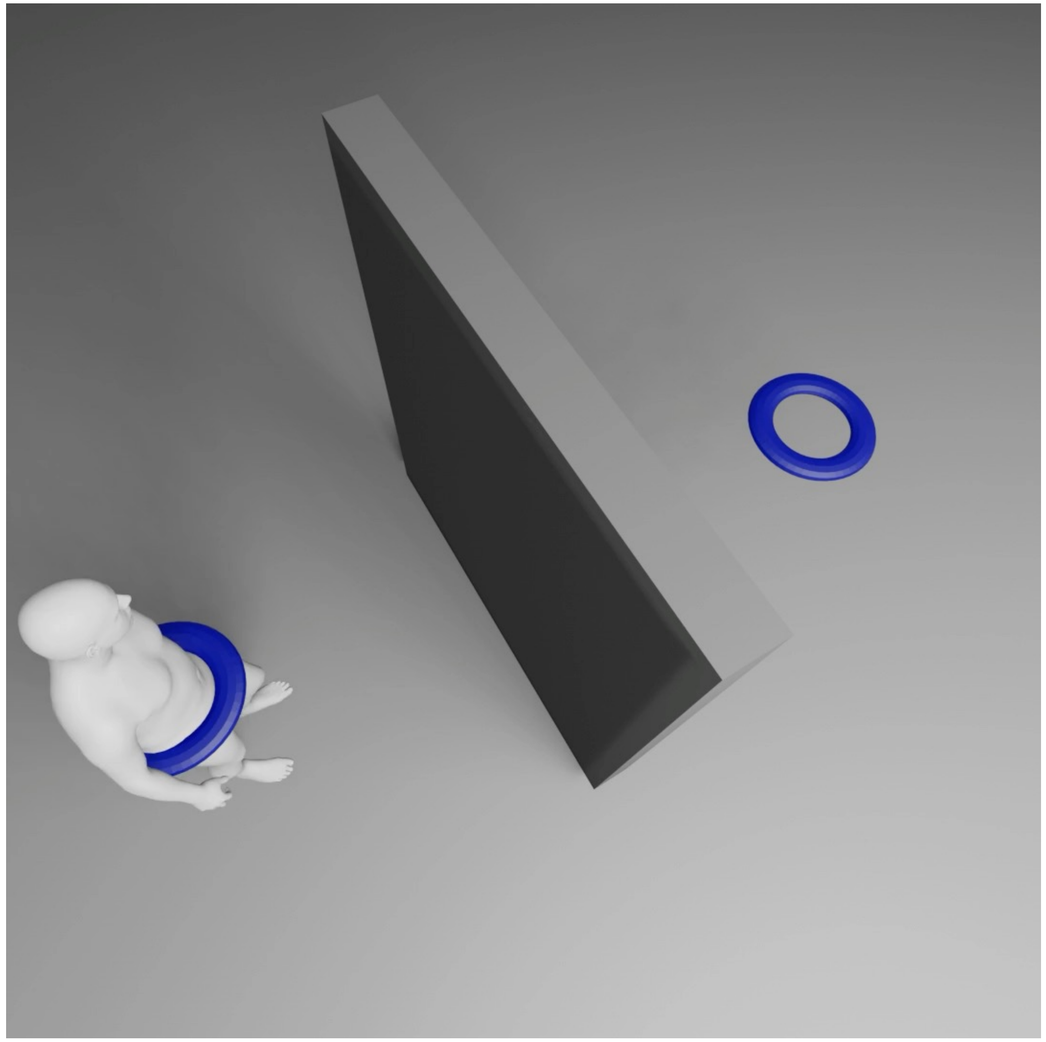
Ego-Sensing Driven Motion Model



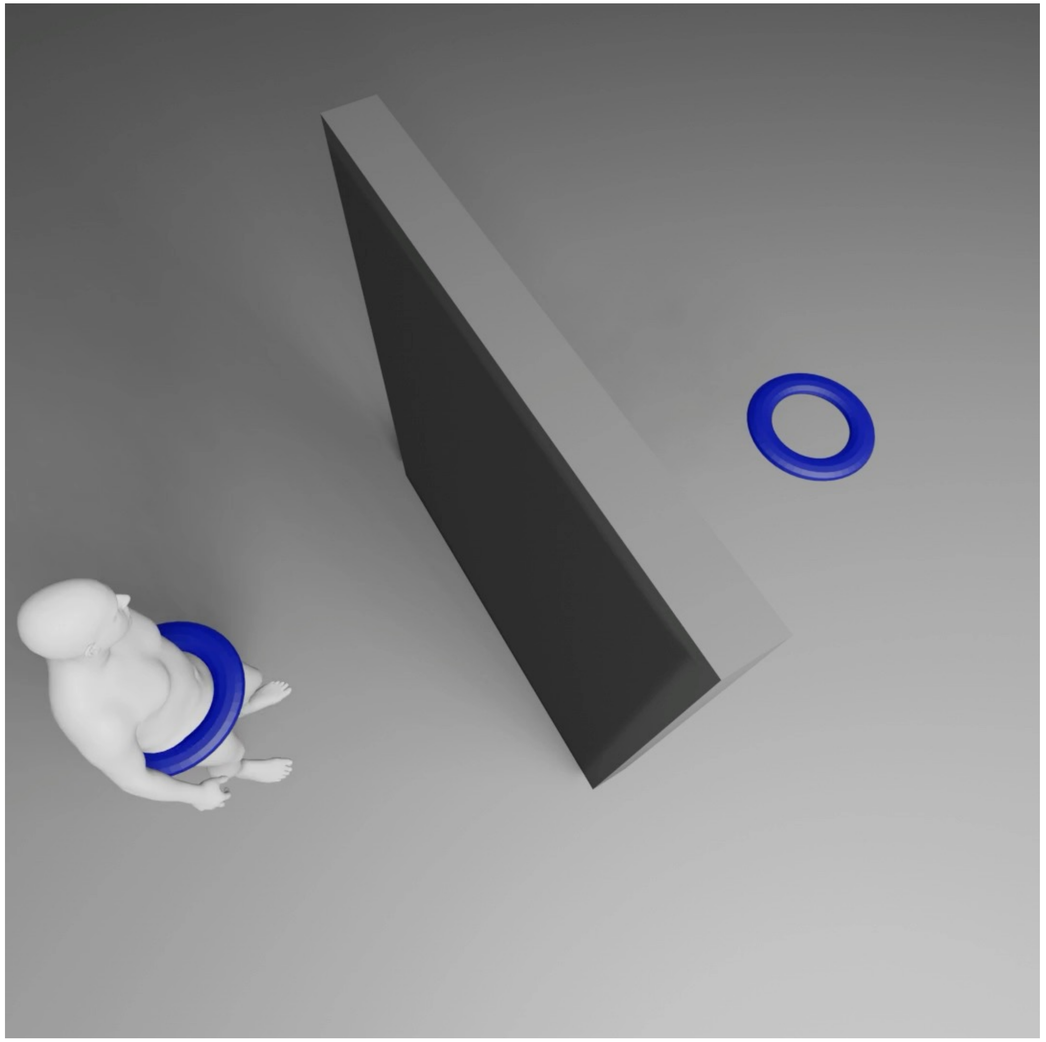
MDP formulation

- State space: our marker-based body representation, embodied perception, etc.
- Action space: latent action space of C-VAE
- Transition Dynamics P(s' | s, a): C-VAE decoder
- Reward: foot-floor contact, distance, penetration, attention, etc.

