Learning Multimodal Human Foundation Models

Challenges and Opportunities

Siyu Tang
Department of Computer Science
ETH Zürich

Are we experiencing the Cambrian explosion of artificial intelligence?

DeepSeek

SegmentAnything (Meta)

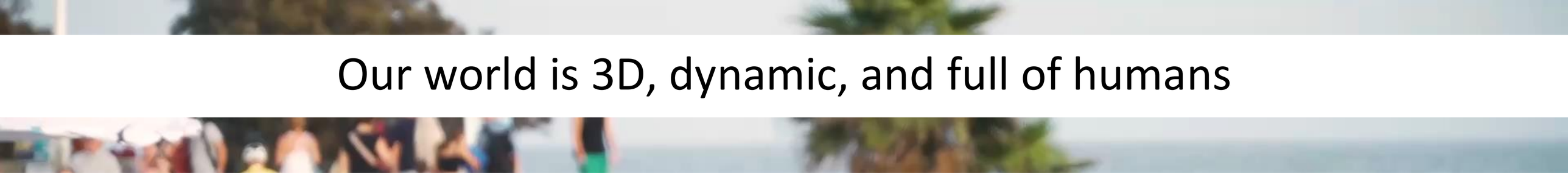
Sora (Open AI)

Are we experiencing the Cambrian explosion of artificial intelligence?

What is still missing?

AI’s ability to perceive the world in 3D

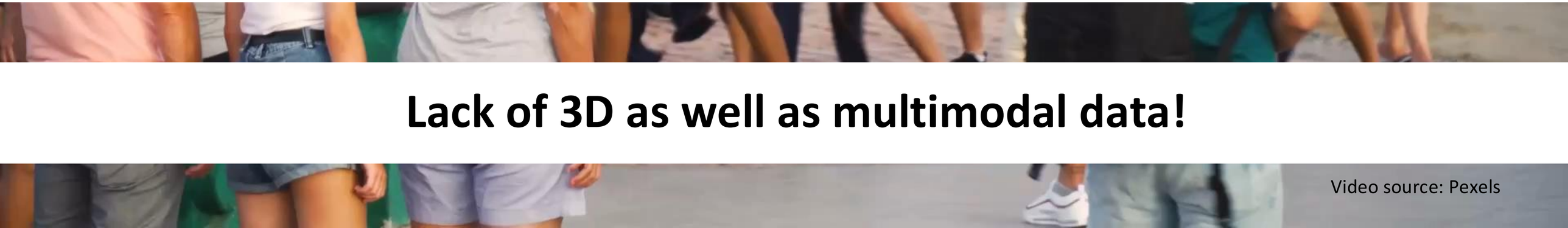
Image source: Pexels



Our world is 3D, dynamic, and full of humans

How can we learn a multimodal human foundation model for perceiving humans in 3D environments?

-“Human-centric multimodality”: 3D human motion, facial expression, speech, gaze, sound, even tactile?
-“in 3D environments”: human-centric 3D scene reconstruction
-“perceiving”: reconstruction, understanding (natural languages), prediction



Lack of 3D as well as multimodal data!

Video source: Pexels

Learning multimodality human foundation models
A data request

EgoGen, Li et al.

PROX, Hassan et al.

Nymeria, Ma et al.

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

In-the-wild videos
- Diverse motion and appearance
- Rich semantics, text, sound
- Limited 3D pseudo ground truth

Embodied egocentric captures
- Extended temporal duration
- Unique and close observations of hand-object interaction
- Multi-modality data
- Human motion capture with very limited observations

Furthermore, how can we handle such highly heterogeneous data sources?

EgoGen:

