ARIA

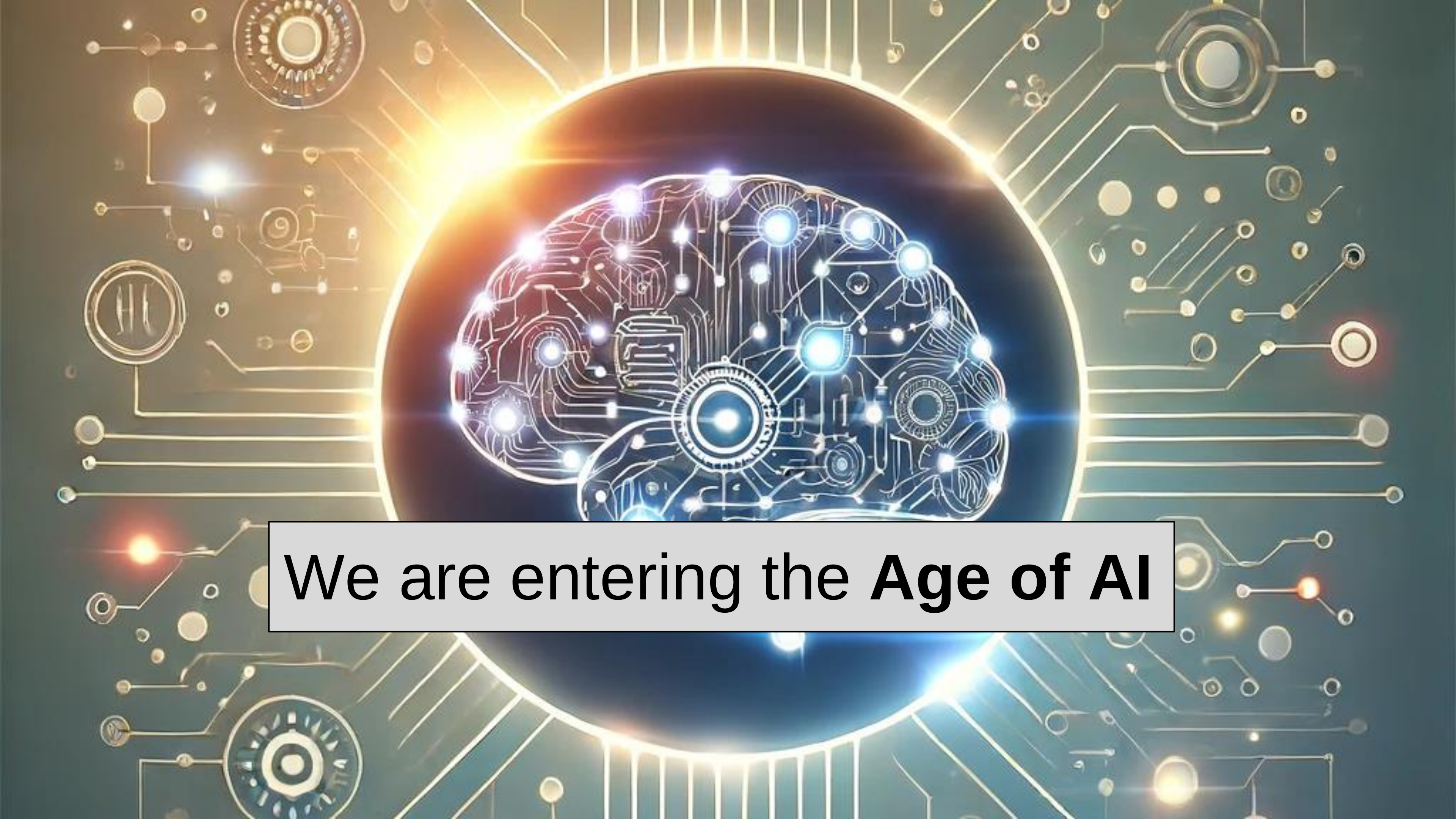
Localization & Mapping for **Contextual AI**

Jakob Engel

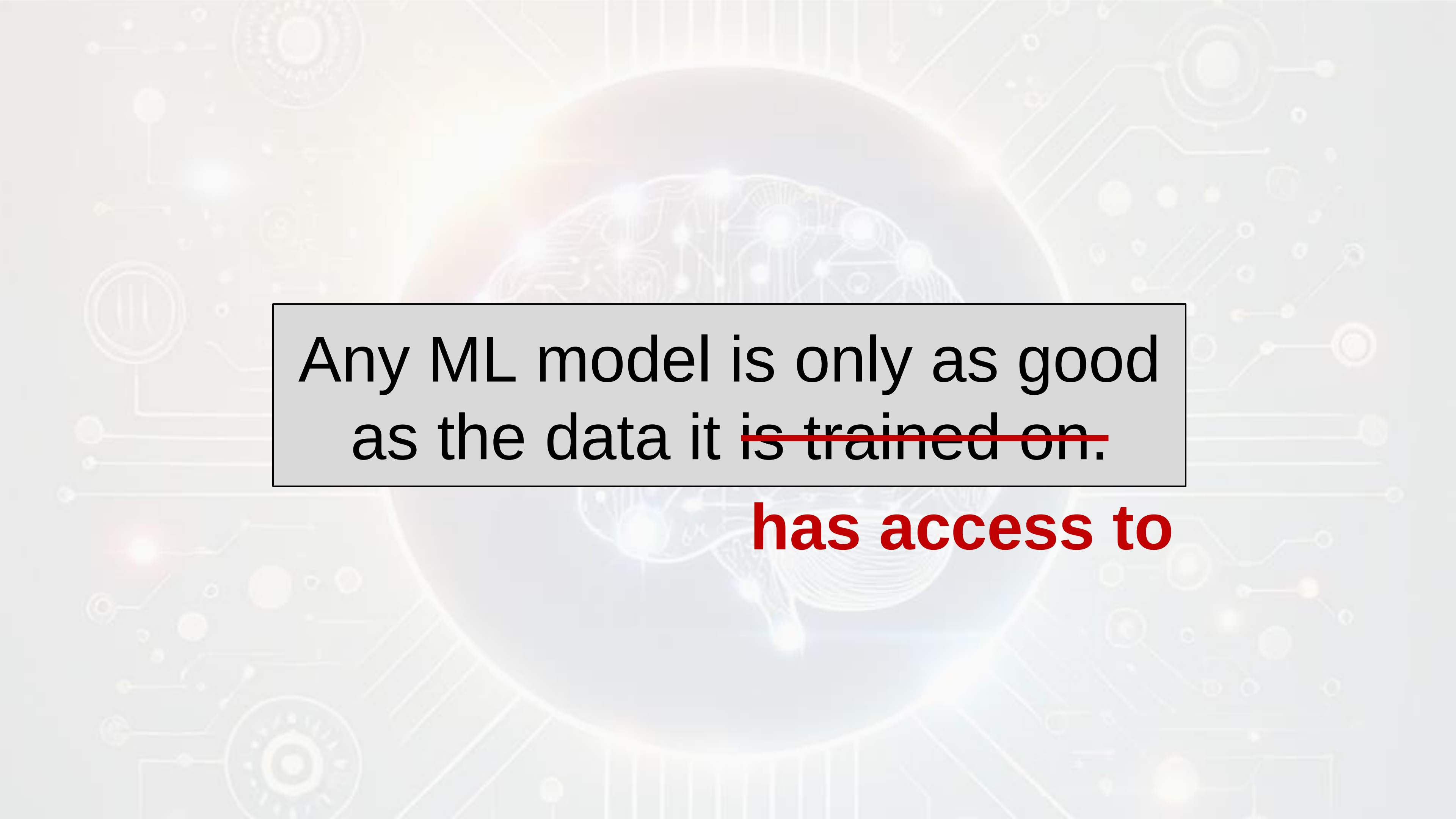
Director of Research

REALITY LABS

Reality Labs Research (RL-R)

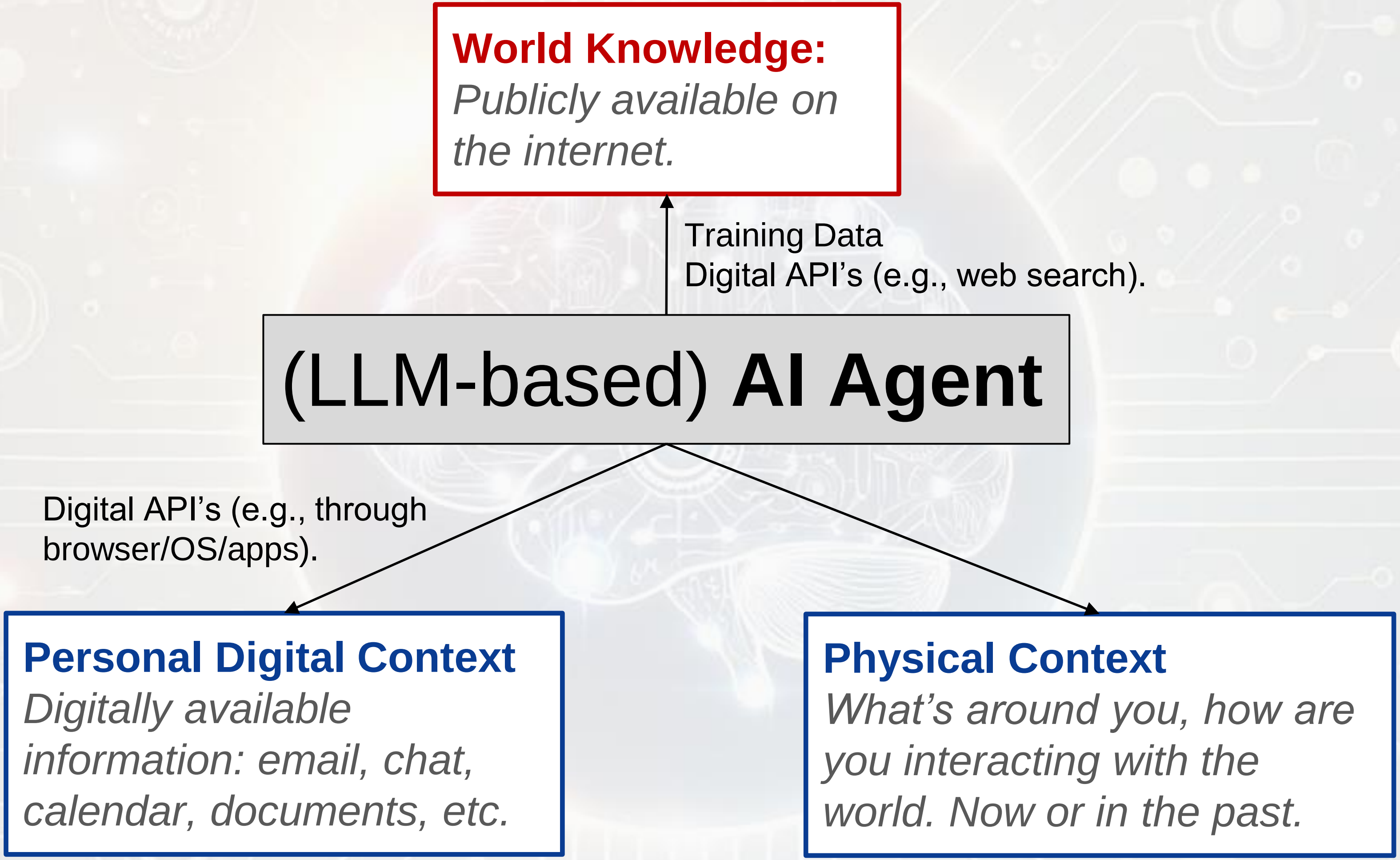


We are entering the **Age of AI**



Any ML model is only as good as the data it ~~is trained on.~~

has access to



World Knowledge:
Publicly available on the internet.

Training Data
Digital API’s (e.g., web search).

(LLM-based) **AI Agent**

Digital API’s (e.g., through browser/OS/apps).

Personal Digital Context
Digitally available information: email, chat, calendar, documents, etc.

Physical Context
What’s around you, how are you interacting with the world. Now or in the past.

Physical Context is *necessary* to make personal **AI Assistants** truly useful.

Where did I leave my keys?

Who did I bump into at the party last week?

What do I like to eat?

What do I typically do on Wednesdays?

What dishes can I cook in my kitchen?

When is my mom’s birthday?

Did I already put salt into the food?

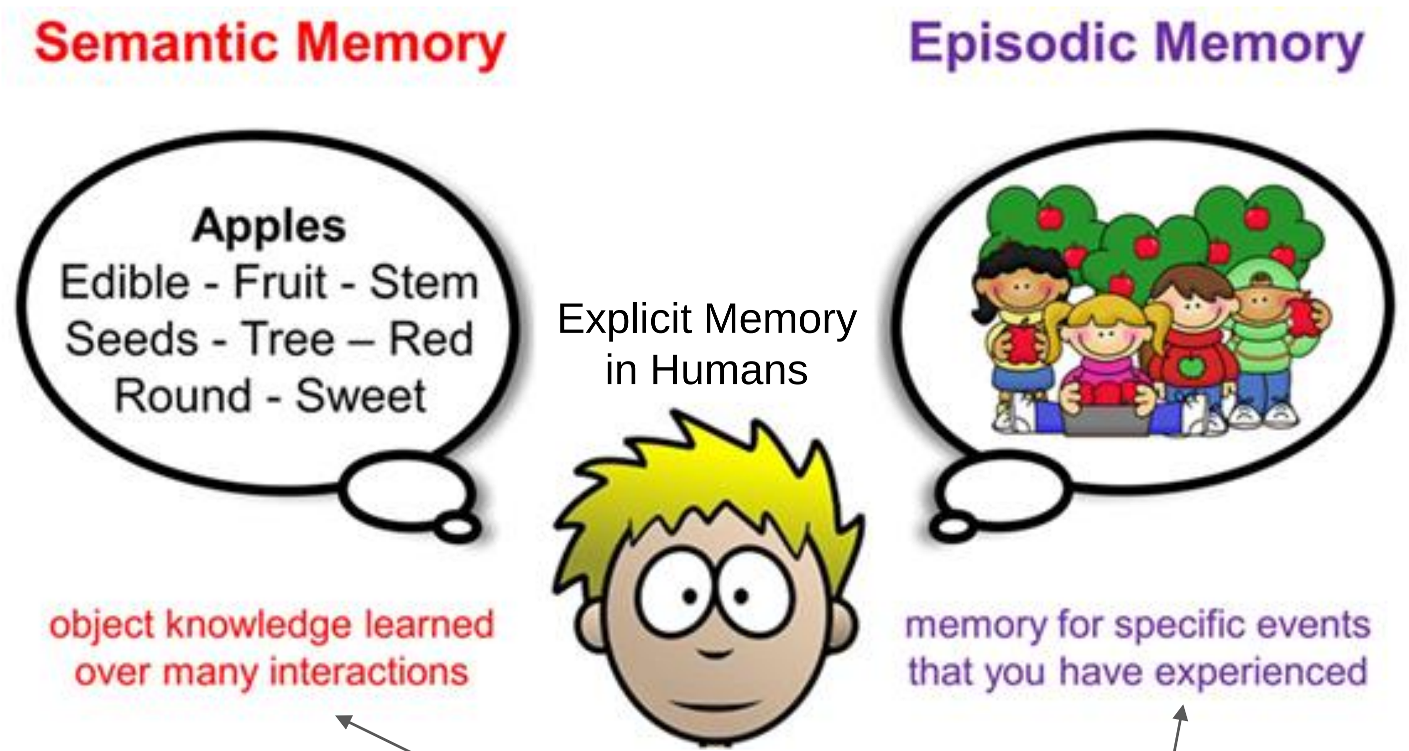
What did I do 2 weeks ago?