Ego-Sensing helps exploration



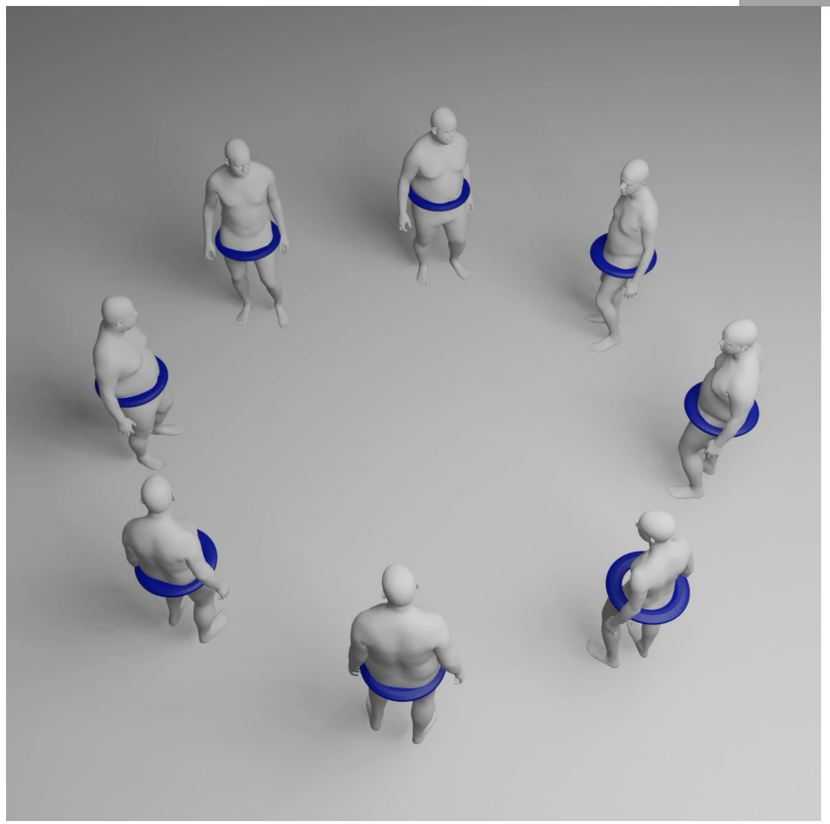
[DIMOS Zhao et al. 2023]



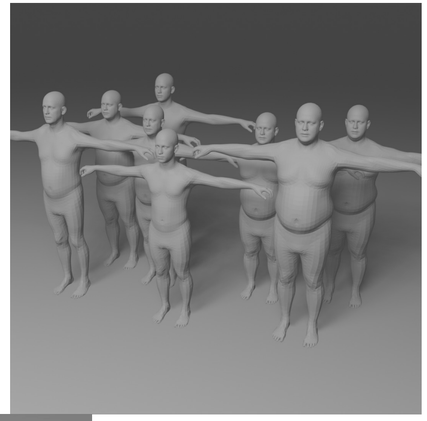
Egocentric Sensing

Comparison: Diversity

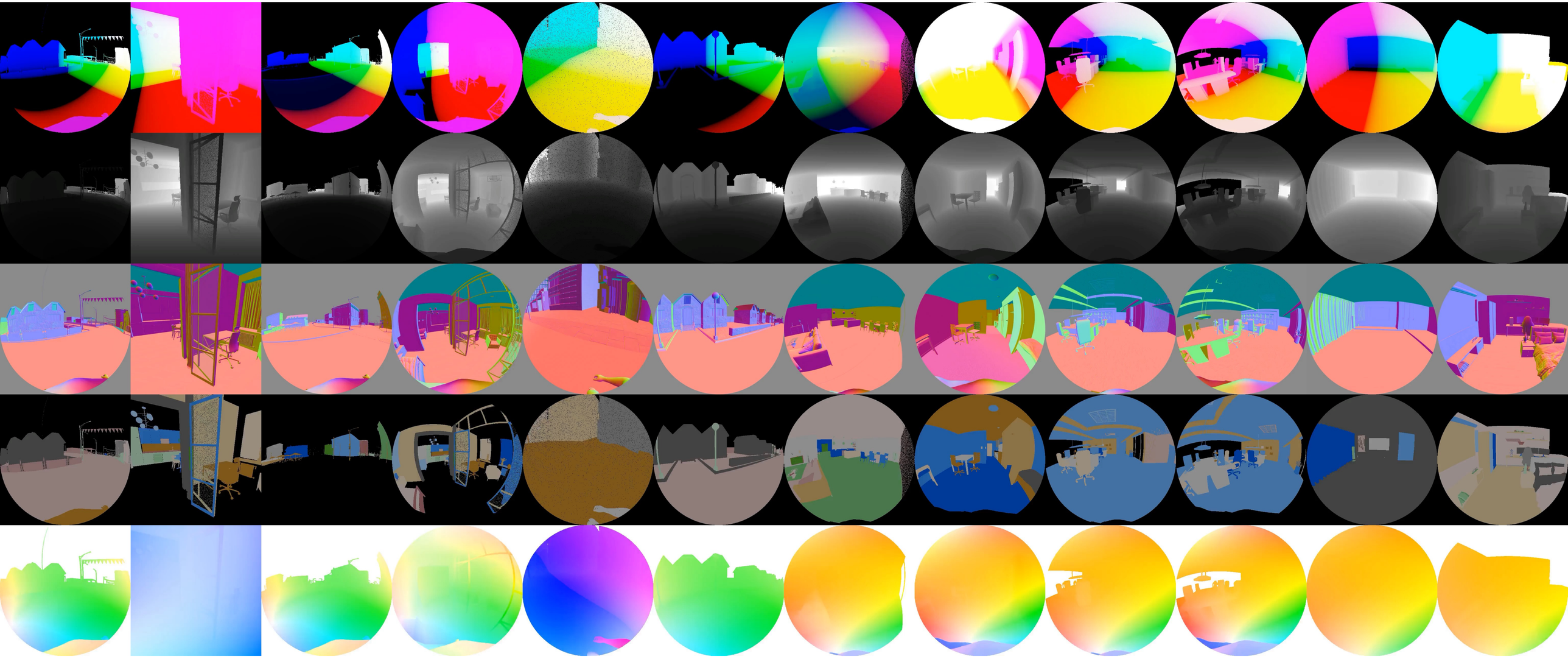
PhysicsVAE [Won et al. SIGGRAPH 2022]