An Egocentric Synthetic Data Generator

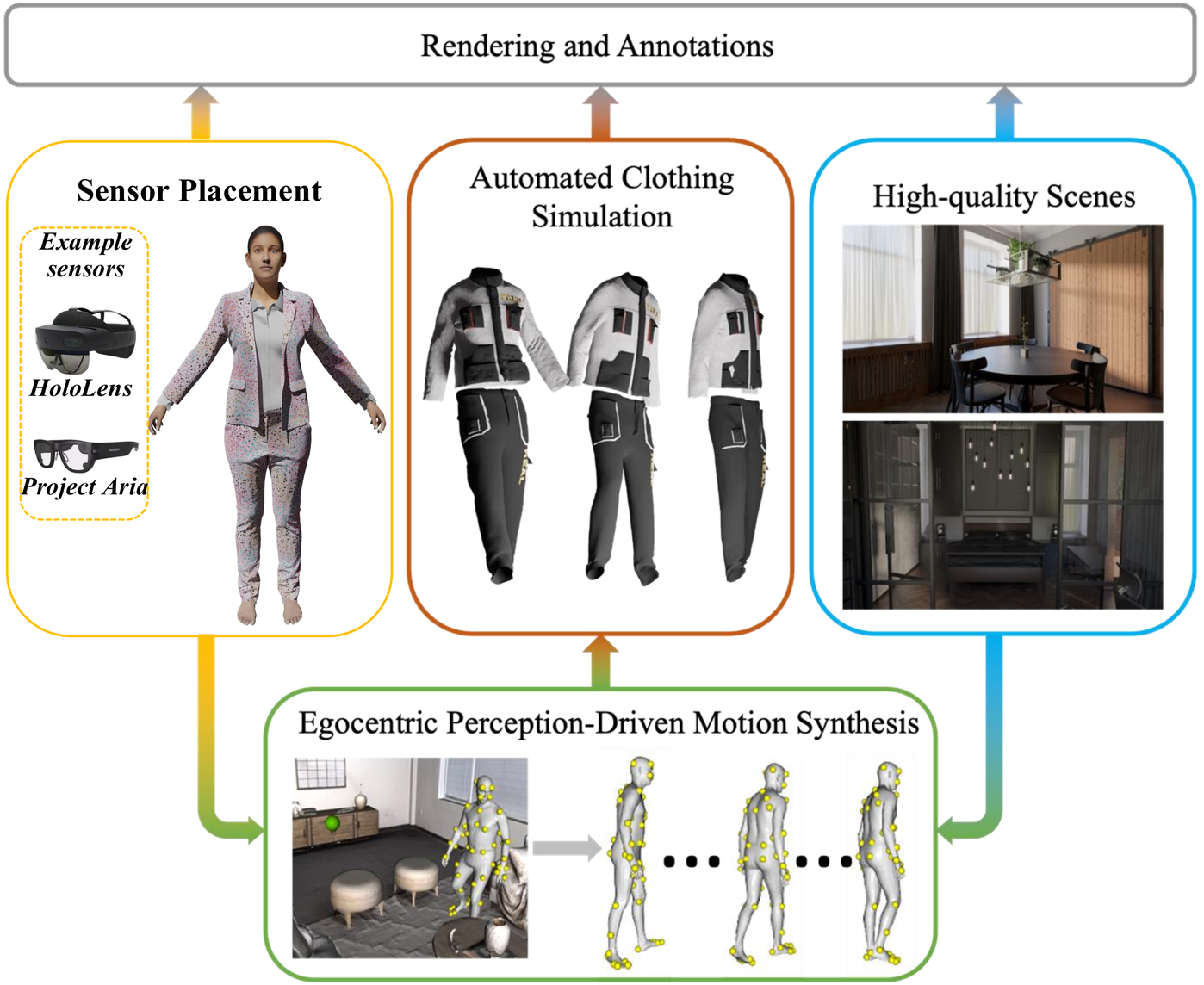
CVPR 2024 oral

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