Do I still have milk at home?

Do I do enough exercise?

Where am I right now?

Who or what is around me?



Semantic Memory

Apples
Edible - Fruit - Stem
Seeds - Tree – Red
Round - Sweet

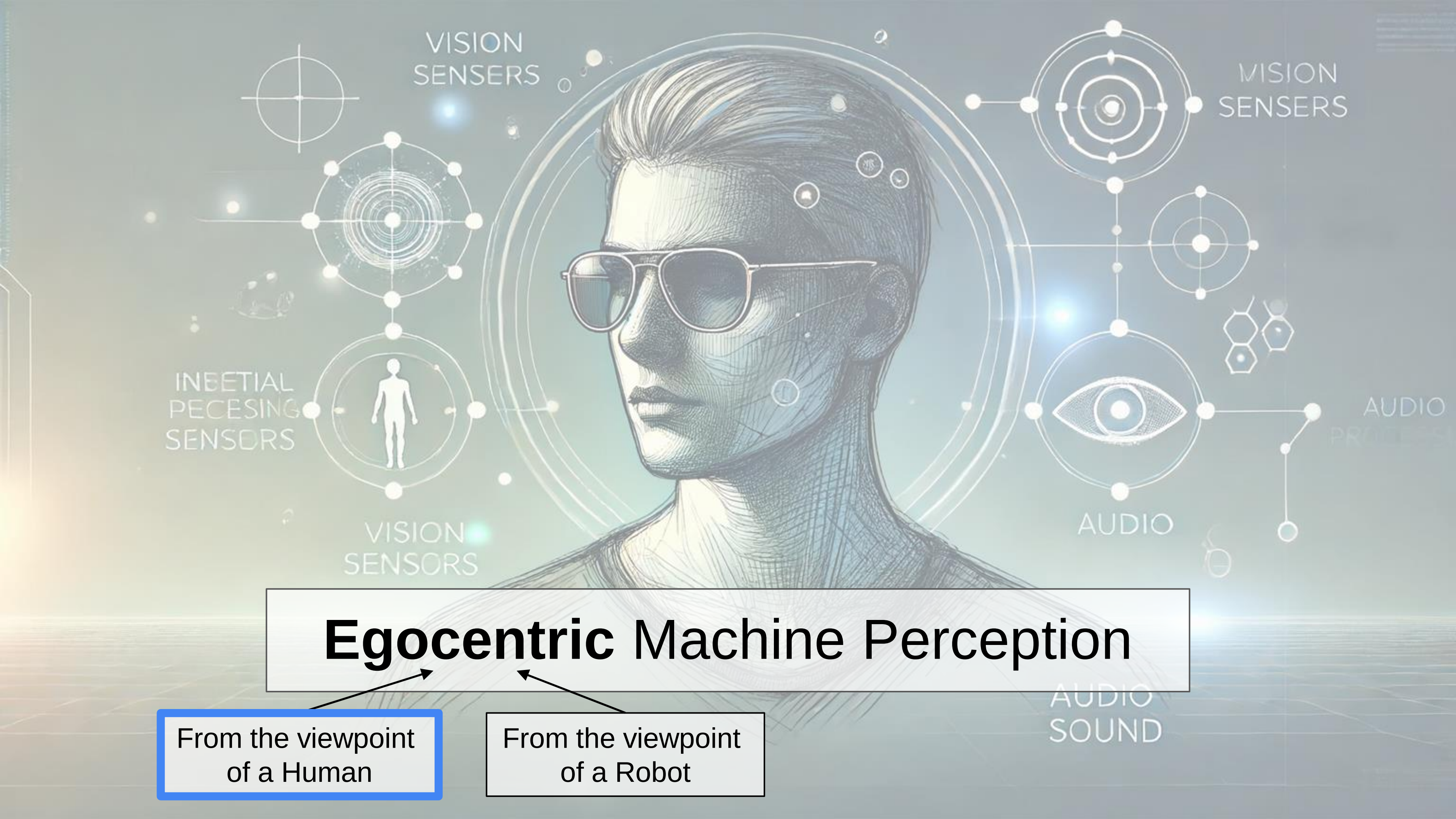
object knowledge learned over many interactions

Explicit Memory in Humans

Episodic Memory

memory for specific events that you have experienced

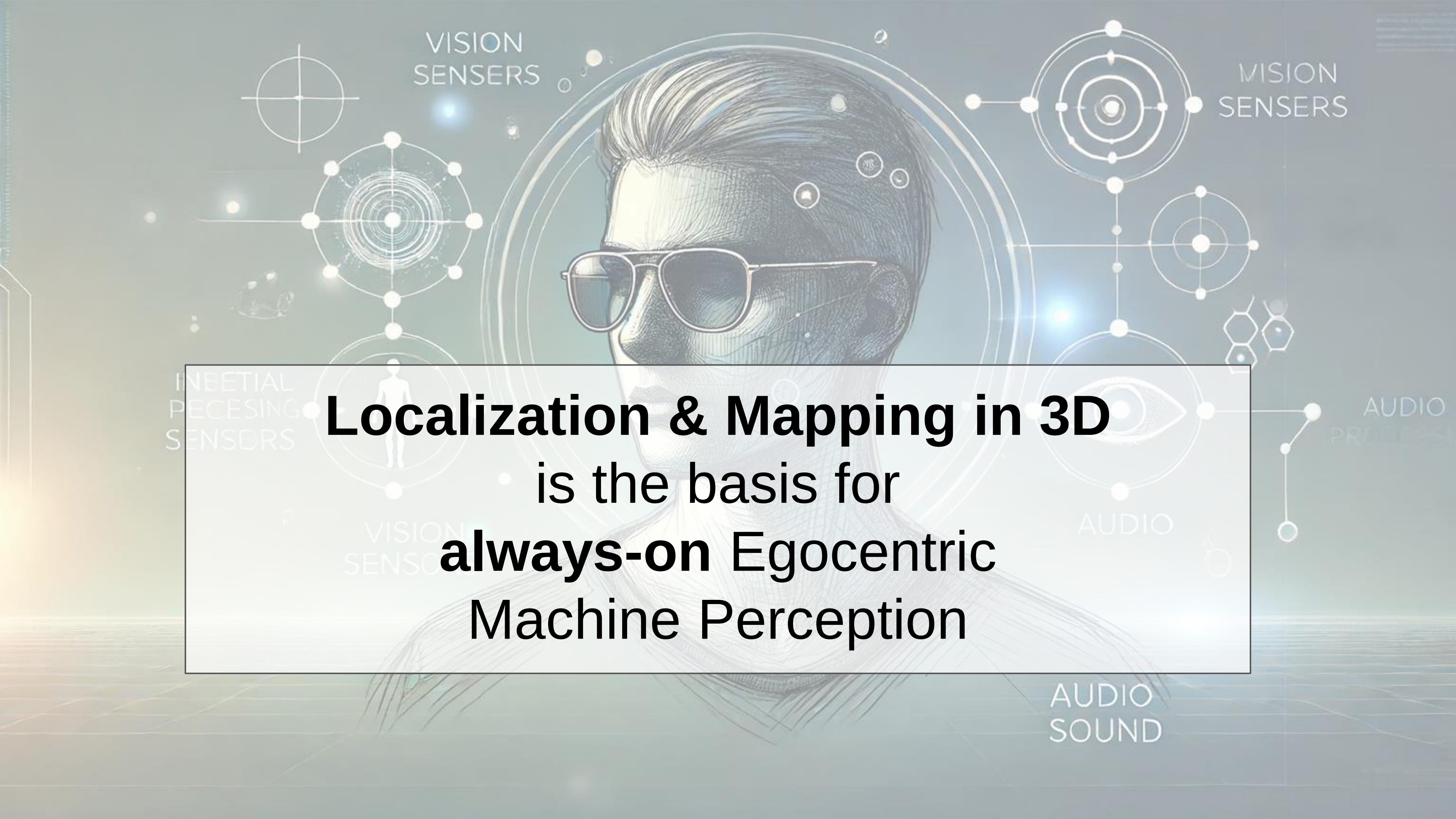
Current AI Agents have a lot of **this**, and almost none of **this**.



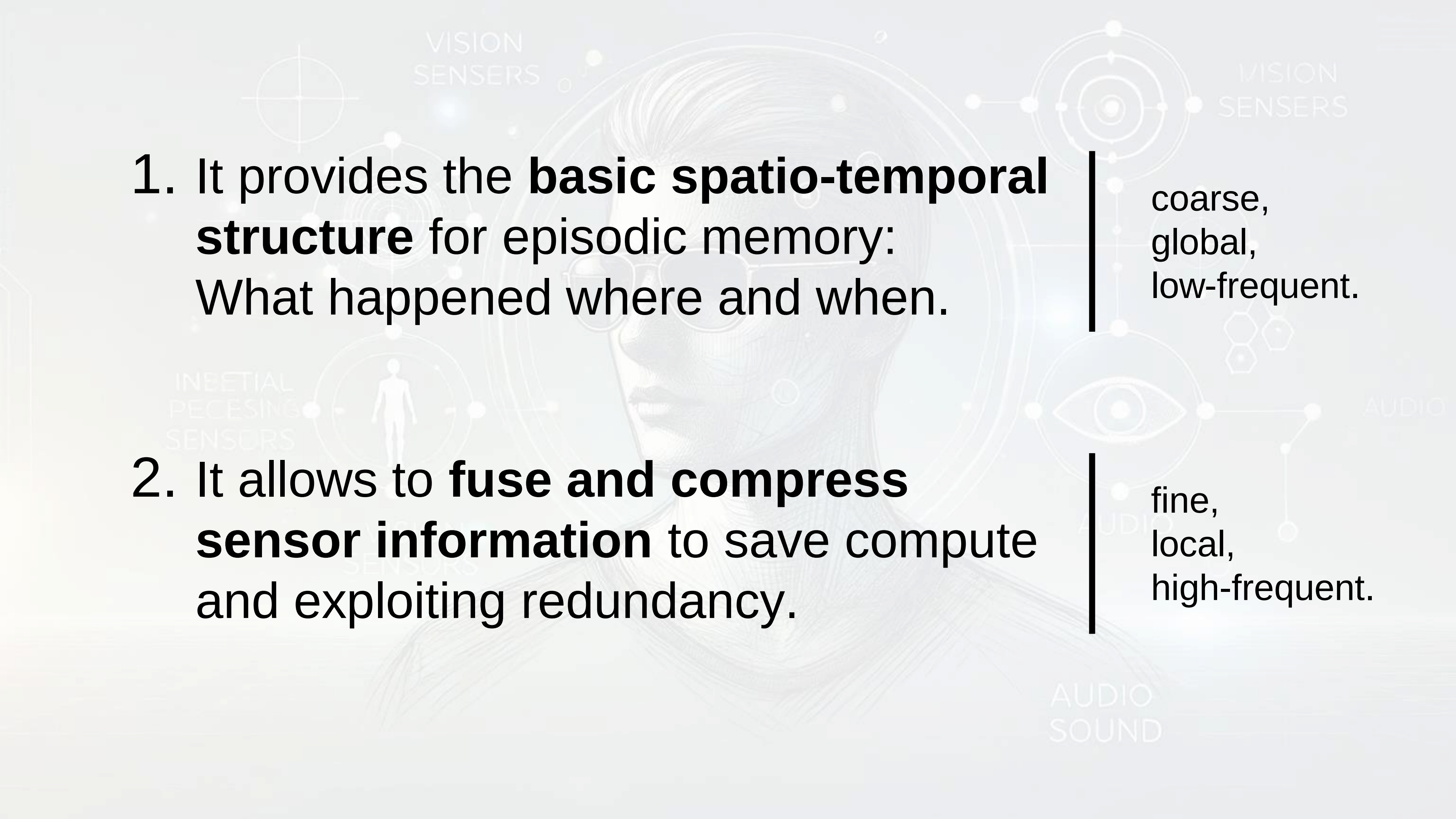
Egocentric Machine Perception

From the viewpoint of a Human

From the viewpoint of a Robot



Localization & Mapping in 3D
is the basis for
always-on Egocentric
Machine Perception



1. It provides the **basic spatio-temporal structure** for episodic memory:
   What happened where and when.

coarse,
global,
low-frequent.

2. It allows to **fuse and compress sensor information** to save compute and exploiting redundancy.

fine,
local,
high-frequent.

Introducing Project Aria

A Research Device to accelerate Machine Perception and AI Research

www.projectaria.com

[1] Project Aria: A New Tool for Egocentric Multi-Modal AI Research; ArXiv, Oct 2023

Project Aria: Multimodal egocentric sensing

1 RGB camera
2 SLAM cameras
2 eye tracking cameras
7 microphones
2 IMUs
Barometer
Magnetometer
WPS, BT, GPS

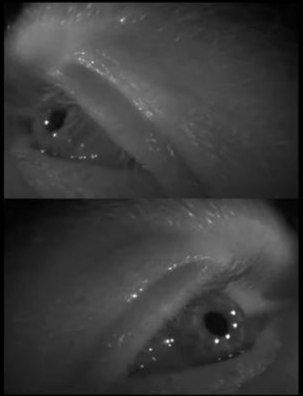
~75 gram
~1h recording time

Aria is a **Recording Device** only.

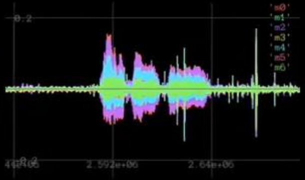
No display, no on-device compute.

mono left

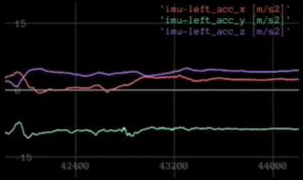
mono right



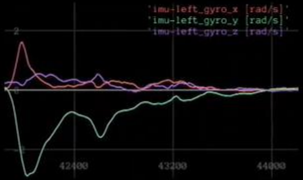
eye cameras



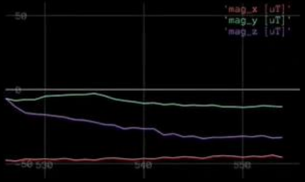
microphones



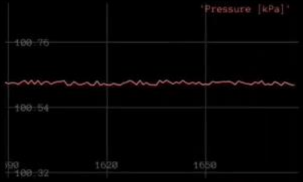
accelerometer



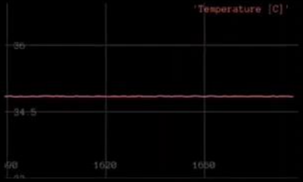
gyroscope



magnetometer



barometer



thermometer

ARIA

PROJECT ARIA'S GOAL IS TO ACCELERATE MP AND AI RESEARCH FOR FUTURE AR GLASSES.