Ours: crowd with random body shapes



EgoGen: An Egocentric Synthetic Data Generator

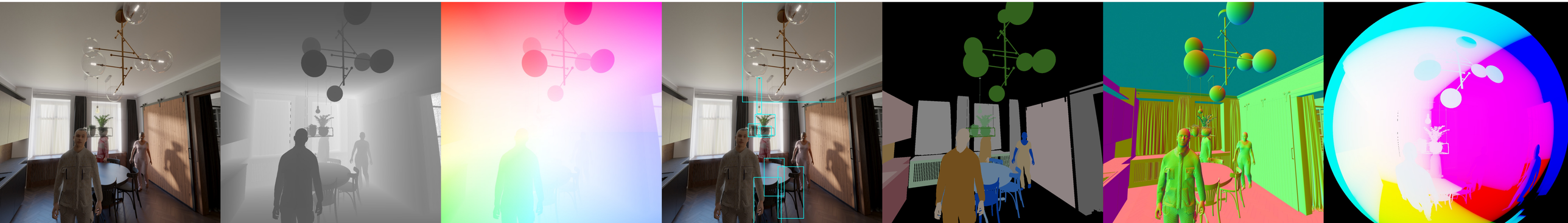


Applications

Egocentric perception tasks with EgoGen

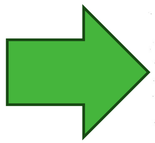
1, Mapping and localization for AR

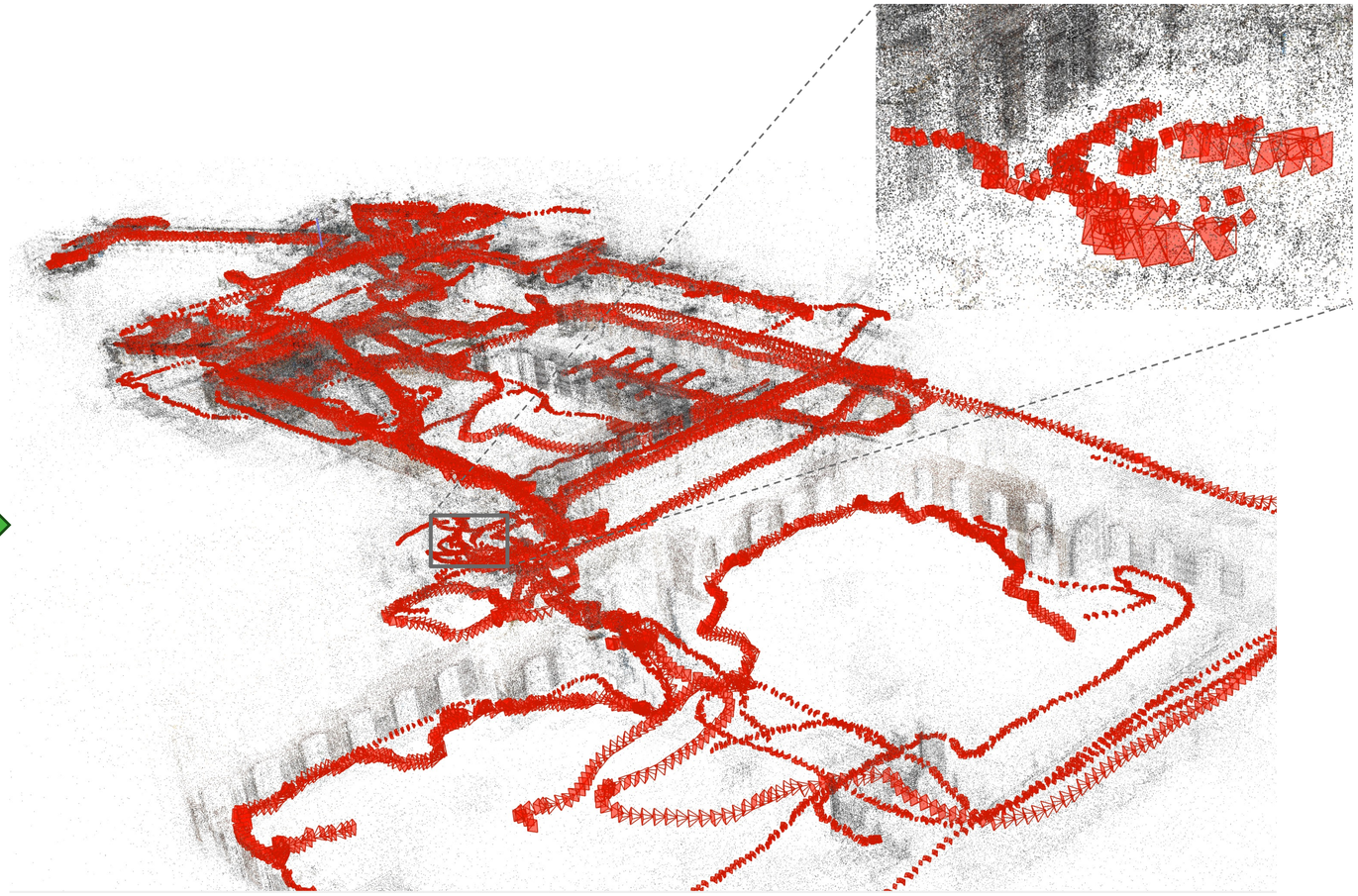
2, Human mesh recovery from egocentric views