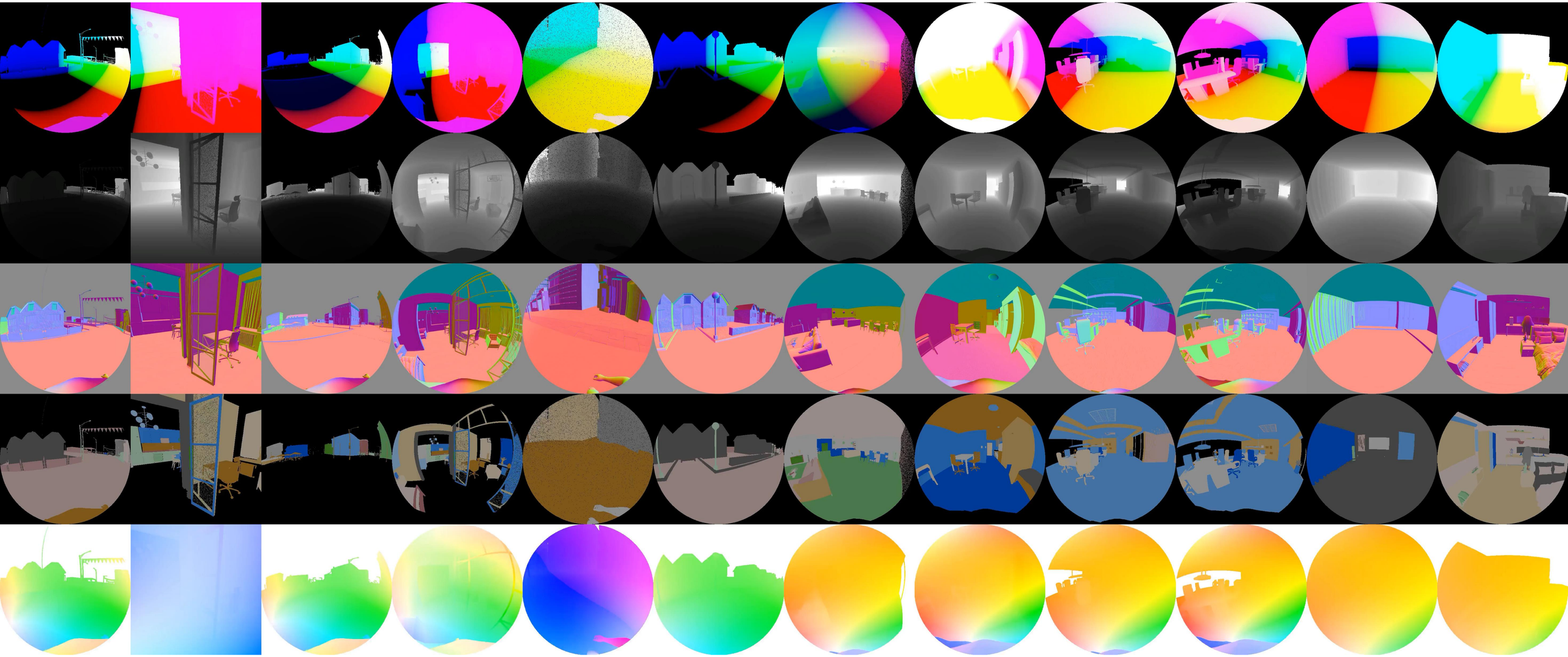
ETH zürich VLG Microsoft

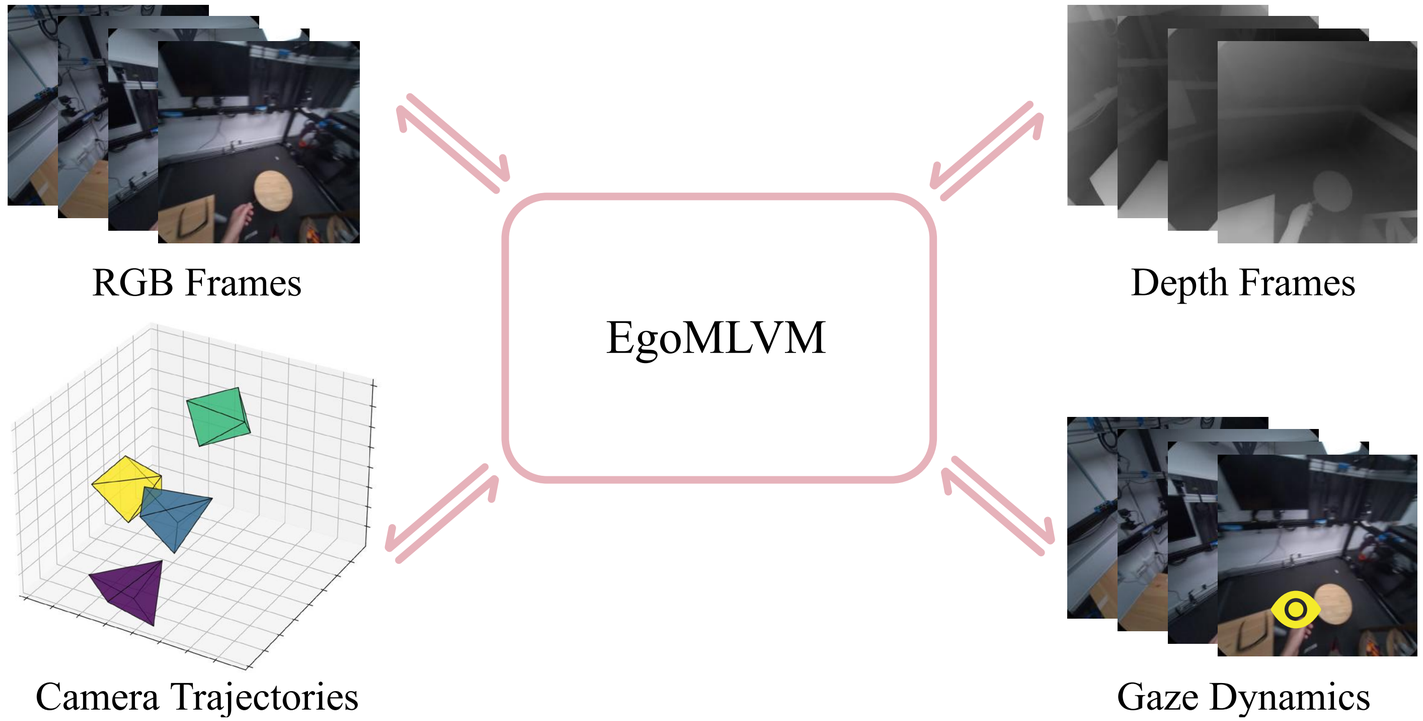
EgoGen: An Egocentric Synthetic Data Generator