Aria Research Kit

(for approved partners)

Open Science Resources

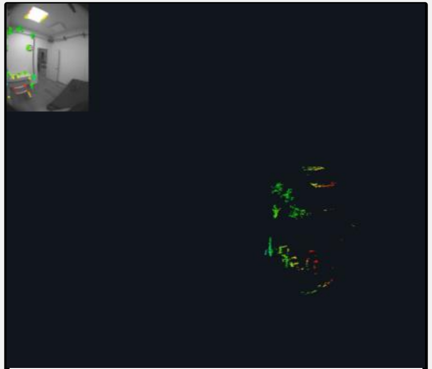
(no devices required)

Visit the **Meta Booth** to learn more.

www.projectaria.com

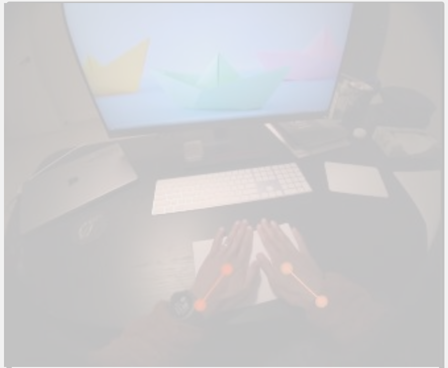
Aria Machine Perception Services

6DoF Location



3D Pointclouds

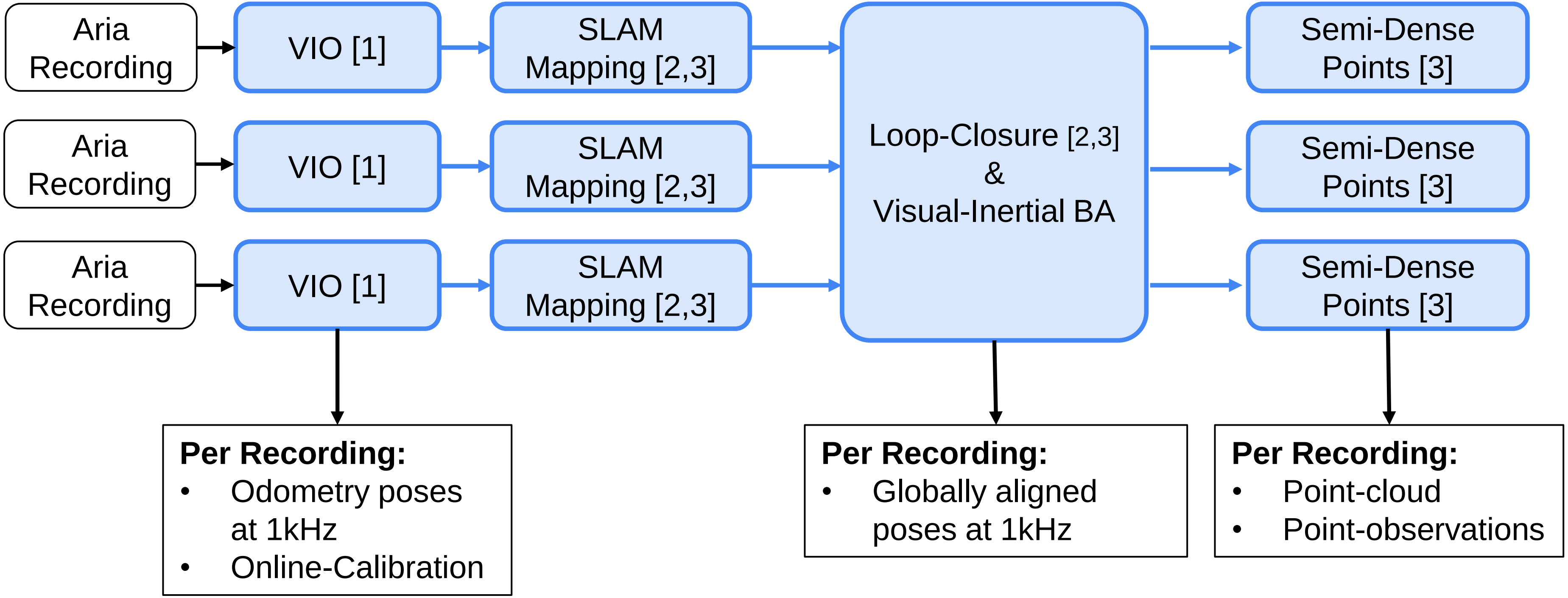
Eye Gaze



Palm & Wrist

+ Factory Calibration
+ Accurate Sensor Models
+ Time Sync across all sensors

Aria Machine Perception Services: Points & Poses



Aria Recording

VIO [1]

SLAM Mapping [2,3]

Aria Recording

VIO [1]

SLAM Mapping [2,3]

Aria Recording

VIO [1]

SLAM Mapping [2,3]

Loop-Closure [2,3] & Visual-Inertial BA

Semi-Dense Points [3]

Semi-Dense Points [3]

Semi-Dense Points [3]

Per Recording:

- Odometry poses at 1kHz
- Online-Calibration

Per Recording:

- Globally aligned poses at 1kHz

Per Recording:

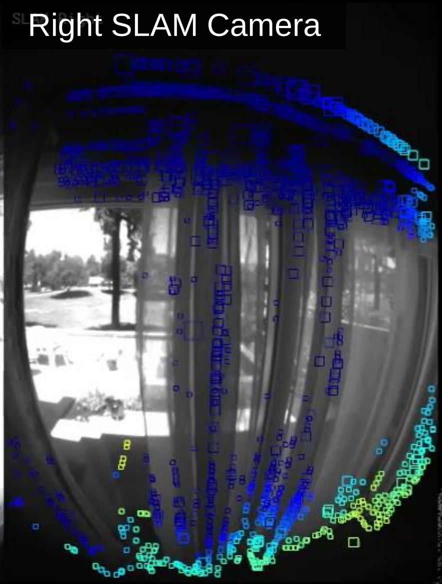
- Point-cloud
- Point-observations

[1] “A multi-state constraint Kalman filter for vision-aided inertial navigation”; Mourikis et.al.; ICRA 2007

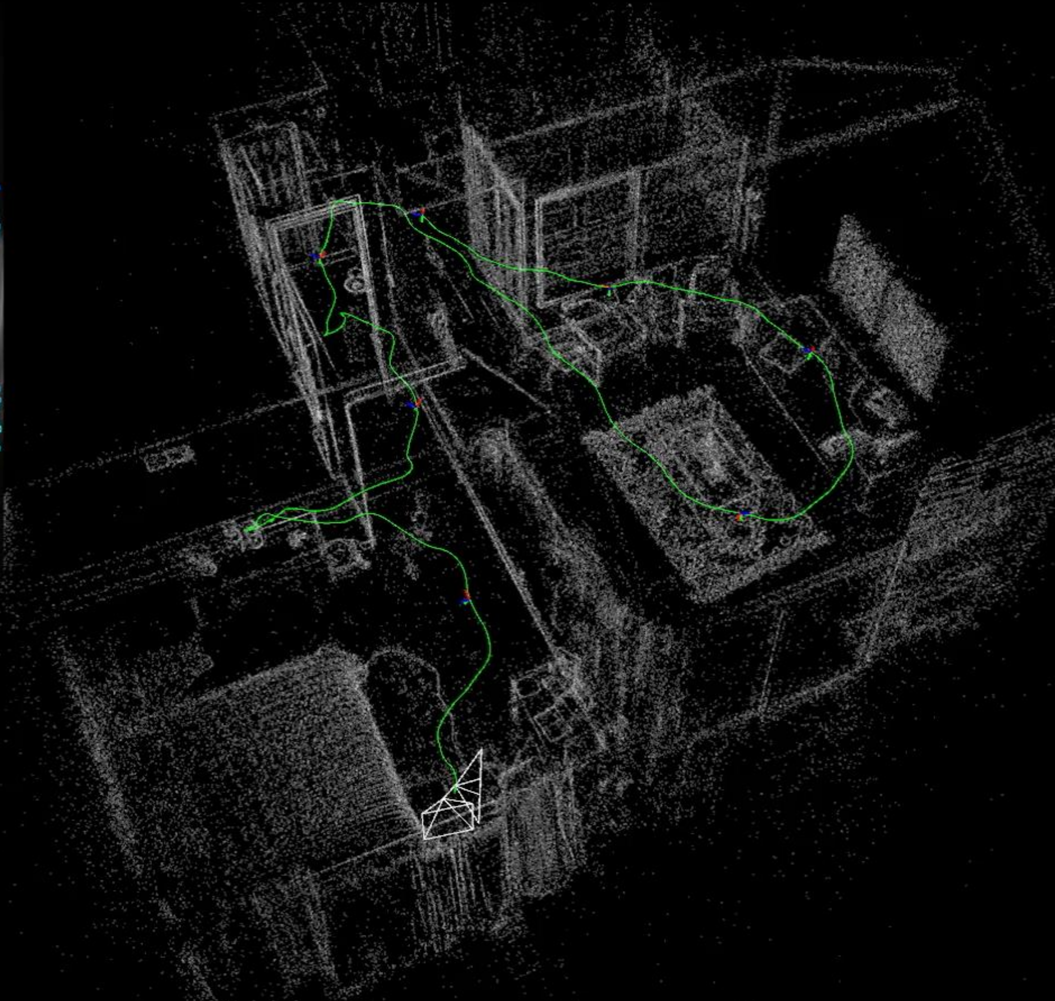
[2] “ORB-SLAM: A Versatile and Accurate Monocular SLAM System”; Mur-Artal et.al.; TRO 2015

[3] “Direct Sparse Odometry”; Engel et.al.; TPAMI 2016

Left SLAM Camera



Right SLAM Camera



Location + **Point Cloud** from Aria Machine Perception Services (MPS)

Why
Online Calibration
matters!

New Aria-based Datasets

Bringing Egocentric Machine Perception into 3D

Ego-Exo 4D

Skilled human activity understanding.

ego-exo4d-data.org

Nymeria

Human Motion in the wild.

projectaria.com/datasets/nymeria

HOT3D

Hand / Object interaction.

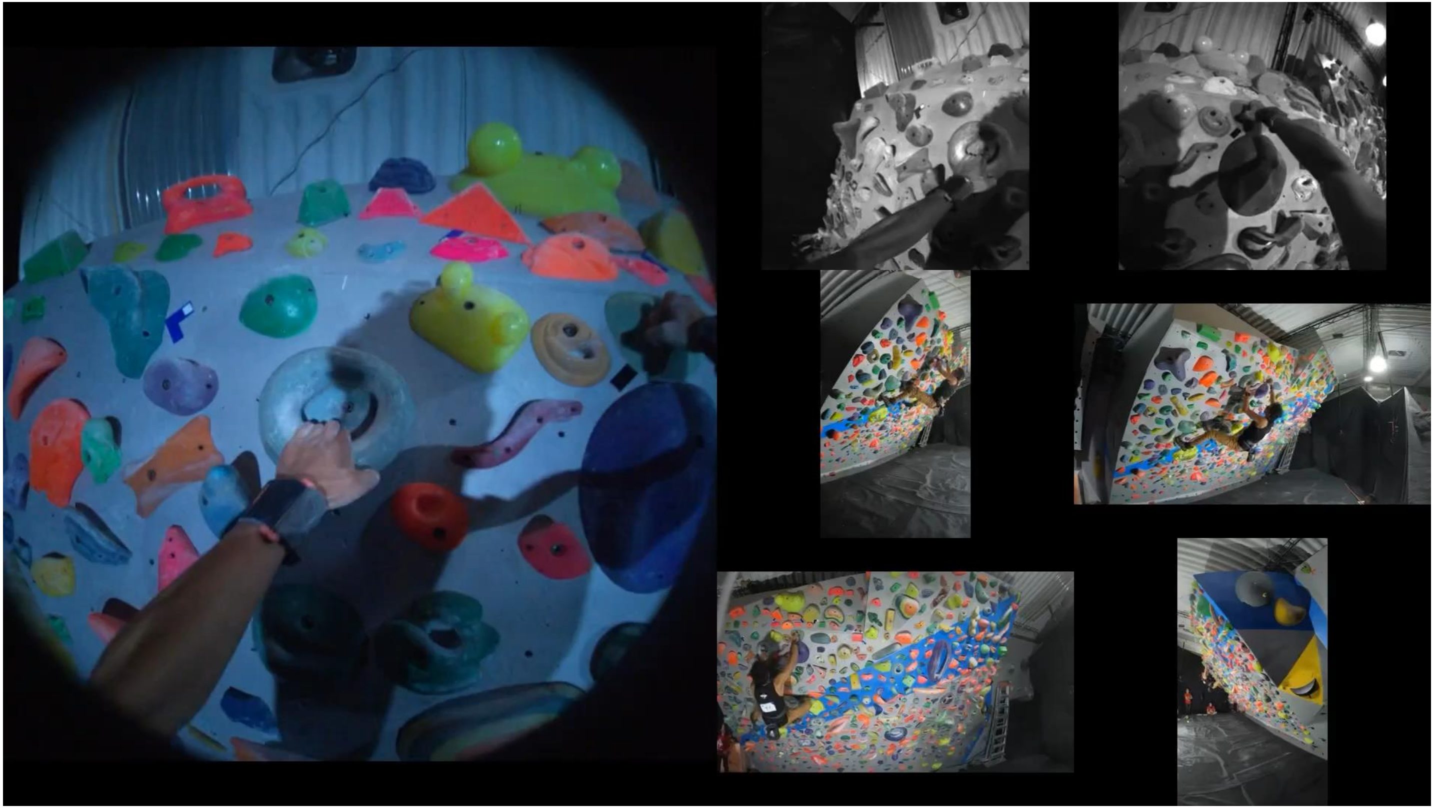
projectaria.com/datasets/hot3d

* CVPR 2024 paper!

* Released this week!

* Released this week!