Mapping & Localization for AR

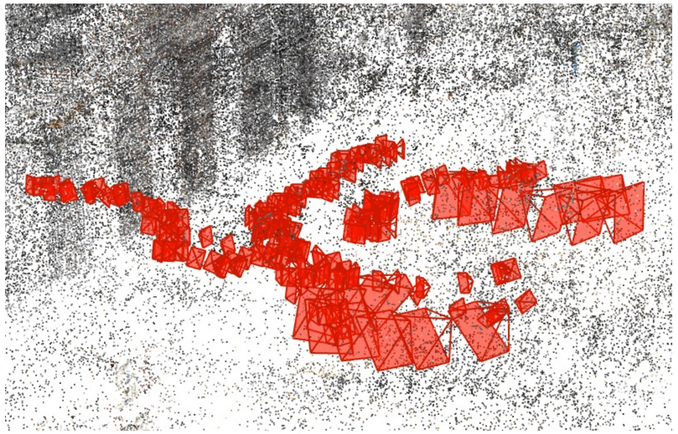
Input Images





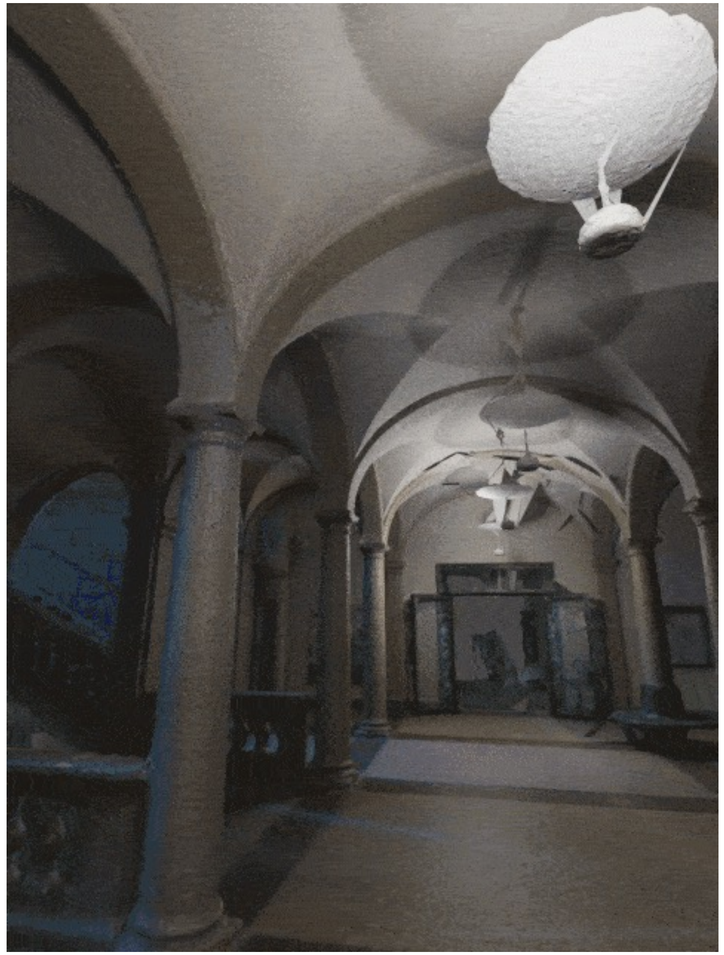
Output SfM Map

Mapping & Localization for AR

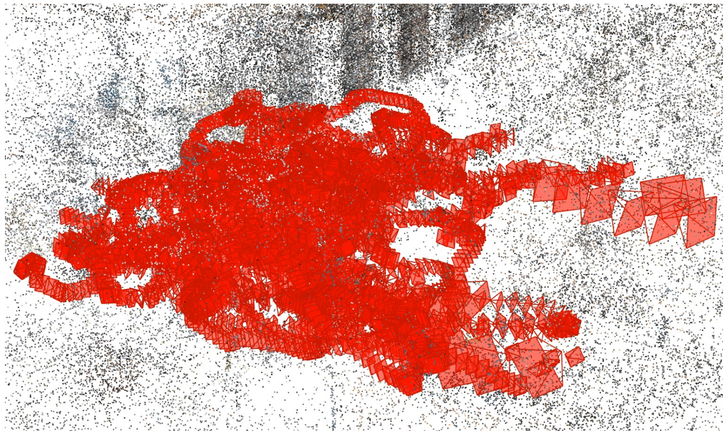


Initial SfM Map

Third-person view



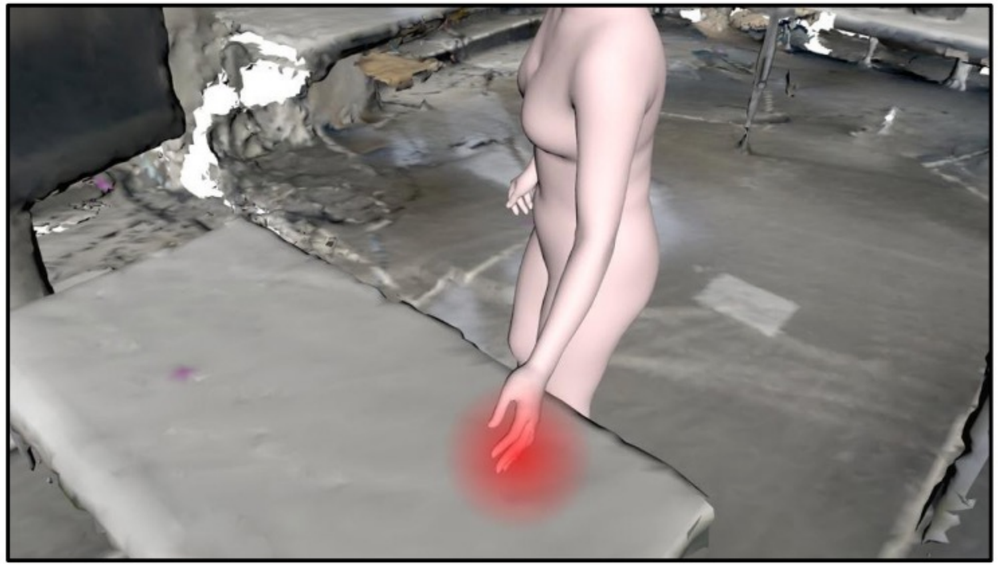
EgoGen synthetic data



More complete SfM Map

Egocentric Human Mesh Recovery

Input egocentric image

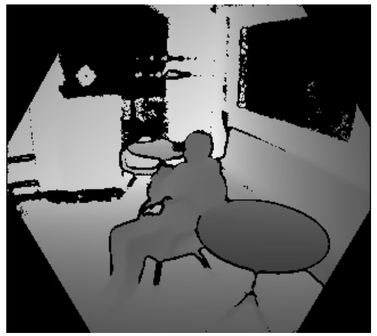
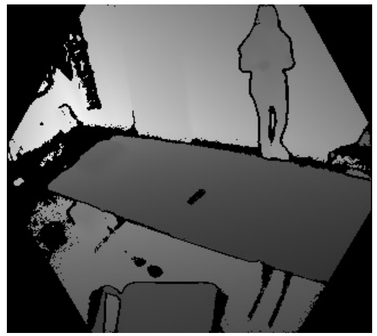


Output recovered human mesh

Egocentric Human Mesh Recovery

Our solution: generating synthetic training data. We leverage EgoGen to generate 300k RGB and 105k depth training images of humans moving in 3D scenes.

Egocentric Human Mesh Recovery


Ground truth

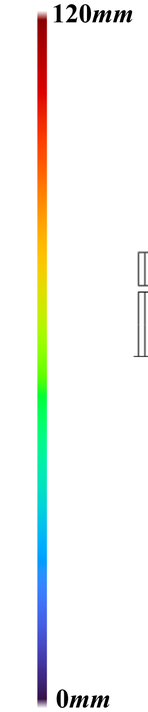


ProHMR Train from scratch



ProHMR Pretrained with our synthetic data





	G-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	V2V ↓
Depth-scratch	117.7	82.2	54.1	100.6
Depth-ft	**90.7**	**65.2**	**47.3**	**81.0**

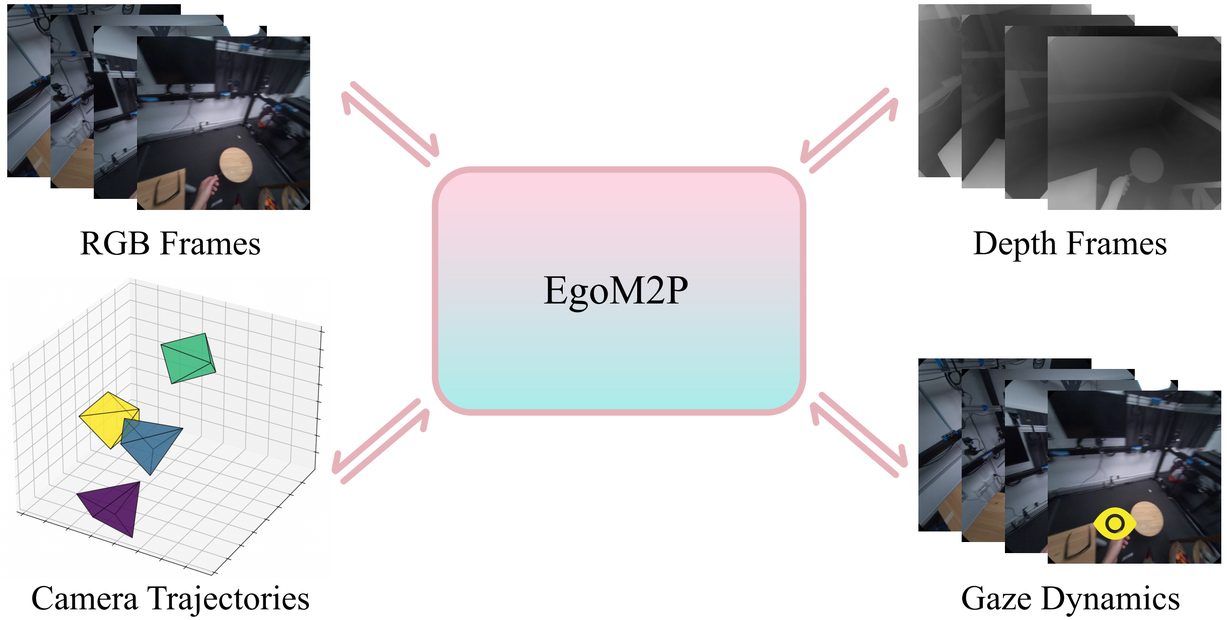
~20% improvement over accuracy

Take home messages

Egocentric synthetic data generation is another interesting application of generative human motion synthesis.

Egocentric-perception-driven modeling is challenging, but it may be the key to scaling up human behavior synthesis.