Overview of EgoGen



EgoGen: An Egocentric Synthetic Data Generator





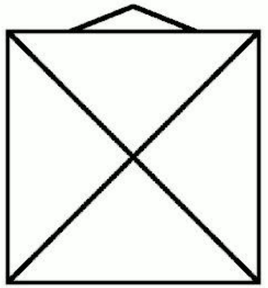
An Egocentric Multitask Multimodal Model

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

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)

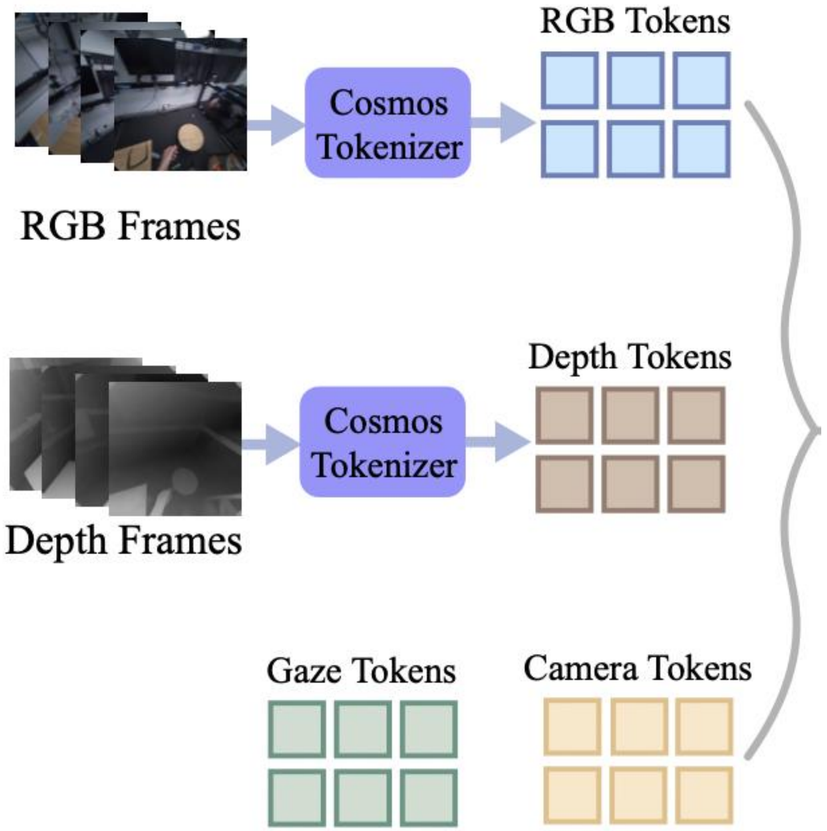
- Feasible to train large multimodal/task egocentric vision models