Ego-Exo 4D



Egocentric

Exocentric

Large-scale ego-exo capture of skilled activity

800+ participants

130+ real-world environments

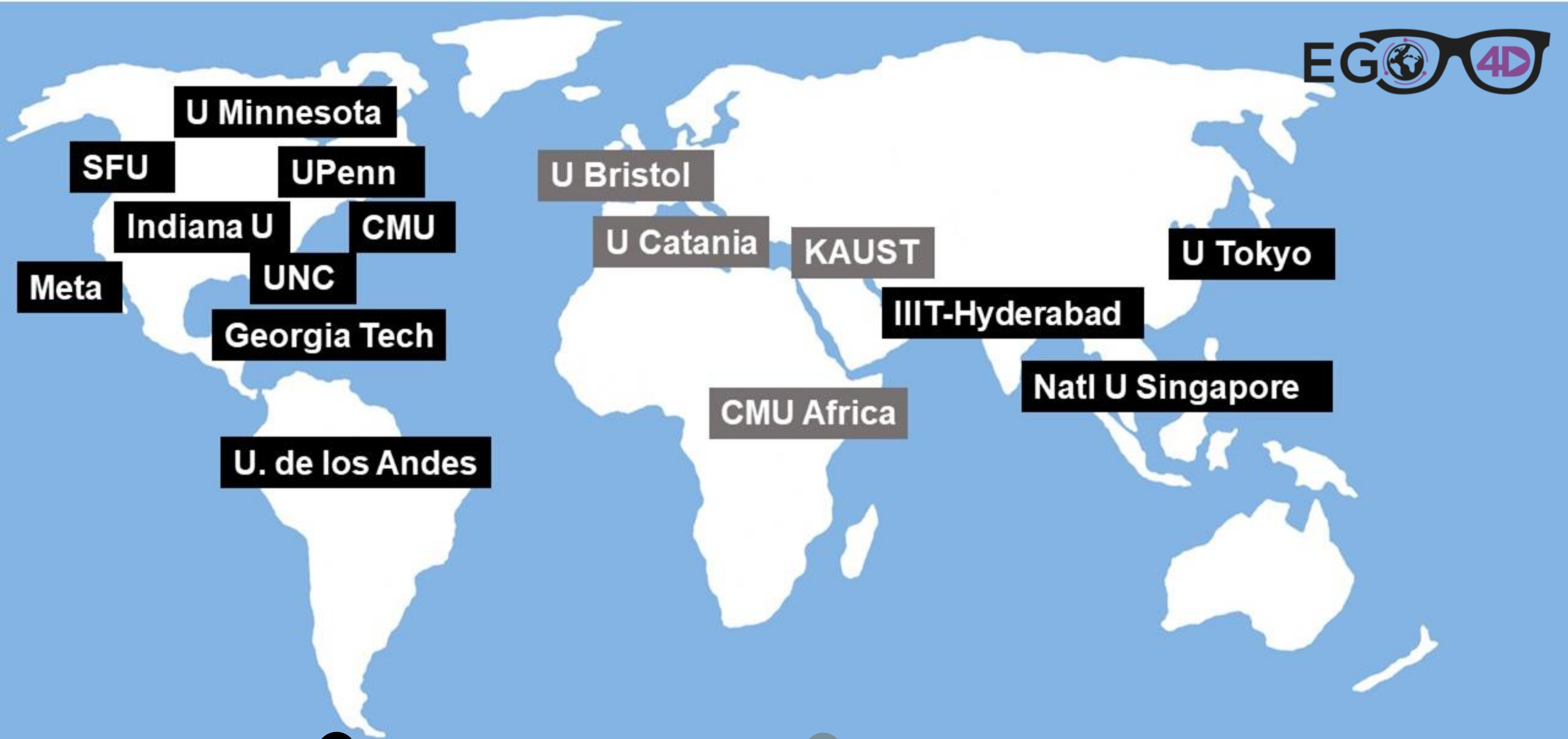
15 cities worldwide

5,000+ takes

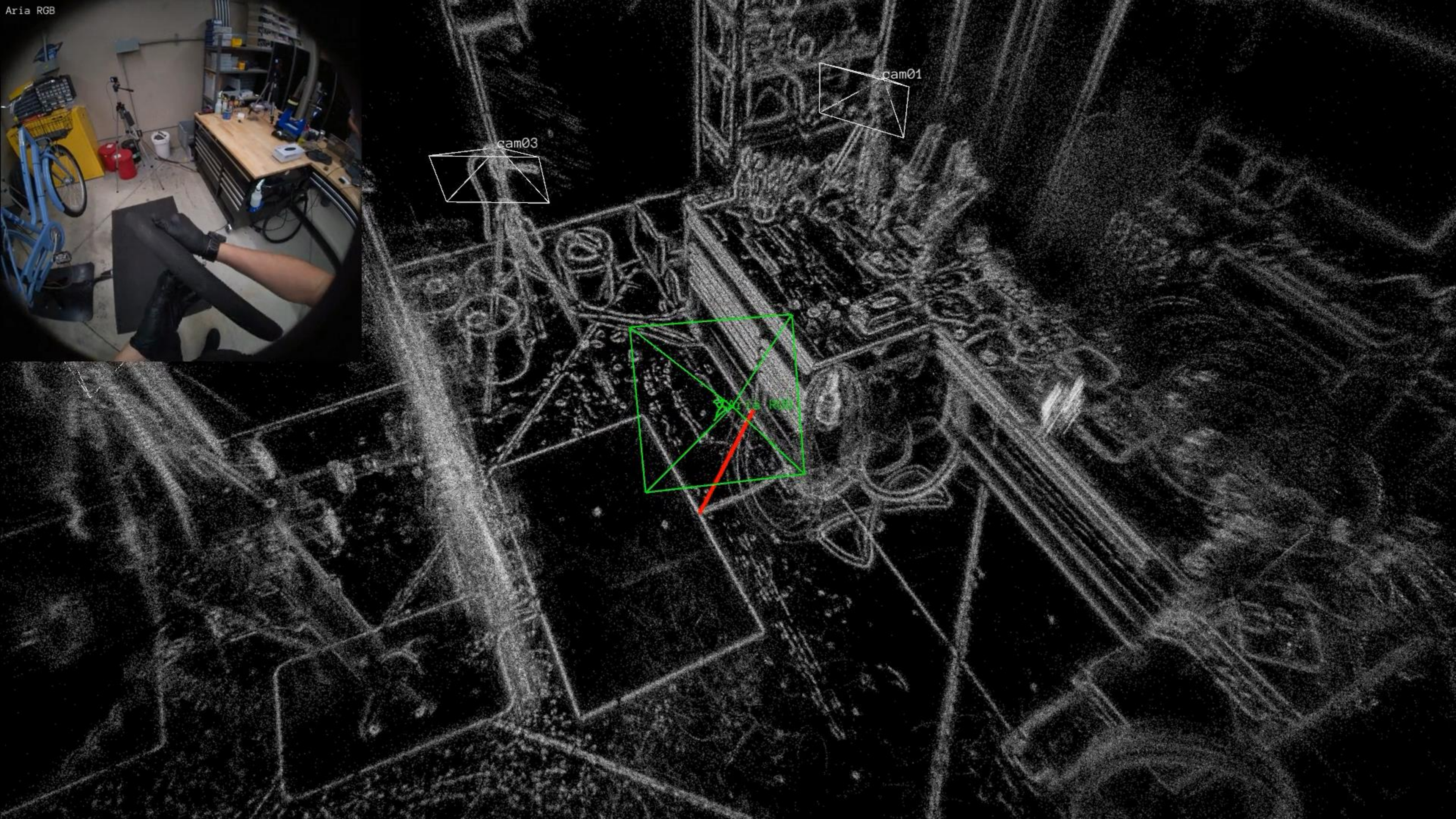
1,400 hours of video (ego + exo)

Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives; Grauman et.al.; CVPR 2024

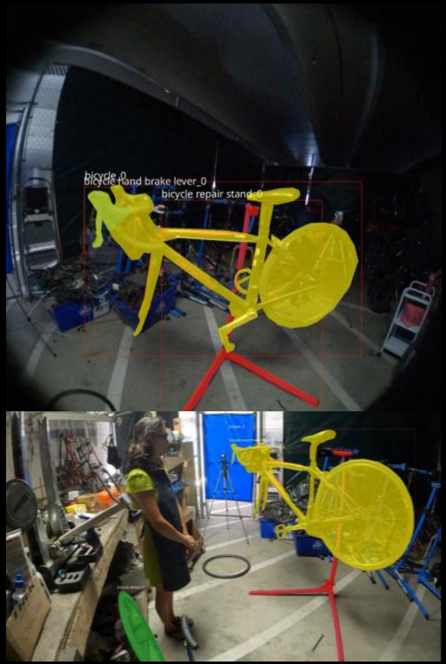
Ego4D university consortium

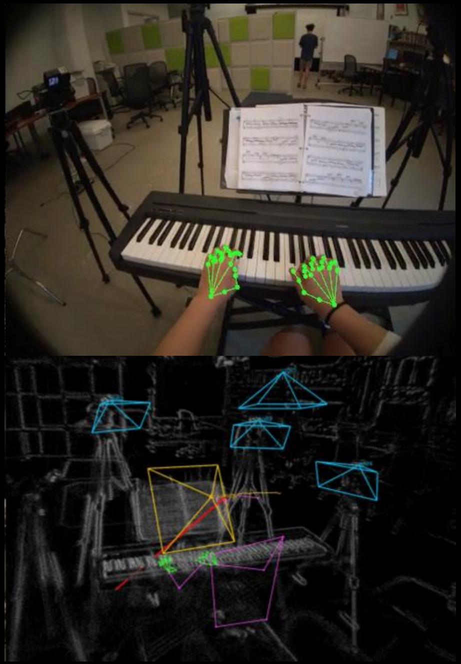


● Data collecting university ● Non-data collecting university



Aria RGB
cam03
cam01
Aria RGB





Keystep recognition

Ego-exo relation

Ego pose

Proficiency

New Aria-based Datasets

Bringing Egocentric Machine Perception into 3D

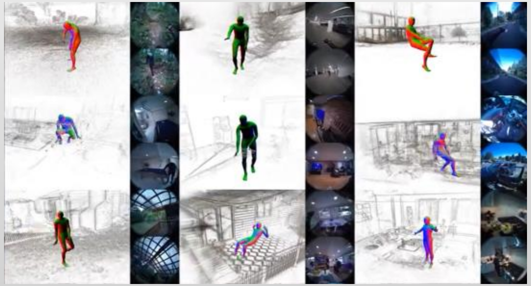
Ego-Exo 4D

Skilled human activity understanding.

ego-exo4d-data.org

* CVPR 2024 paper!

Nymeria



Human Motion in the wild.

projectaria.com/datasets/nymeria

* Released this week!

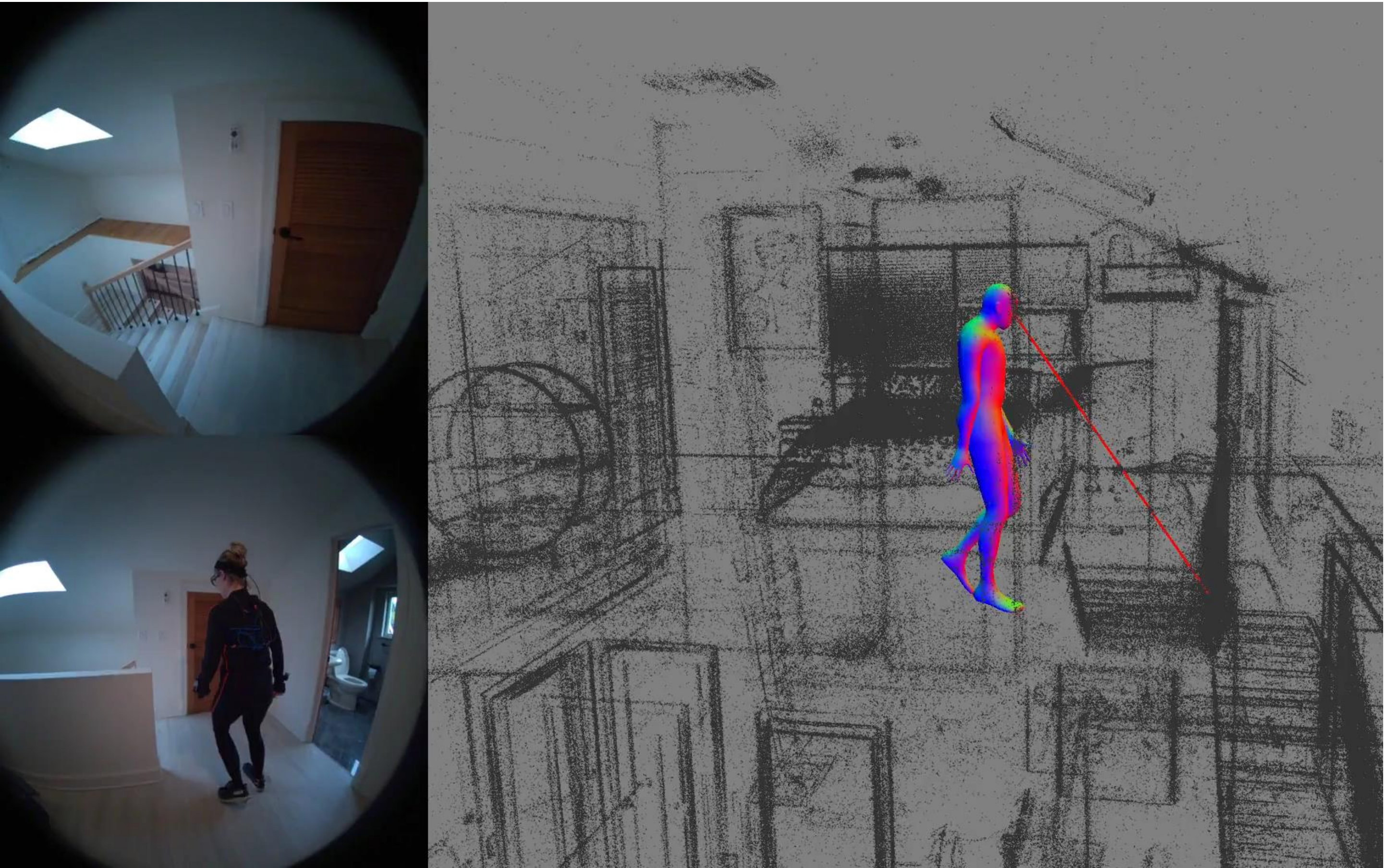
HOT3D

Hand / Object interaction.

projectaria.com/datasets/hot3d

* Released this week!