We have shown very simple human motions here; there are many more to explore…



EgoM2P: Egocentric Multimodal Multitask Pretraining

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

ICCV 2025

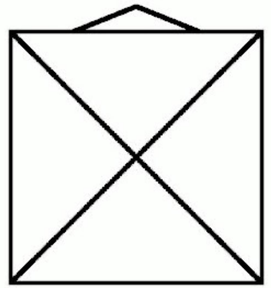
* Equal contribution, alphabetic order

ETH zürich VLG Microsoft

Motivation

- Egocentric captures contain rich multimodal data
  - RGB, depth, gaze, camera trajectory, …



Motivation

- Egocentric captures contain rich multimodal data
- Data amount is scaling up:
  - Real-world data: semantic rich and diverse. (HoloAssist, EgoExo4D, etc)
  - Synthetic data: precise GT annotation, cheap to scale. (EgoGen, Habitat)

- Feasible to train large multimodal/task egocentric vision models

Challenges

- Heterogeneous modality annotations
  - Lack effective pseudo labelers

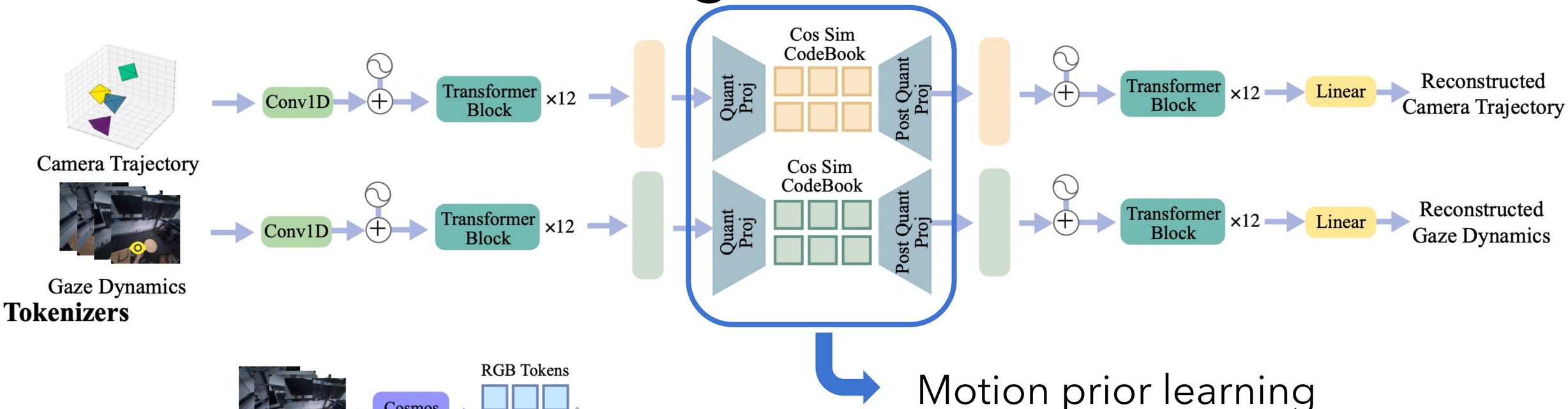
Datasets \ Modalities	RGB	Depth	Gaze	Camera
EgoExo4D [29]	✓	×	✓	✓
HoloAssist [103]	✓	✓*	✓	✓
HOT3D (Aria) [10]	✓	✓*	✓	✓
HOT3D (Quest) [10]	gray	✓*	×	✓
ARCTIC [23]	✓	✓*	×	✓
TACO [59]	✓	✓*	×	✓
H2O [48]	✓	✓	×	✓
EgoGen [51]	✓	✓	×	✓

Challenges

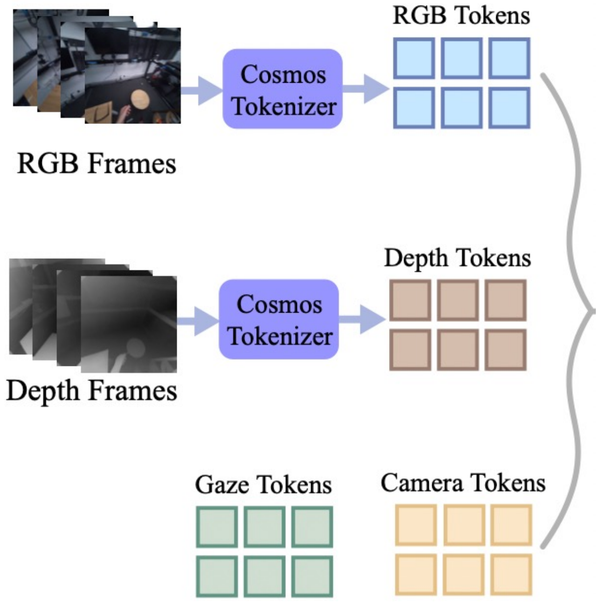
- Heterogeneous modality annotations
  - Lack effective pseudo labelers
- Temporal consistency compared to multitask image foundation models
  - Fast-changing camera poses
  - Spatiotemporal complexity
  - Token explosion