Challenges

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

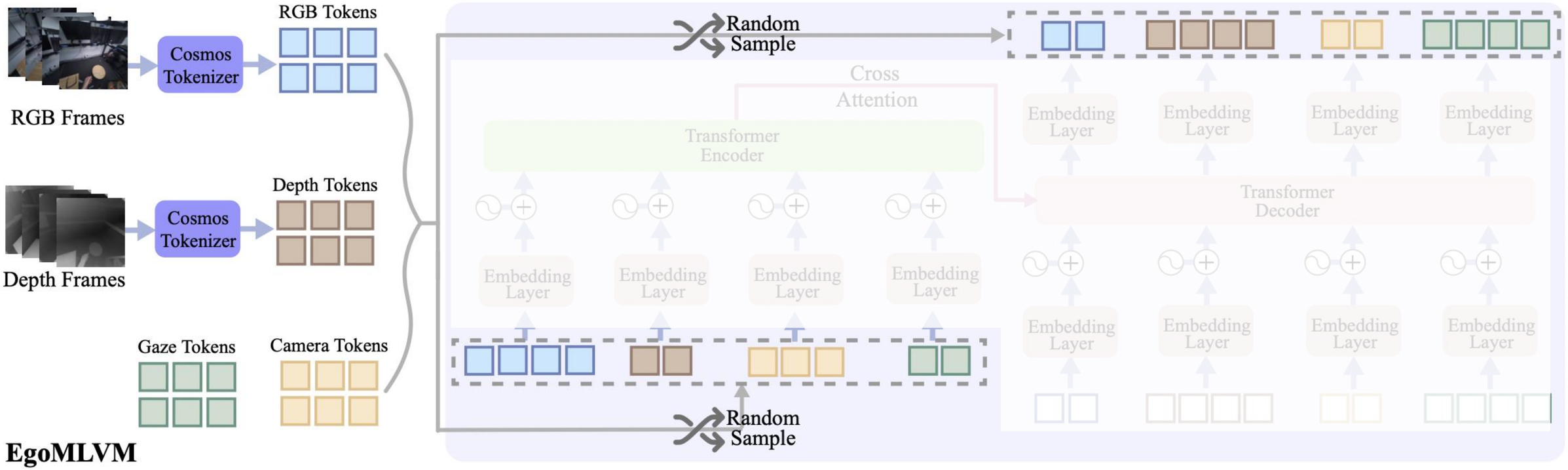
Datasets \ Modalities	RGB	Depth	Gaze	Camera
EgoExo4D [29]	✓	×	✓	✓
HoloAssist [103]	✓	×	✓	✓
HOT3D (Aria) [10]	✓	×	✓	✓
HOT3D (Quest) [10]	gray	×	×	✓
ARCTIC [23]	✓	×	×	✓
TACO [59]	✓	×	×	✓
H2O [48]	✓	✓	×	✓
EgoGen [51]	✓	✓	×	✓

EgoMLVM



4 billion multimodal token database

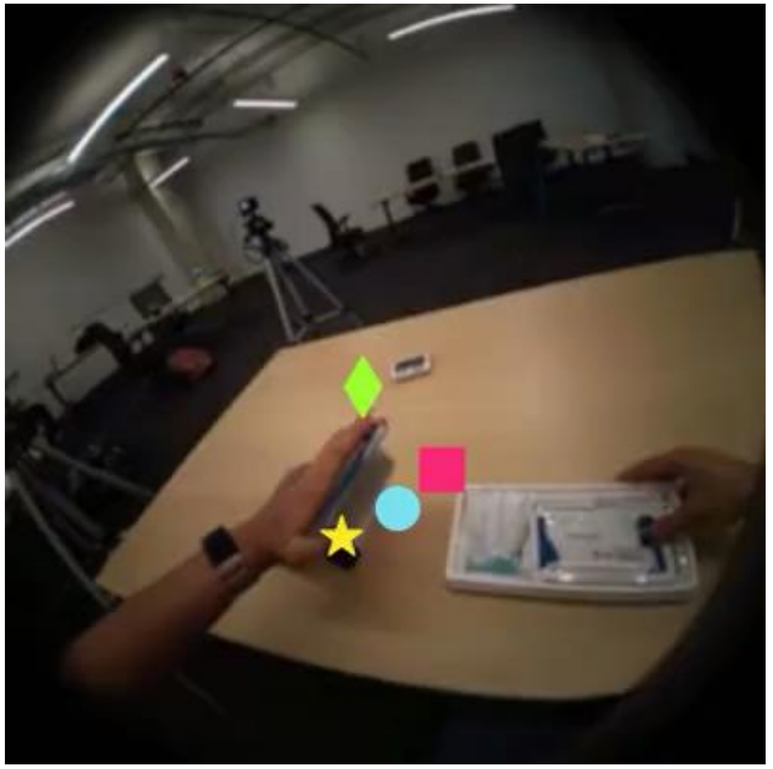
EgoMLVM



Variable masking rates to random mask out multimodal tokens

Egocentric gaze estimation

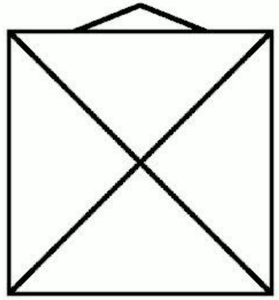
Ground Truth
Huang et al. 2018
Lai et al. 2022
Ours