Nymeria



Massive Human Motion in the Wild

300h daily activities

264 participants

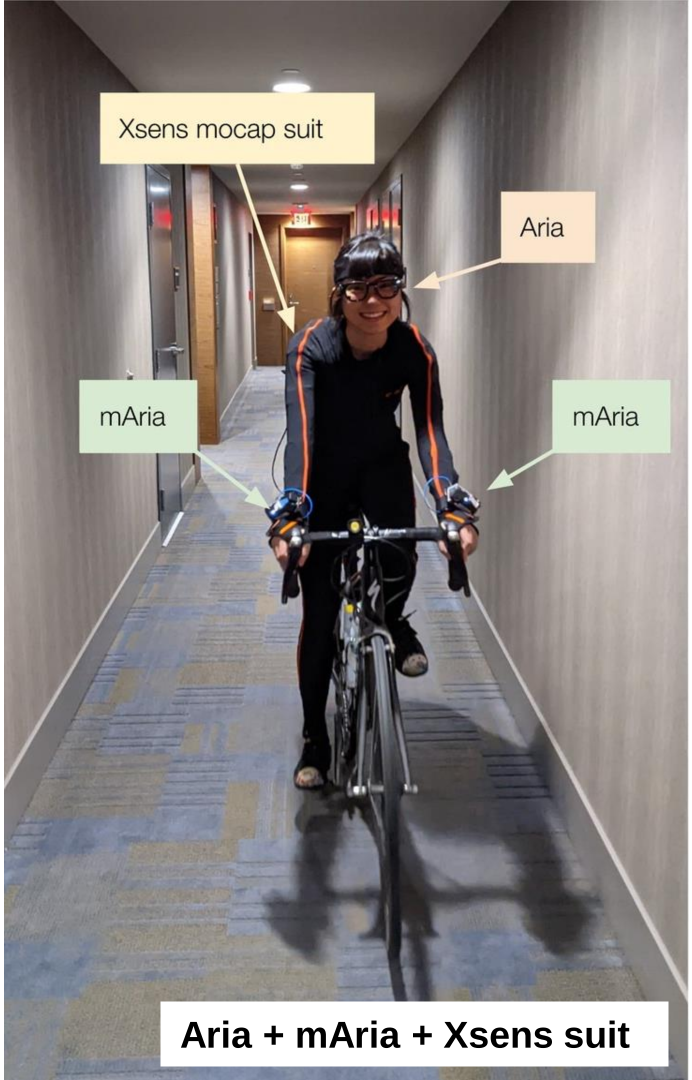
50 locations

3,600h video total

400km head motion

1053km wrist motion

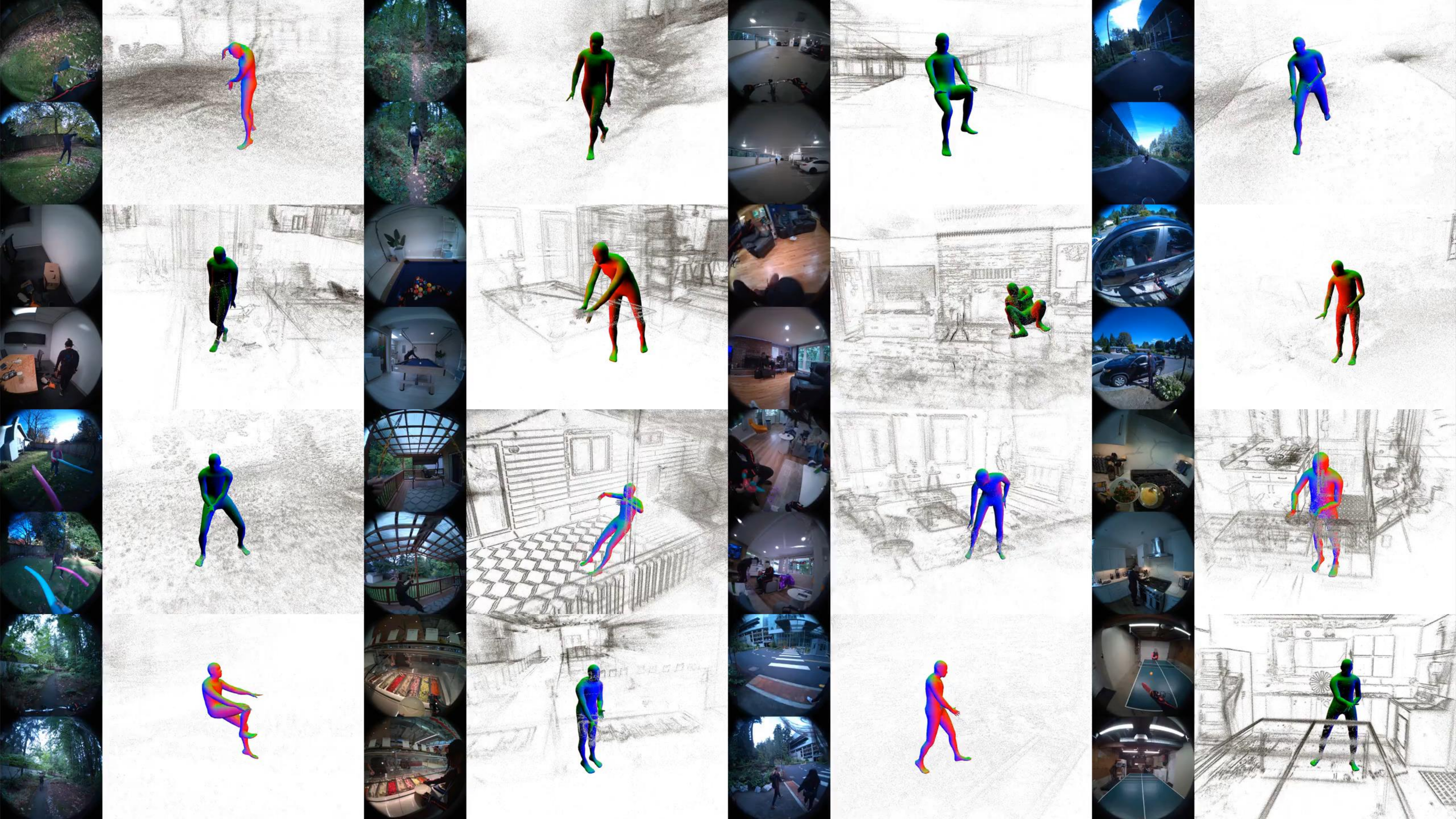
Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild; Ma et.al.; arXiv June 2024

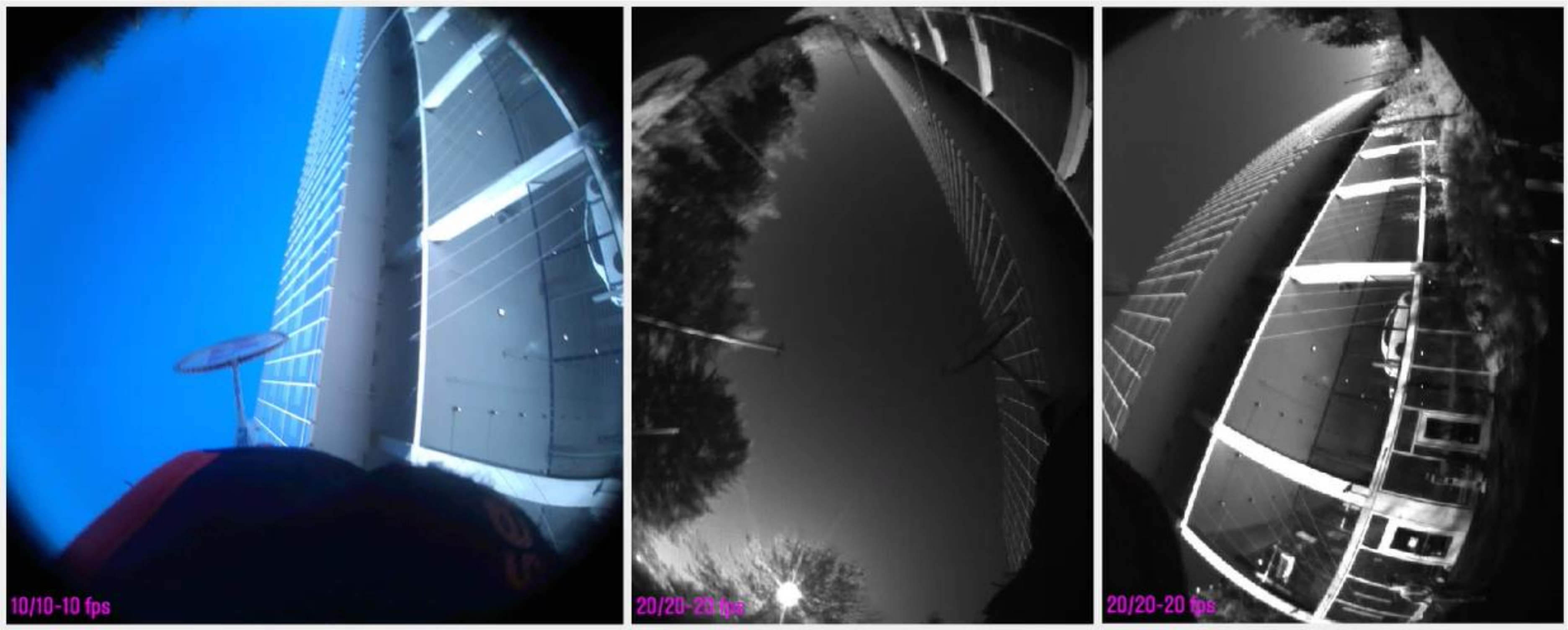


Aria + mAria + Xsens suit

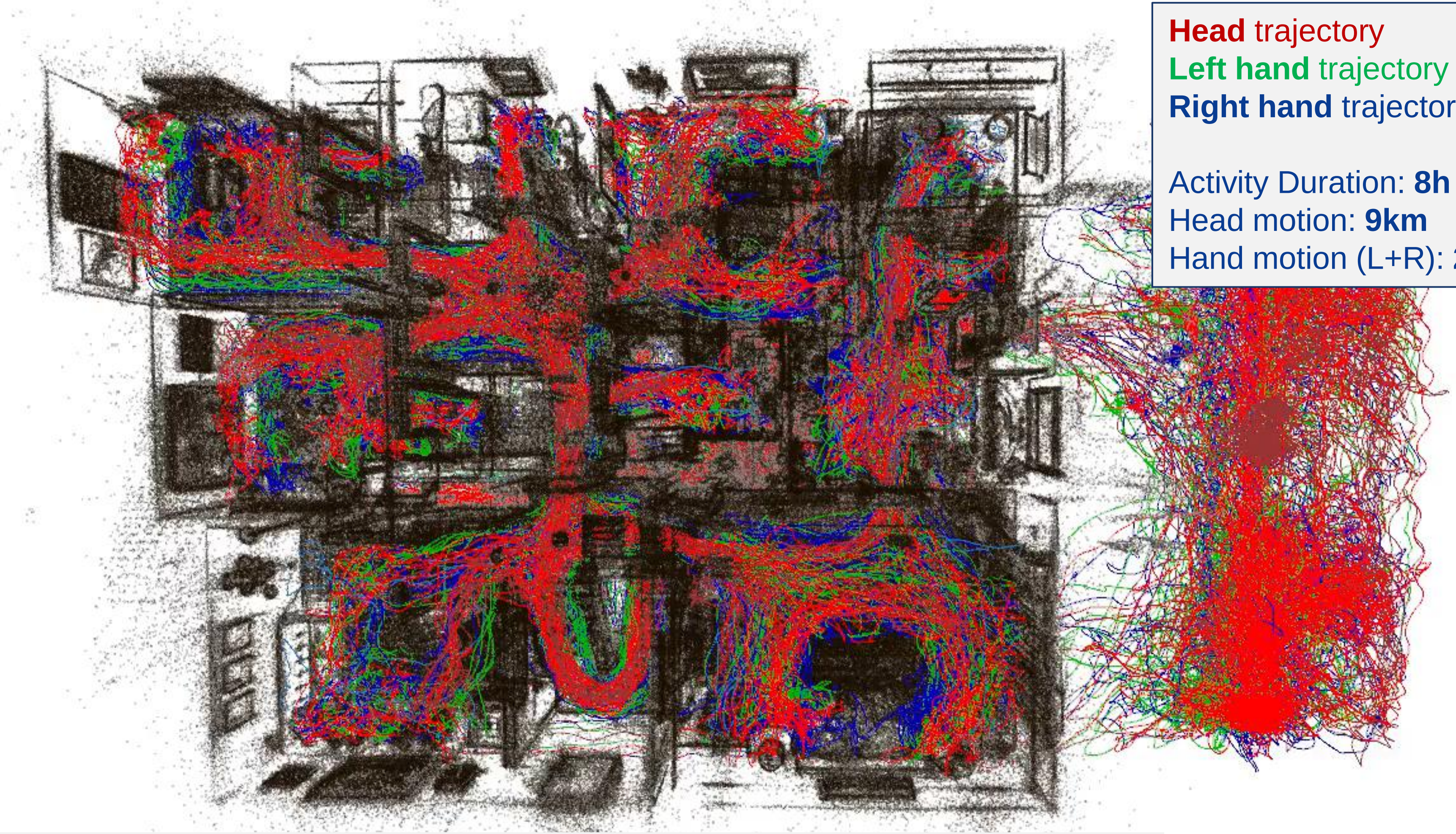
Aria

mAria





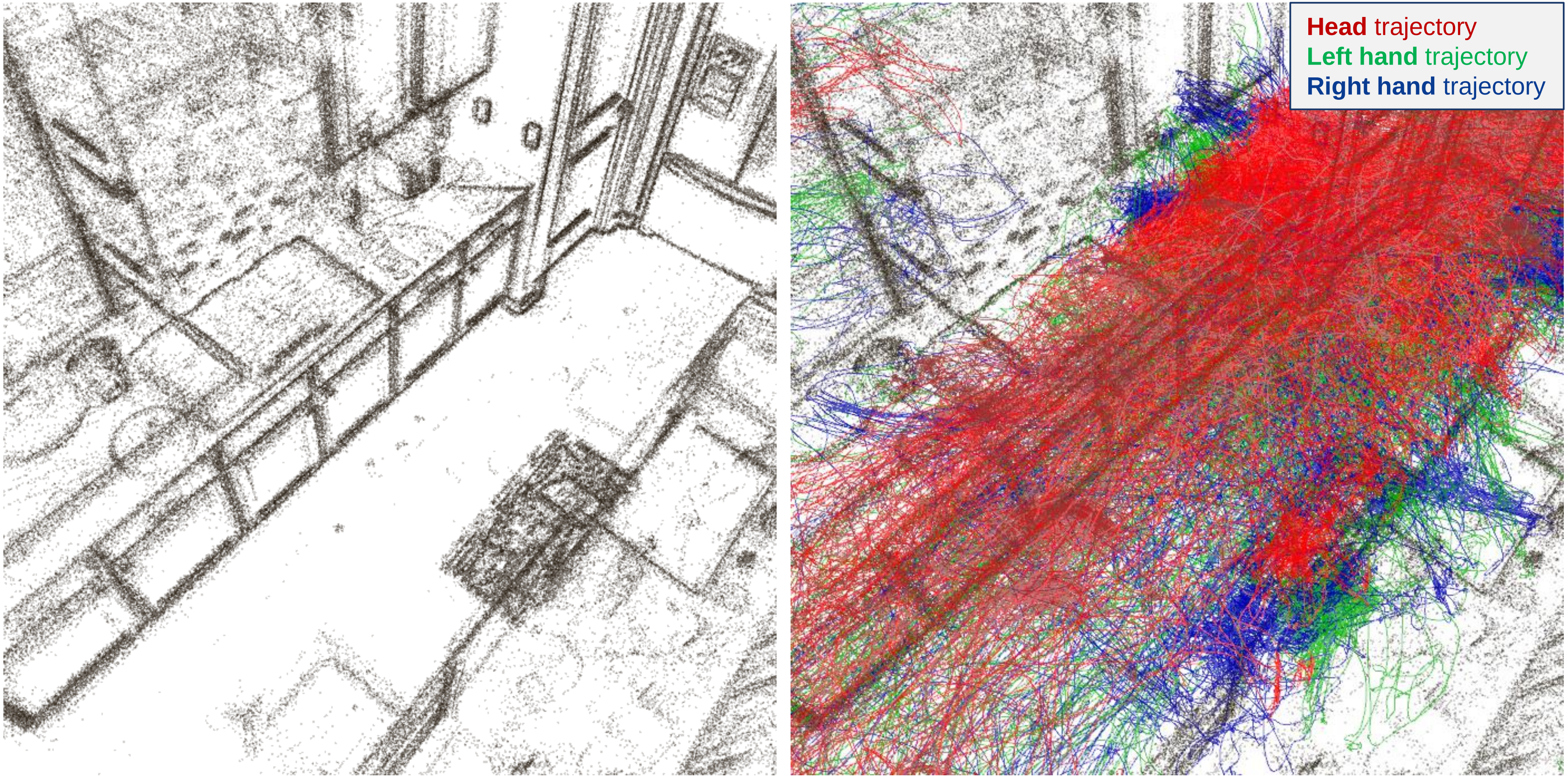
Playing Badminton (Right Wrist)