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

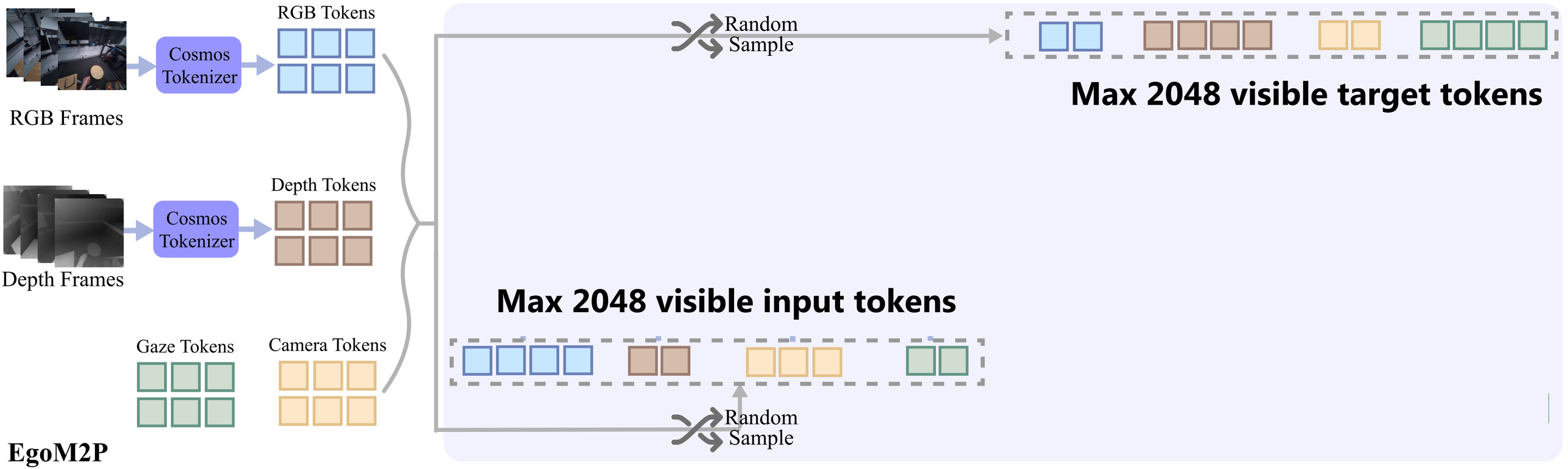


Motion prior learning

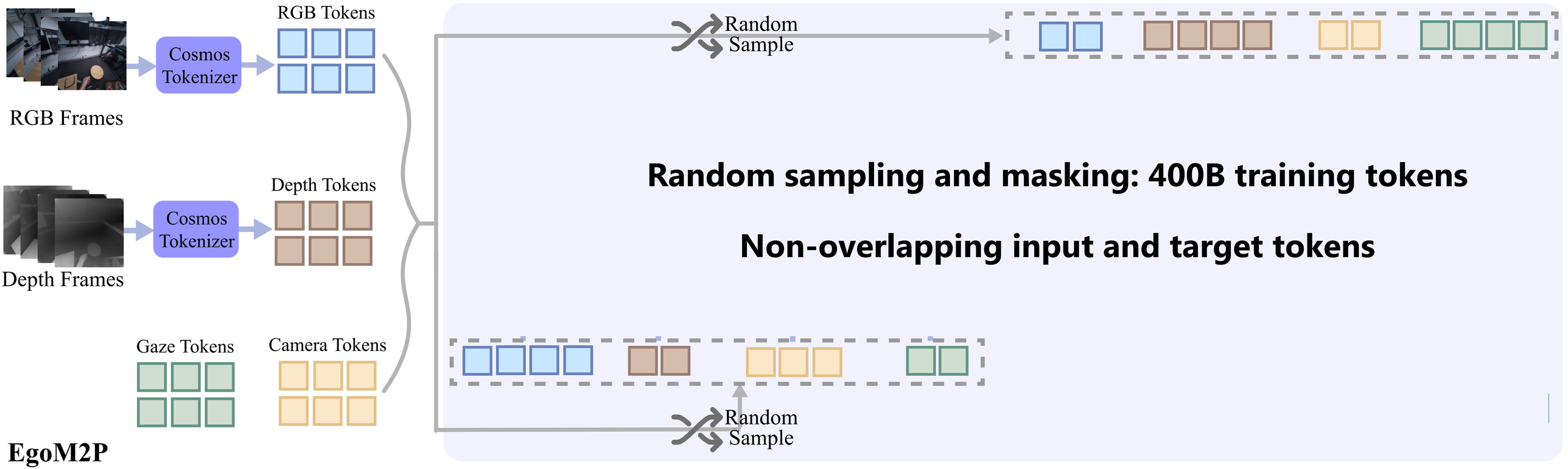


4 billion multimodal token database

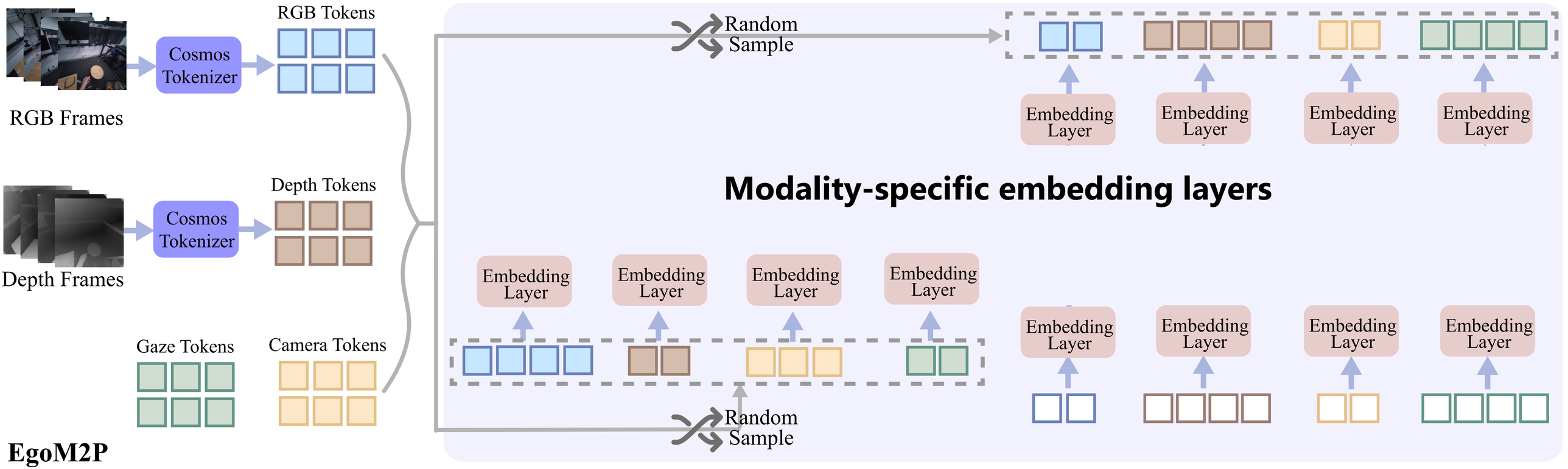
EgoM2P Training



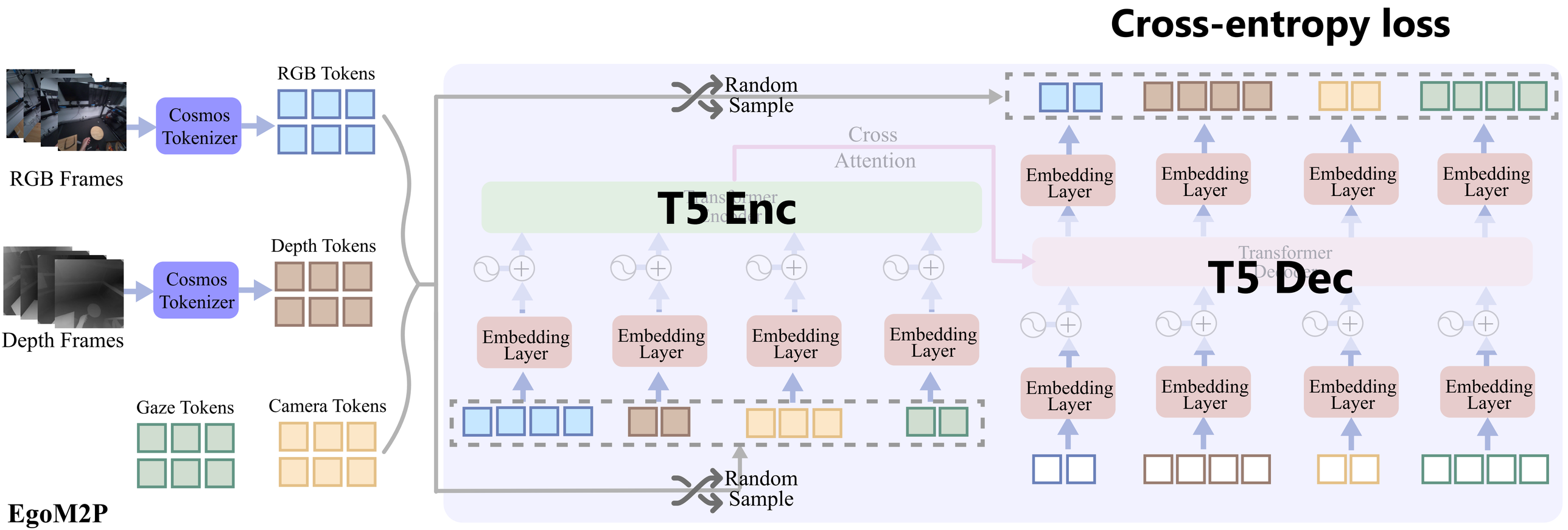
EgoM2P Training



EgoM2P Training



EgoM2P Training



256 H100 GPUs, 16 hours

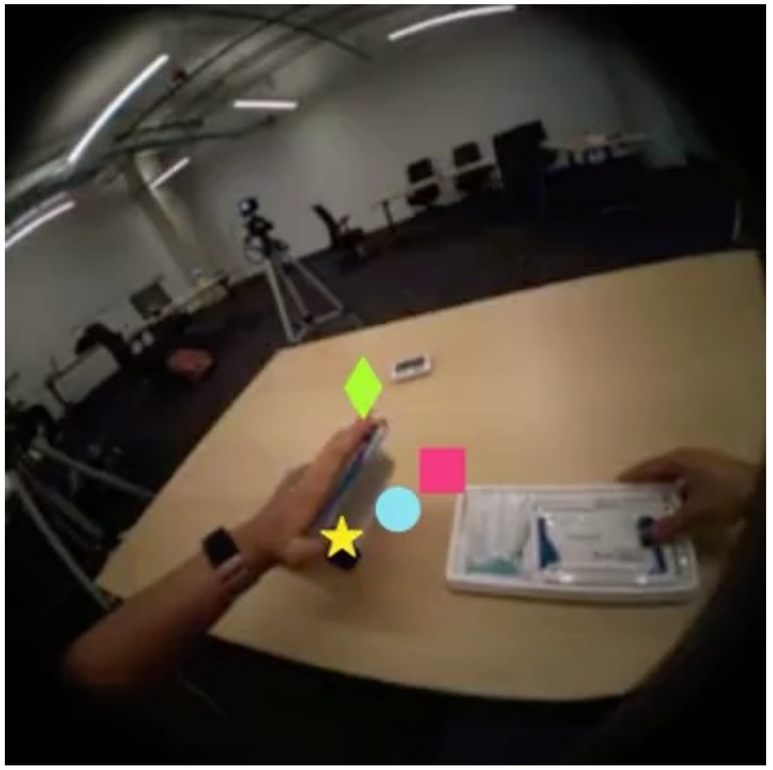
EgoM2P Inference

➡️Efficient inference with parallel decoding

- Variable masking rates to random mask out multimodal tokens
- => order-agnostic autoregressive model
- 300 FPS+ inference speed for egocentric camera tracking

Egocentric gaze estimation

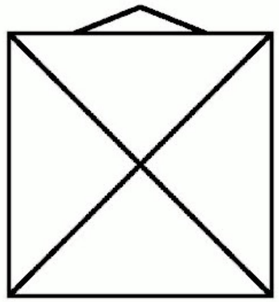
Ground Truth
Huang et al. 2018
Lai et al. 2022
Ours



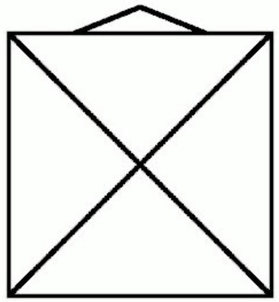
Egocentric camera tracking

Black wireframe: GT

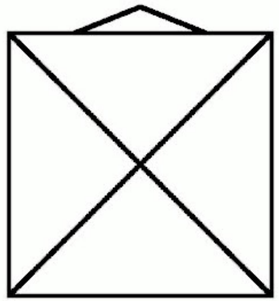
➡️EgoExo4D



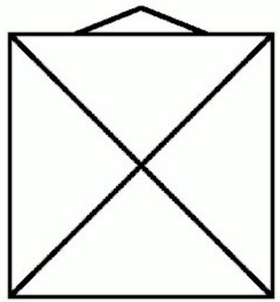
DROID-SLAM



ACE-Zero



Align3R

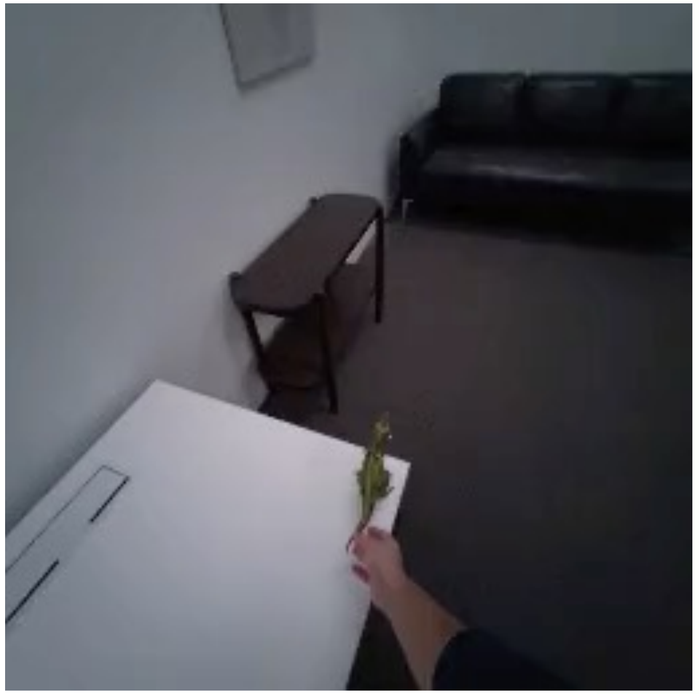