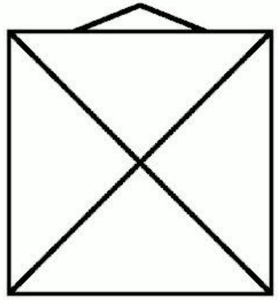
Egocentric camera tracking

Black wireframe: GT

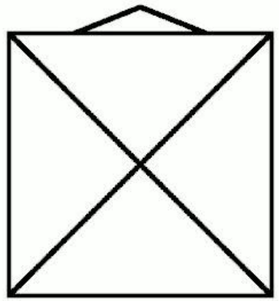
➡EgoExo4D



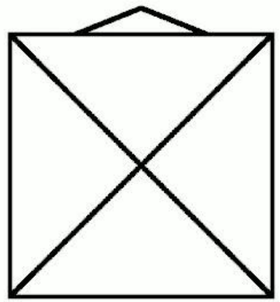
DROID-SLAM



ACE-Zero



Align3R

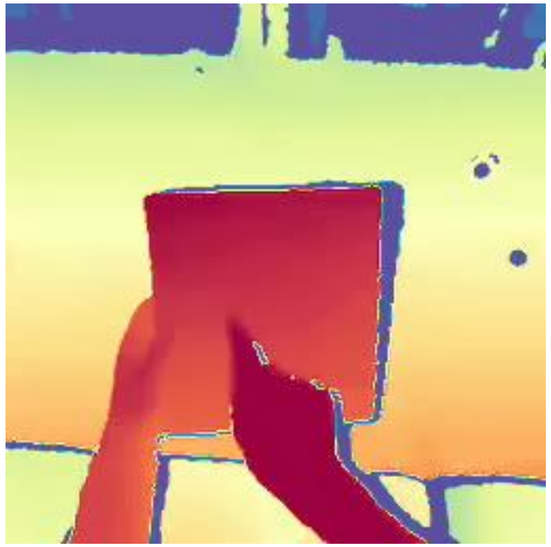


Ours

Egocentric depth estimation

→H2O

Input RGB



GT Depth

RollingDepth

Ours

Align3r

Egocentric video synthesis

Method	HoloAssist [103]				ASE [6] (*unseen*)			
	FVD* ↓	SSIM ↑	PSNR ↑	LPIPS ↓	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 [123]	1.363	0.235	8.18	0.653	1.392	0.275	10.46	**0.676**
EgoMLVM	**0.759**	**0.592**	**15.26**	**0.339**	**0.767**	**0.394**	8.73	0.679

Table 4. **Evaluation of depth-to-RGB video synthesis.** On the HoloAssist test set, our method produces higher quality videos compared to the baselines. Additionally, on the unseen ASE dataset, our approach generates egocentric videos that more closely resemble real ones, as indicated by a low FVD score. Note that FVD* represents FVD divided by 10^3.

Input
Depth

opportunities

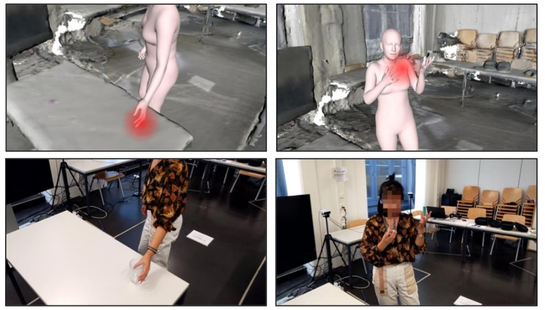
Learning multimodality human foundation models - ~~challenges~~

Bedlam, Black et al.

EgoGen, Li et al.

PROX, Hassan et al.

ProxyCap, Zhang et al.



EgoBody, Zhang et al.

Nymeria, Ma et al.

Synthesized virtual humans
- Rich and accurate 3D ground-truth annotations
- Controllability
- Generalization capability, LLM/VLM

In-the-wild videos
- Diverse motion and appearance
- Rich semantics
- Large feed-forward 3D human and scene reconstruction models

Embodied egocentric captures
- Extended temporal duration
- Unique and close observations of hand-object interaction
- Multi-modality data
- Egocentric hand–object manipulation

Unified models for imbalanced and diverse data sources and modalities

Thanks!