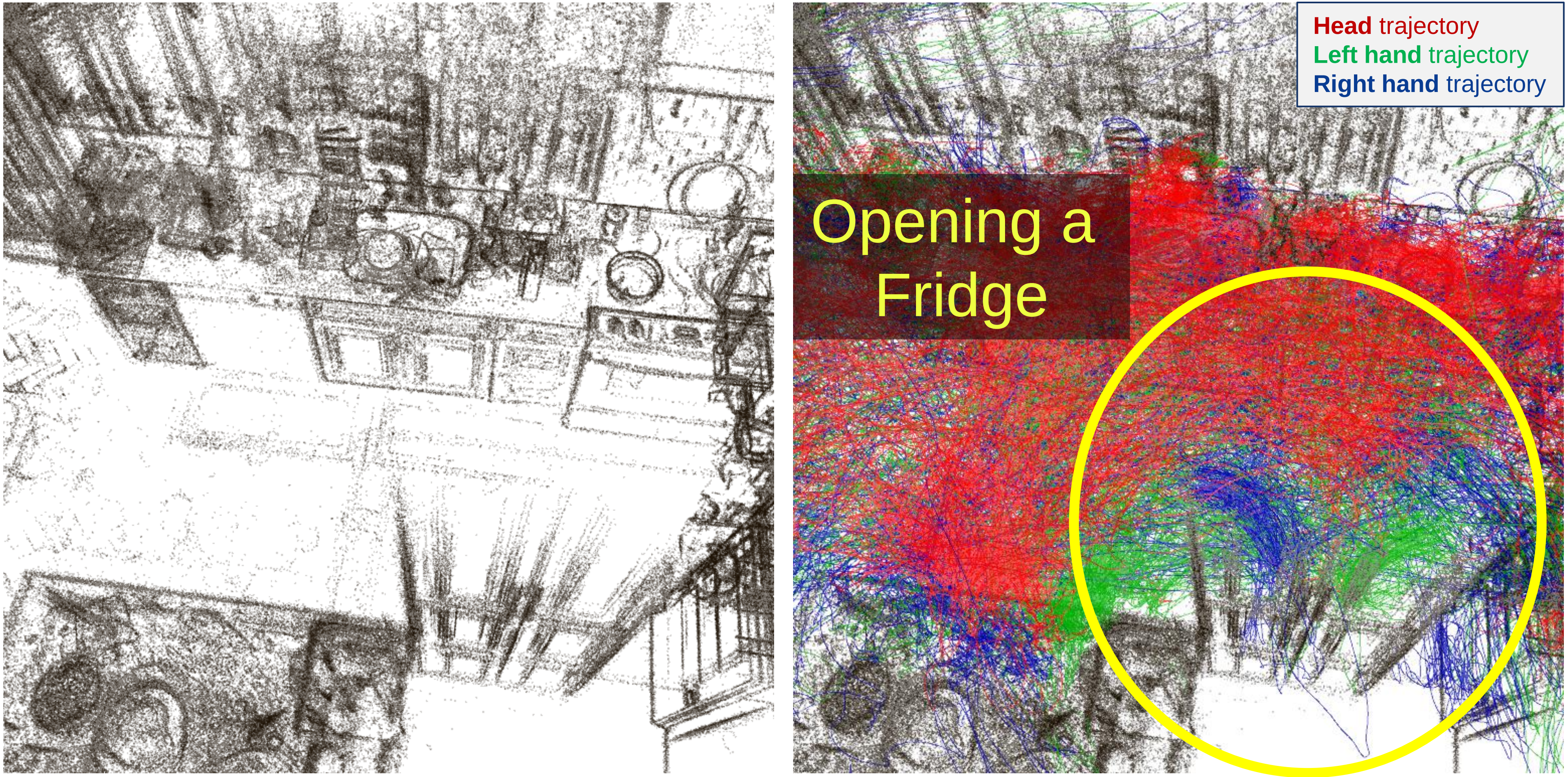
Head trajectory
Left hand trajectory
Right hand trajectory

Activity Duration: **8h**
Head motion: **9km**
Hand motion (L+R): **21km**

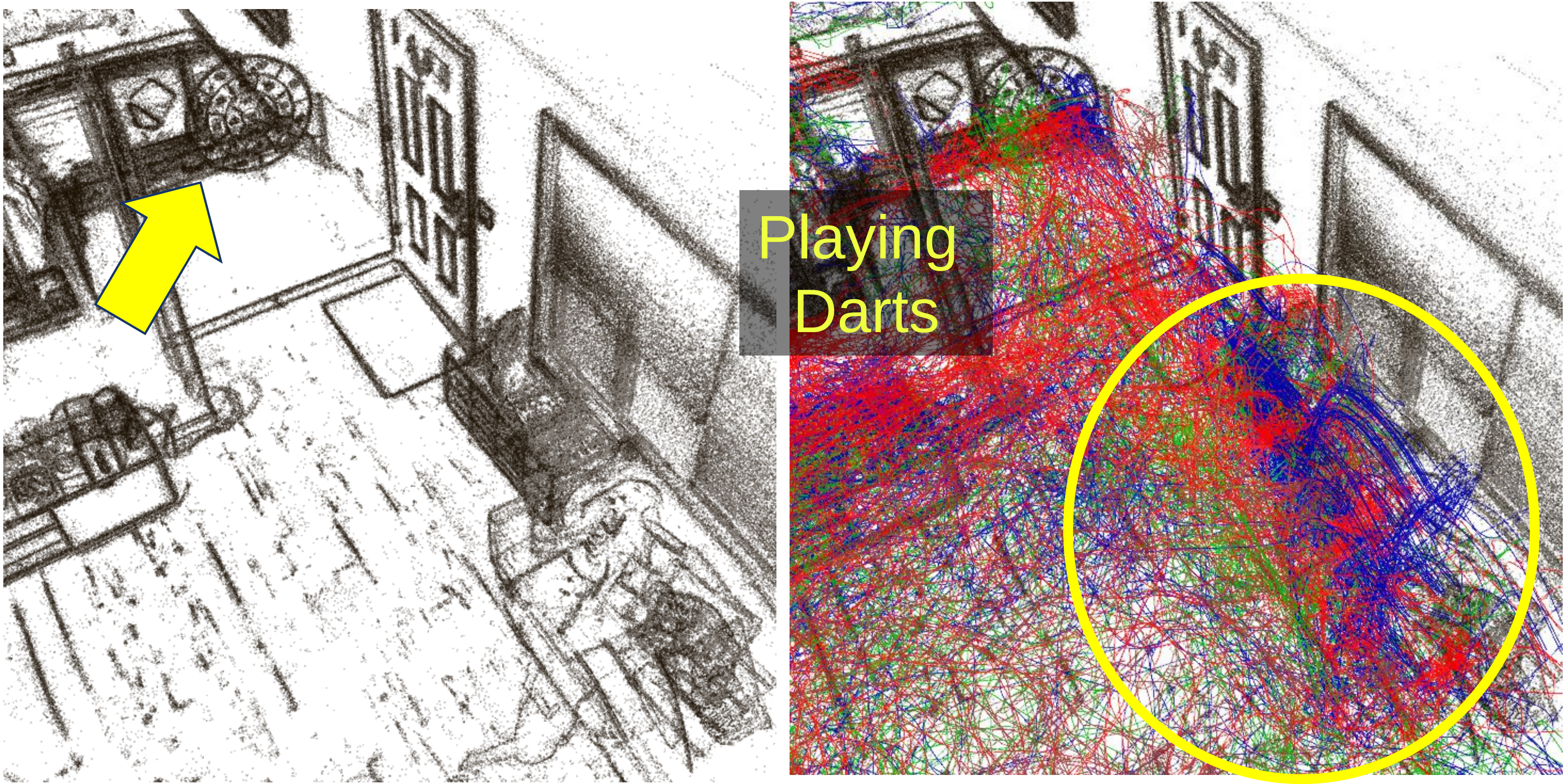
All recordings (in one Location, house35) aligned into the same frame of reference.


Head trajectory
Left hand trajectory
Right hand trajectory

Kitchen


Head trajectory
Left hand trajectory
Right hand trajectory
Opening a Fridge

Kitchen



Living Room

New Aria-based Datasets

Bringing Egocentric Machine Perception into 3D

Ego-Exo 4D

Skilled human activity understanding.

ego-exo4d-data.org

* CVPR 2024 paper!

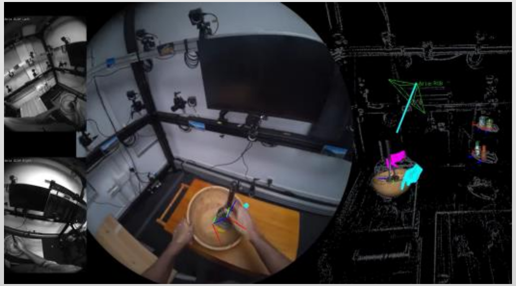
Nymeria

Human Motion in the wild.

projectaria.com/datasets/nymeria

* Released this week!

HOT3D

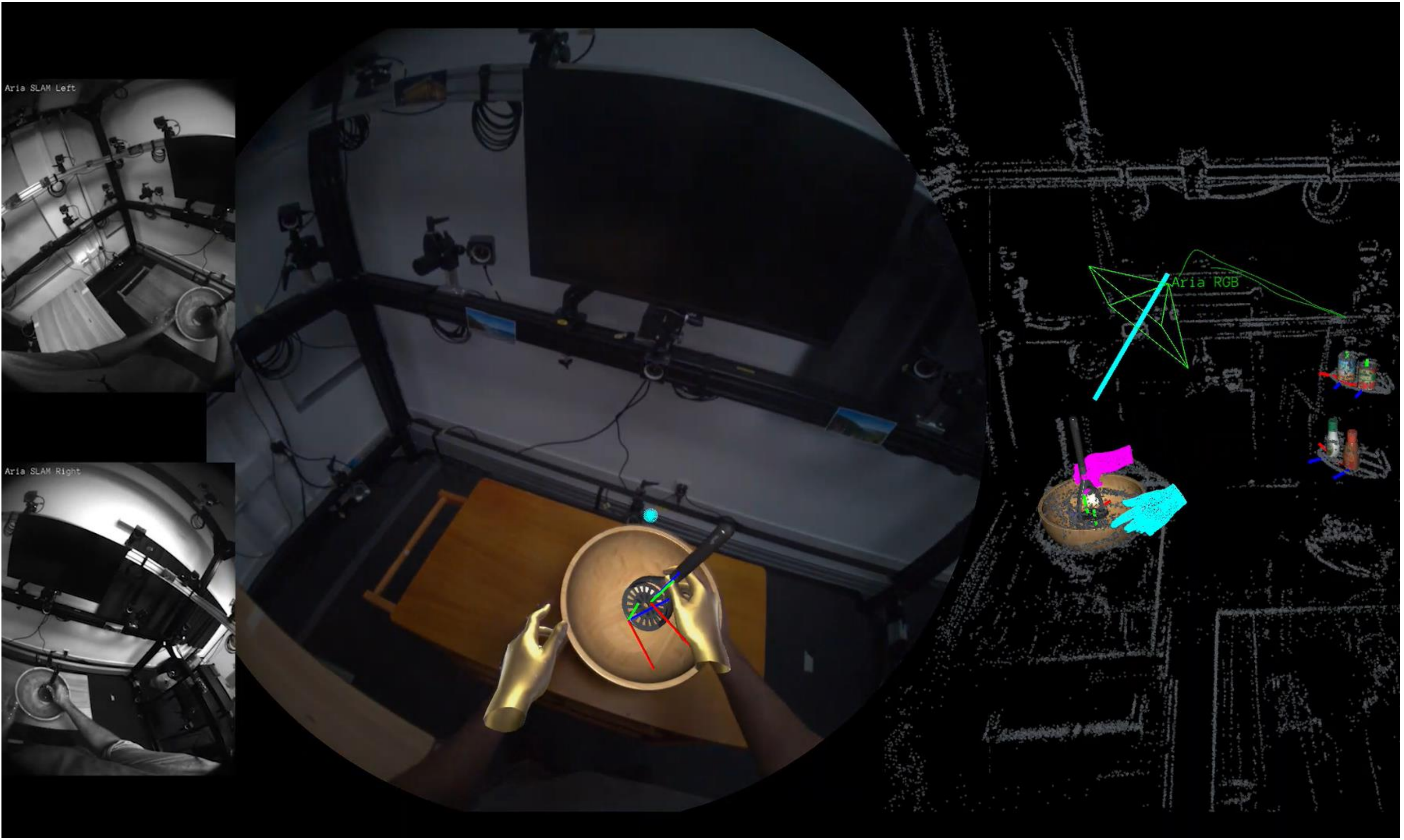


Hand / Object interaction.

projectaria.com/datasets/hot3d

* Released this week!

HOT-3D



A dataset and benchmark for egocentric hand and object tracking.

833 min ego-video

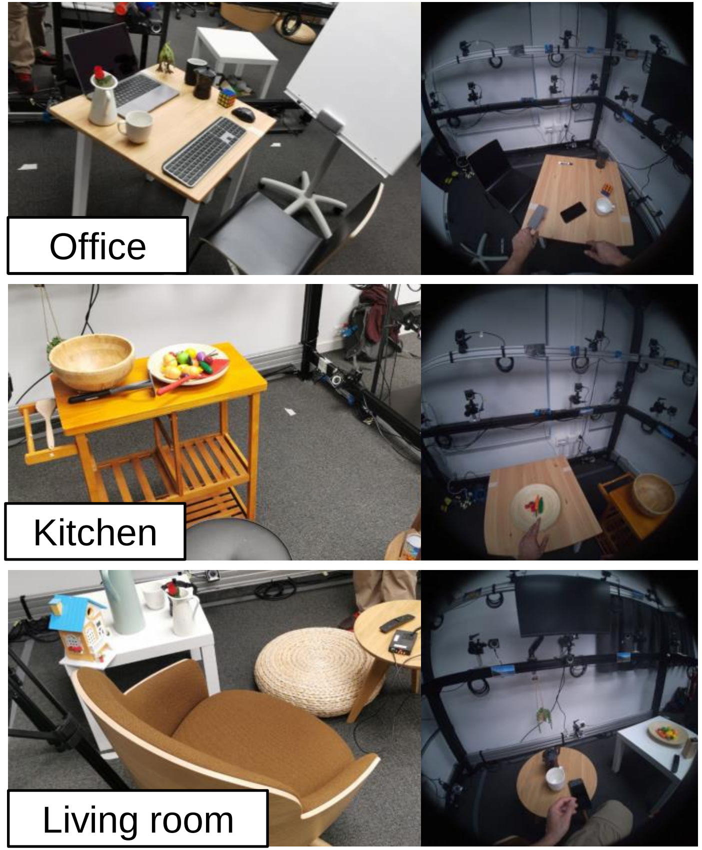
33 distinct objects

19 participants

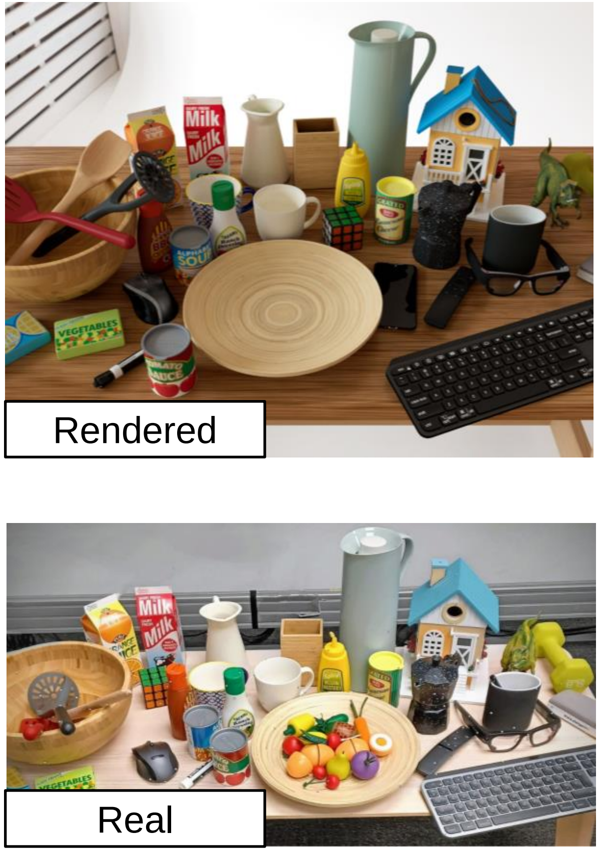
3 environments

13 km object motion

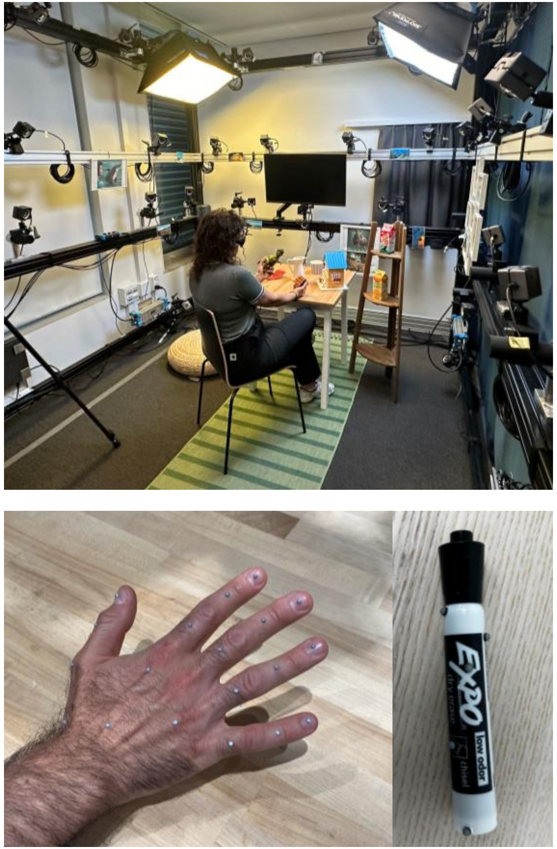
Introducing HOT3D: An Egocentric Dataset for 3D Hand and Object Tracking; Banerjee et.al.; ArXiv June 2024