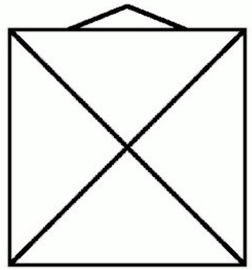
Ours

Egocentric camera tracking

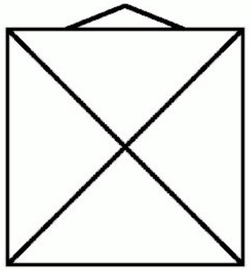
Black wireframe: GT

➡️ADT (unseen dataset)

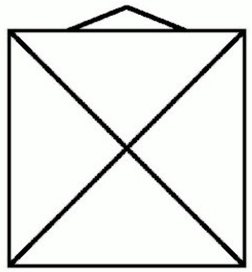




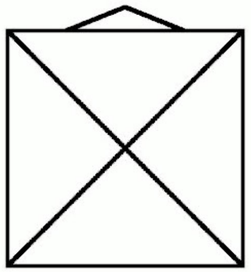
DROID-SLAM



ACE-Zero



Align3R



Ours

Egocentric camera tracking

Method	EgoExo4D [30] ATE↓	EgoExo4D [30] RTE↓	EgoExo4D [30] RRE↓	ADT [79] (*unseen*) ATE↓	ADT [79] (*unseen*) RTE↓	ADT [79] (*unseen*) RRE↓	Time↓
DROID-SLAM [93]	0.018	0.005	0.506	0.034	0.010	0.316	2.7s
ACE-Zero [13]	0.028	0.007	0.672	0.049	0.011	0.333	426s
Align3R [65]	0.019	0.006	0.762	**0.028**	0.010	**0.276**	372s
EgoM2P	**0.017**	**0.004**	**0.429**	0.032 **0.026**	**0.006** **0.005**	0.490 0.480	**0.18s**

Table 2. **Evaluation on camera tracking.** Compared to specialist SOTAs that require geometry test-time optimization, *EgoM2P*'s feed-forward tracking results achieve comparable performance yet with significantly higher efficiency. We report the average runtime per sequence. Underlined denotes post-training results (Sec. 4.5).