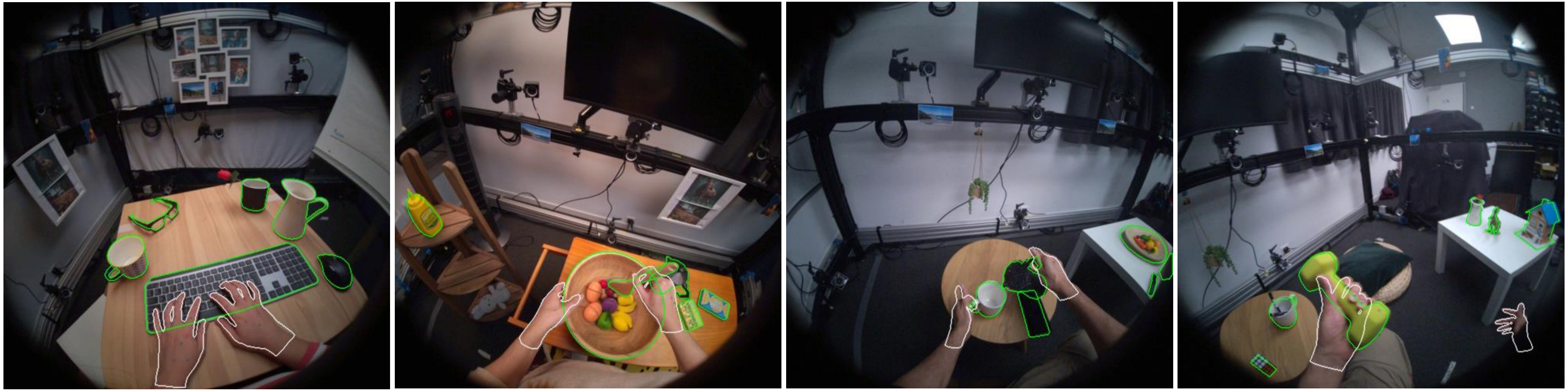
3 Different Scenarios



33 Distinct Objects



High-quality Hand- and Object Poses



2024 BOP Object Tracking Challenge

Challenge Tracks:

- Object detection and pose estimation
- Model-based & model-free
- Seen & unseen objects

Results at ECCV 2024

More at bop.felk.cvut.cz

2024 Hand Tracking Challenge

Challenge Tracks:

- Hand pose estimation with known hand shapes
- Hand Shape Estimation with MANO [1] models

Organized as part of HANDS workshop at ECCV 2024

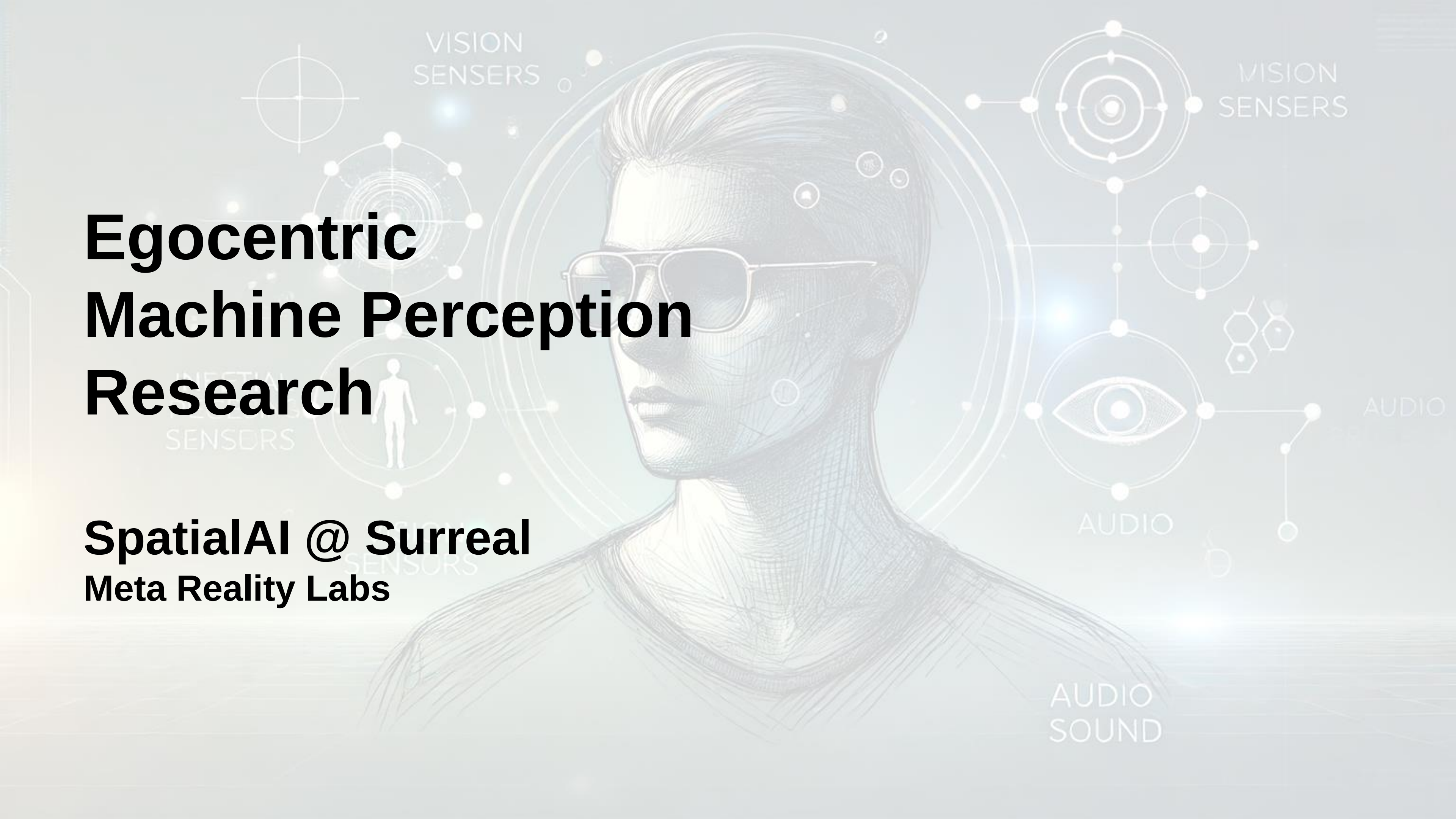
More at github.com/facebookresearch/hand_tracking_toolkit

Joint dataset & test-frames – towards joint, egocentric, hand and object tracking.

Egocentric Machine Perception Research

SpatialAI @ Surreal

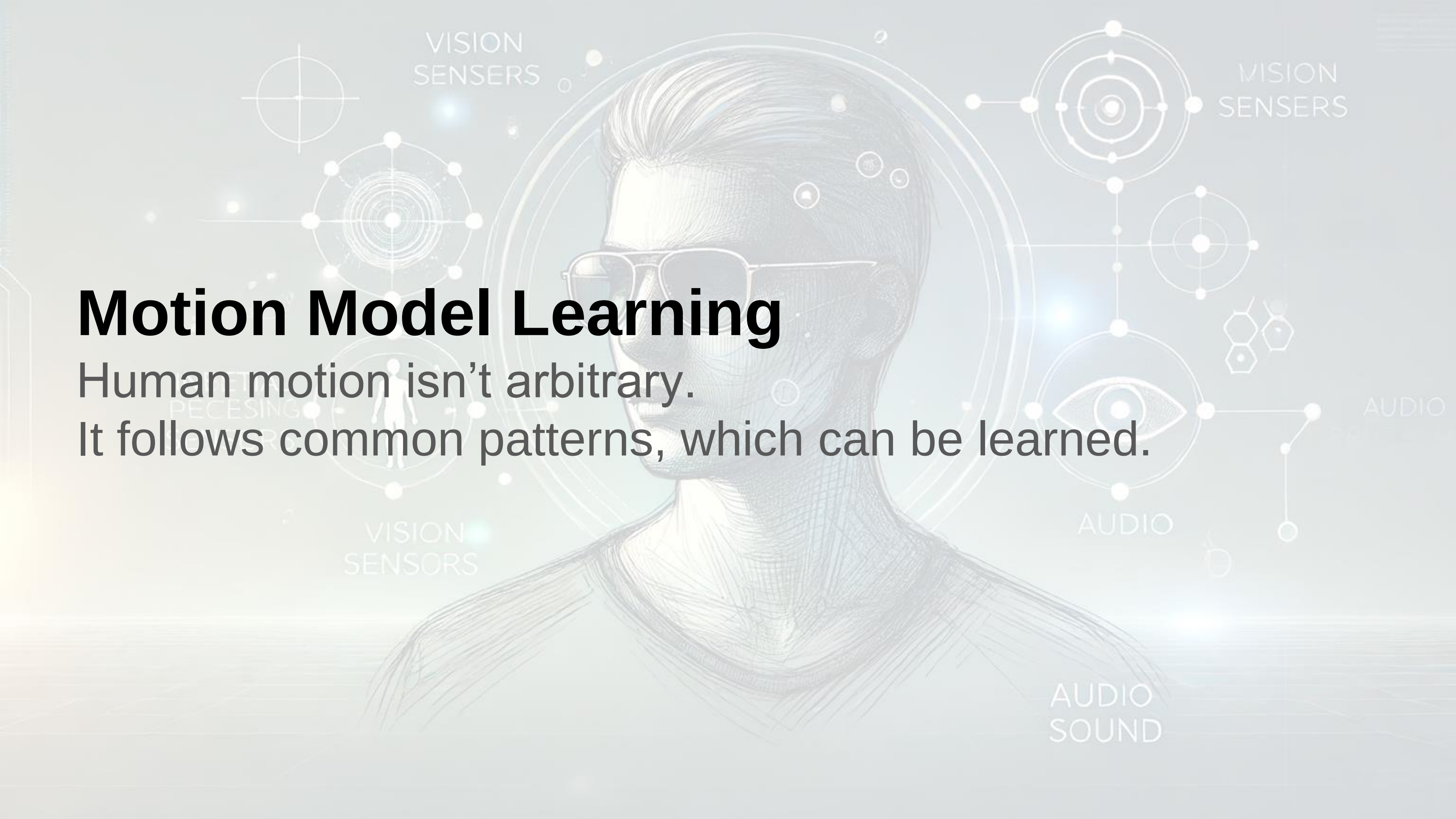
Meta Reality Labs

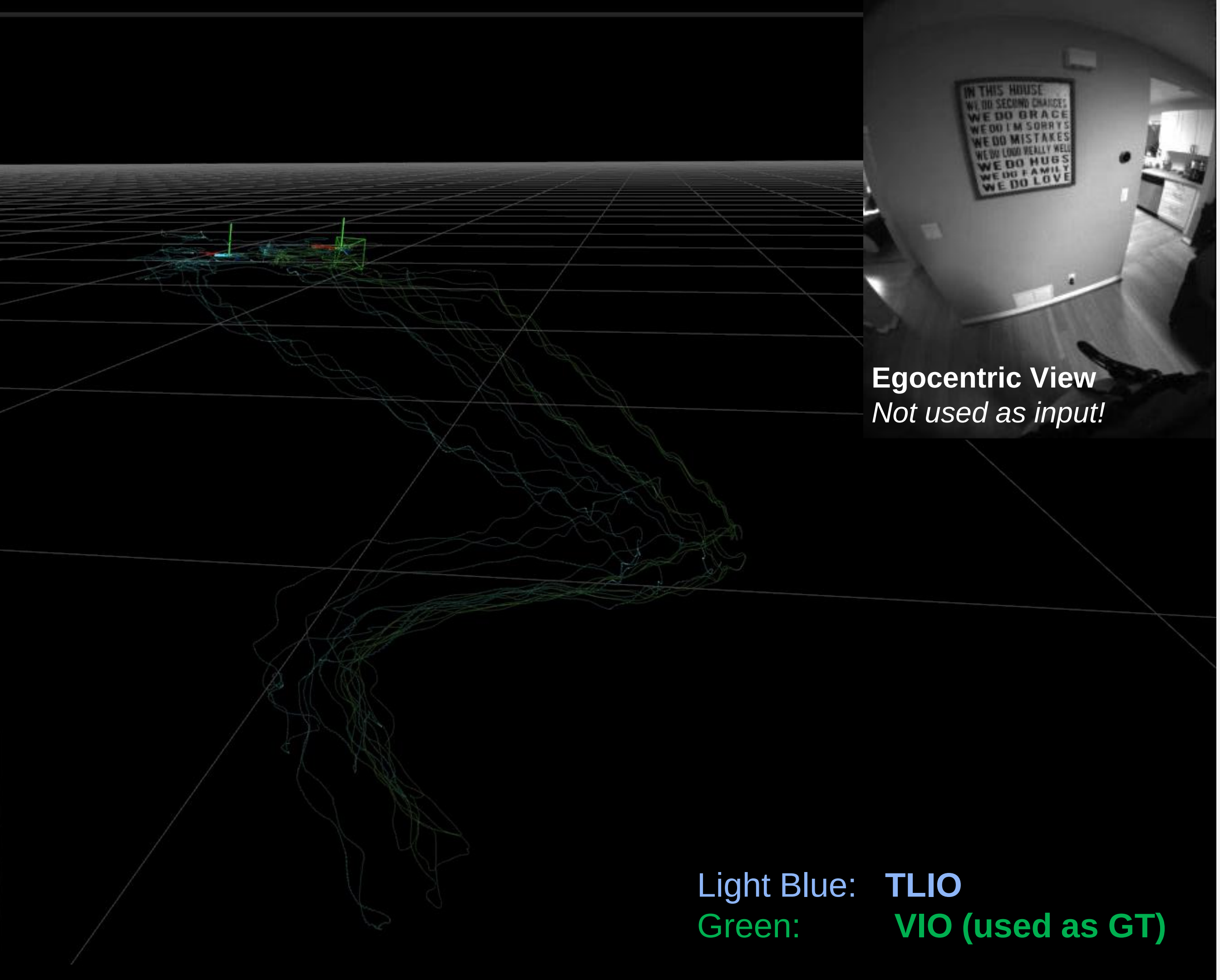


Motion Model Learning

Human motion isn’t arbitrary.