Egocentric depth estimation

➡️H2O

Input RGB

GT Depth

RollingDepth

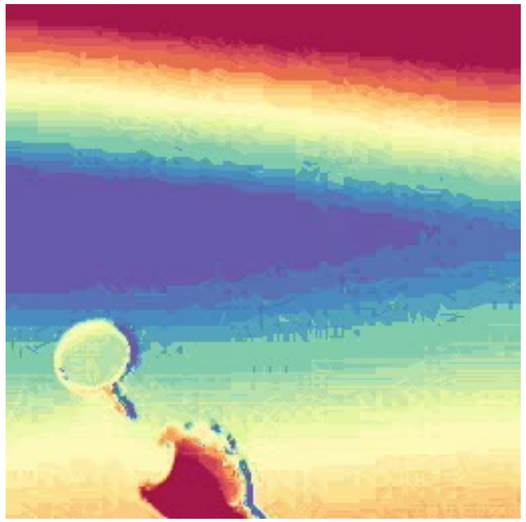
Align3r

Ours

Egocentric depth estimation

➡️HOI4D (unseen dataset)

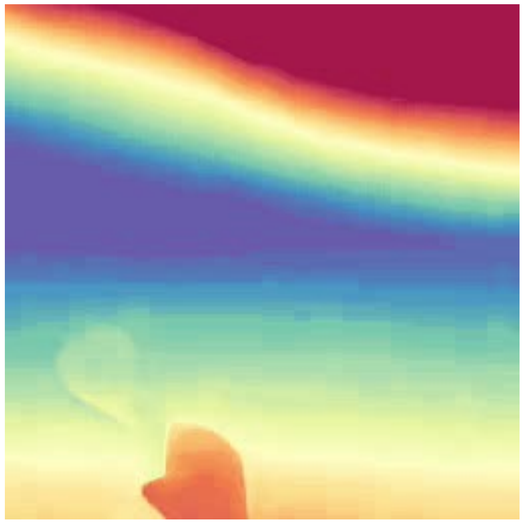
Input RGB



GT Depth

RollingDepth

Align3r



Ours

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	98.0	**0.8s**
			<u>0.041</u>	**<u>99.0</u>**	

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>).

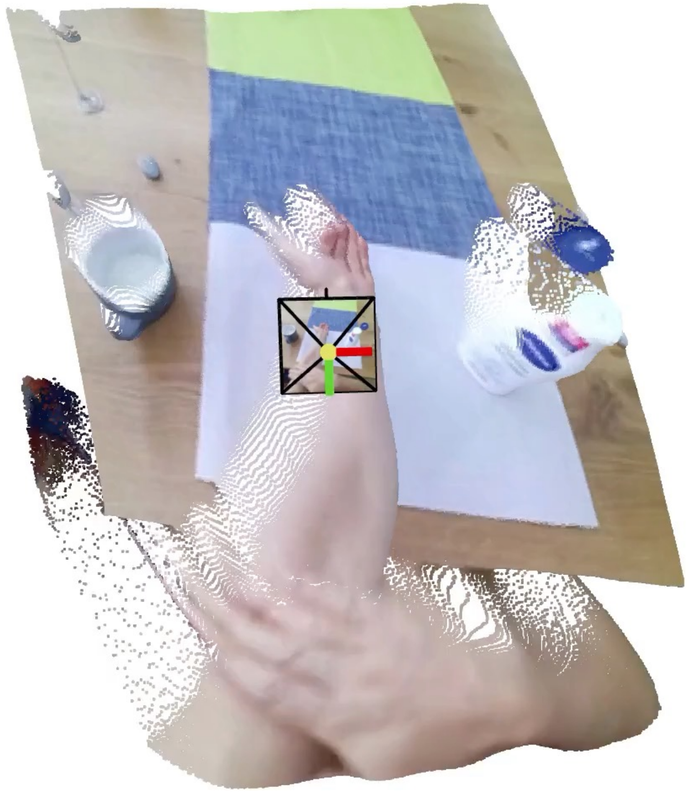
Egocentric video synthesis

Method	HoloAssist [106] FVD* ↓	SSIM ↑	PSNR ↑	LPIPS ↓	ASE [6] (*unseen*) FVD* ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Control-A-Video [19]	2.309	0.185	9.25	0.677	2.226	0.289	**11.11**	0.817
ControlVideo [126]	1.363	0.235	8.18	0.653	1.392	0.275	10.46	**0.676**
EgoM2P	**0.759**	**0.592**	**15.163**	**0.336**	**1.336**	**0.308**	6.923	0.715
					0.525	**0.594**	**16.924**	**0.520**

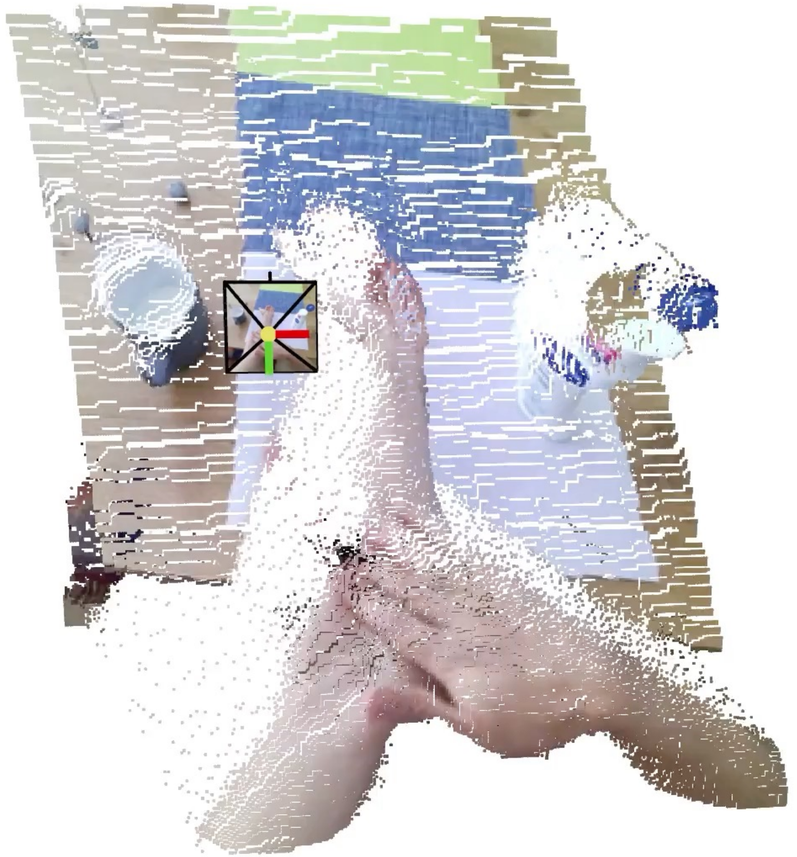
Table 4. **Evaluation on depth-to-RGB video synthesis.** *EgoM2P* outperforms baselines on the HoloAssist test set, producing higher-quality egocentric videos. On the unseen ASE dataset, it generates videos that more closely resemble real ones with a lower FVD* (FVD/10^3). With post-training (Sec. 4.5), *EgoM2P* excels on unseen datasets indicated by underlined results.

Input
Depth

Egocentric 4D Reconstruction



MegaSaM (71s)



Ours (<1s)

EgoGen boosts performance

Method	EgoExo4D [30] ATE↓	RTE↓	RRE↓	ADT [79] (*unseen*) ATE↓	RTE↓	RRE↓
EgoM2P w/o EgoGen	0.028	0.005	0.561	0.053	0.009	0.593
EgoM2P	**0.017**	**0.004**	**0.429**	**0.032**	**0.006**	**0.490**

Table B.2. **Ablation of EgoGen [52] on camera tracking.**

Method	H2O [49] Abs Rel ↓	$\delta_{1.25}$ ↑	HOI4D [59] (*unseen*) Abs Rel ↓	$\delta_{1.25}$ ↑
EgoM2P w/o EgoGen	0.062	94.9	0.067	97.1
EgoM2P	**0.055**	**96.0**	**0.061**	**98.0**

Table B.3. **Ablation of EgoGen [52] on depth estimation.**

Take home messages

- Egocentric synthetic data helps data scaling
- EgoM2P remains scalable to evolving datasets and new heterogeneous modalities.
- Promising zero-shot capability