It follows common patterns, which can be learned.





Egocentric View
Not used as input!

Light Blue: **TLIO**
Green: **VIO (used as GT)**

TLIO:

Tight Learned Inertial Odometry

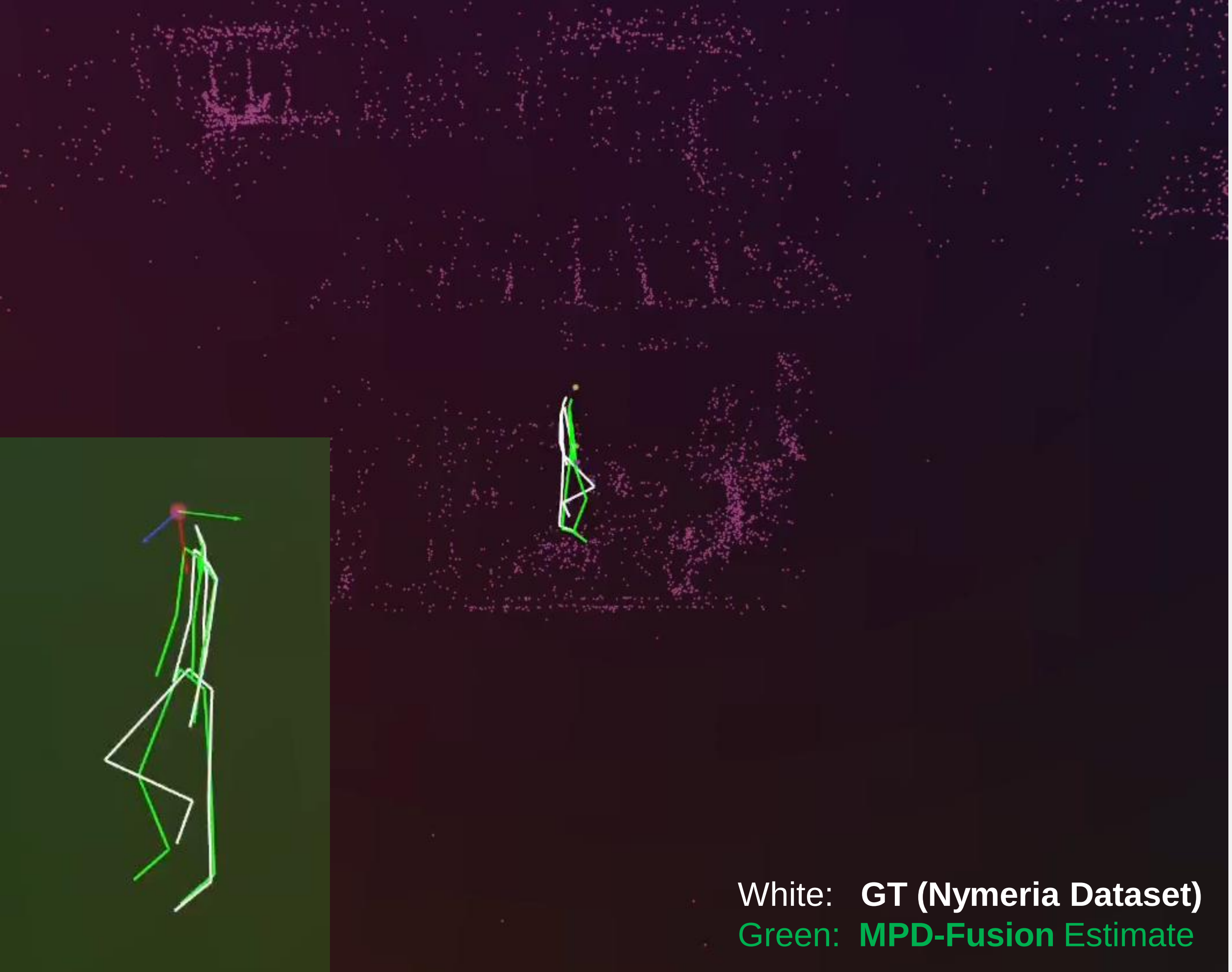
Input:

- IMU data from HMD
- No Vision.

Output:

- Odometry Trajectory

Liu et.al, Robotics and Automation Letters, June 2020



White: **GT (Nymeria Dataset)**

Green: **MPD-Fusion** Estimate

MPD-Fusion

Motion Prior Diffusion

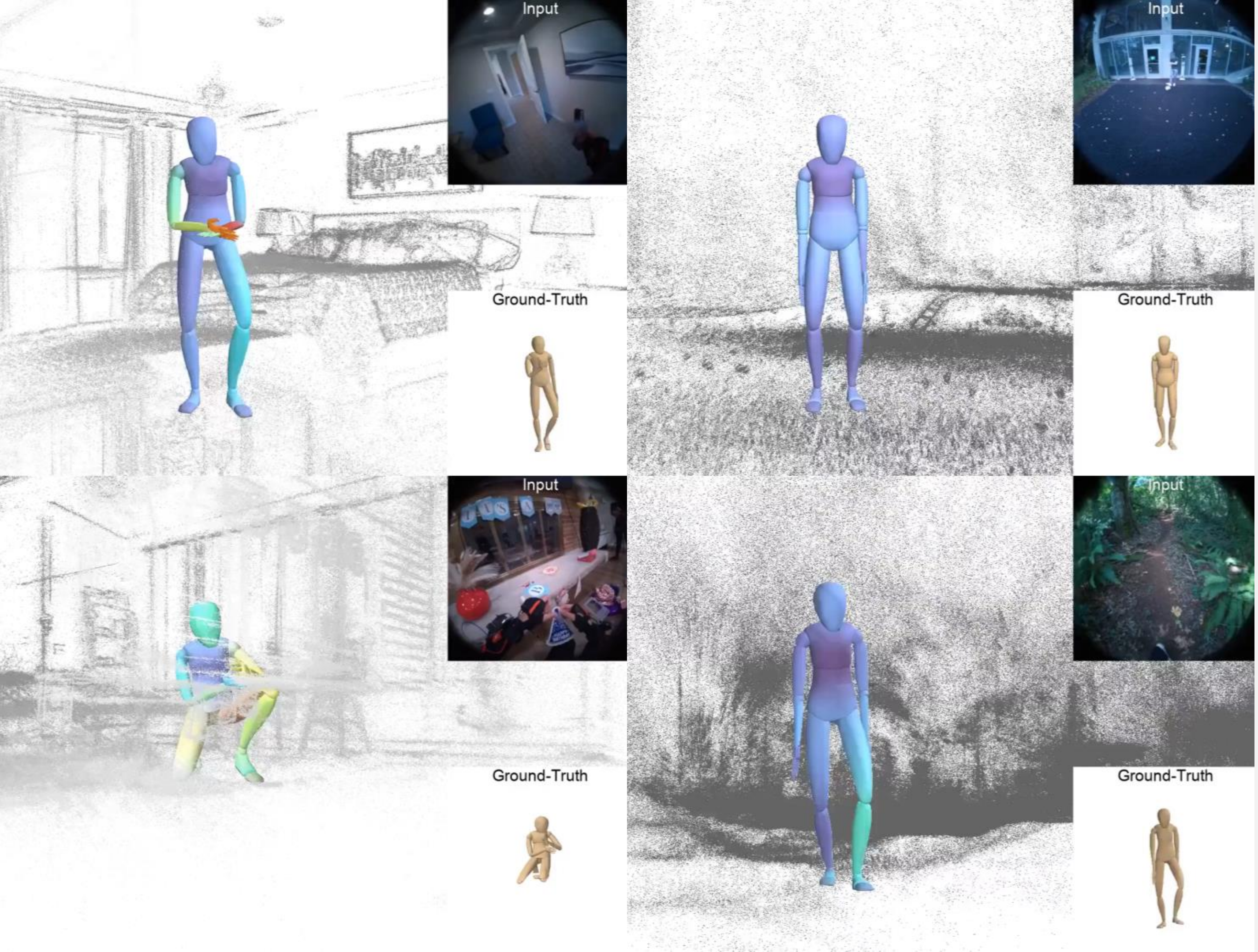
Input:

- IMU data from HMD and left/right wristband
- No vision **at all**

Output:

- Skeleton Pose
- Odometry Trajectory

(under review)



HMD^2: Environment-aware Motion Generation from an HMD

Input:

- 6DoF Aria trajectory
- RGB images
- Point clouds

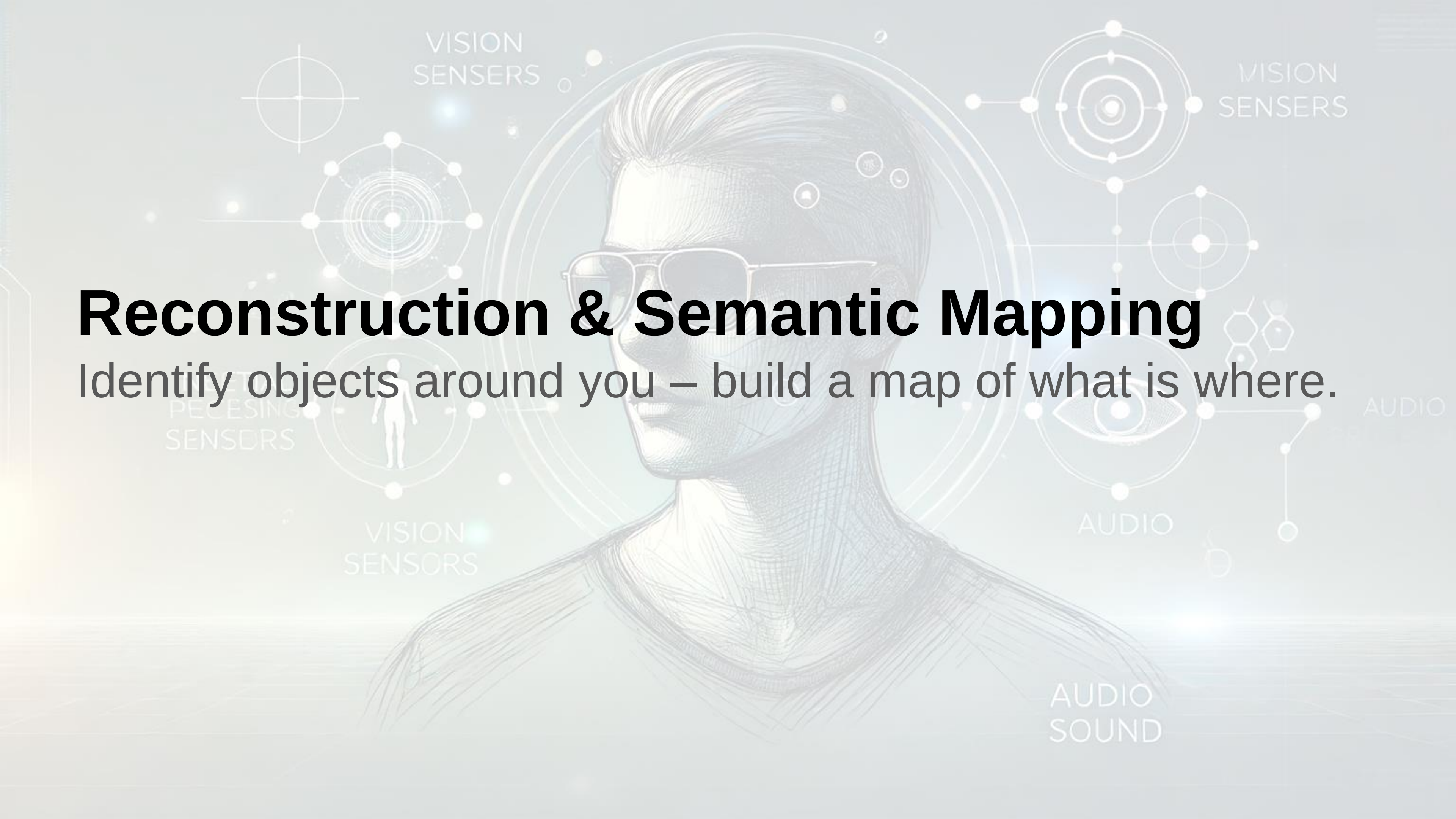
Output:

- Skeleton Pose & Orientation

(under review)

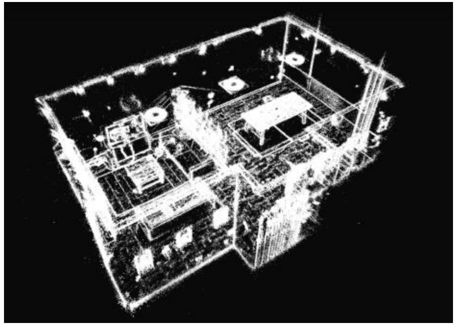
Reconstruction & Semantic Mapping

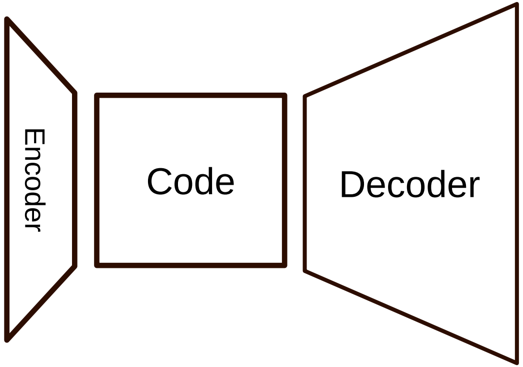
Identify objects around you – build a map of what is where.

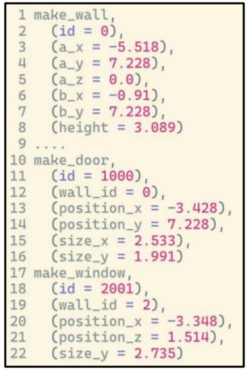


SceneScript

Tokenizing a scene using Language

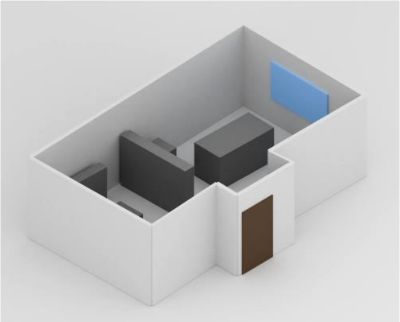






```
make_wall,
  (id = 0),
  (a_x = -5.518),
  (a_y = 7.228),
  (a_z = 0.0),
  (b_x = -0.91),
  (b_y = 7.228),
  (height = 3.089)
....
make_door,
  (id = 1000),
  (wall_id = 0),
  (position_x = -3.428),
  (position_y = 7.228),
  (size_x = 2.533),
  (size_y = 1.991)
make_window,
  (id = 2001),
  (wall_id = 2),
  (position_x = -3.348),
  (position_z = 1.514),
  (size_y = 2.735)
```

=



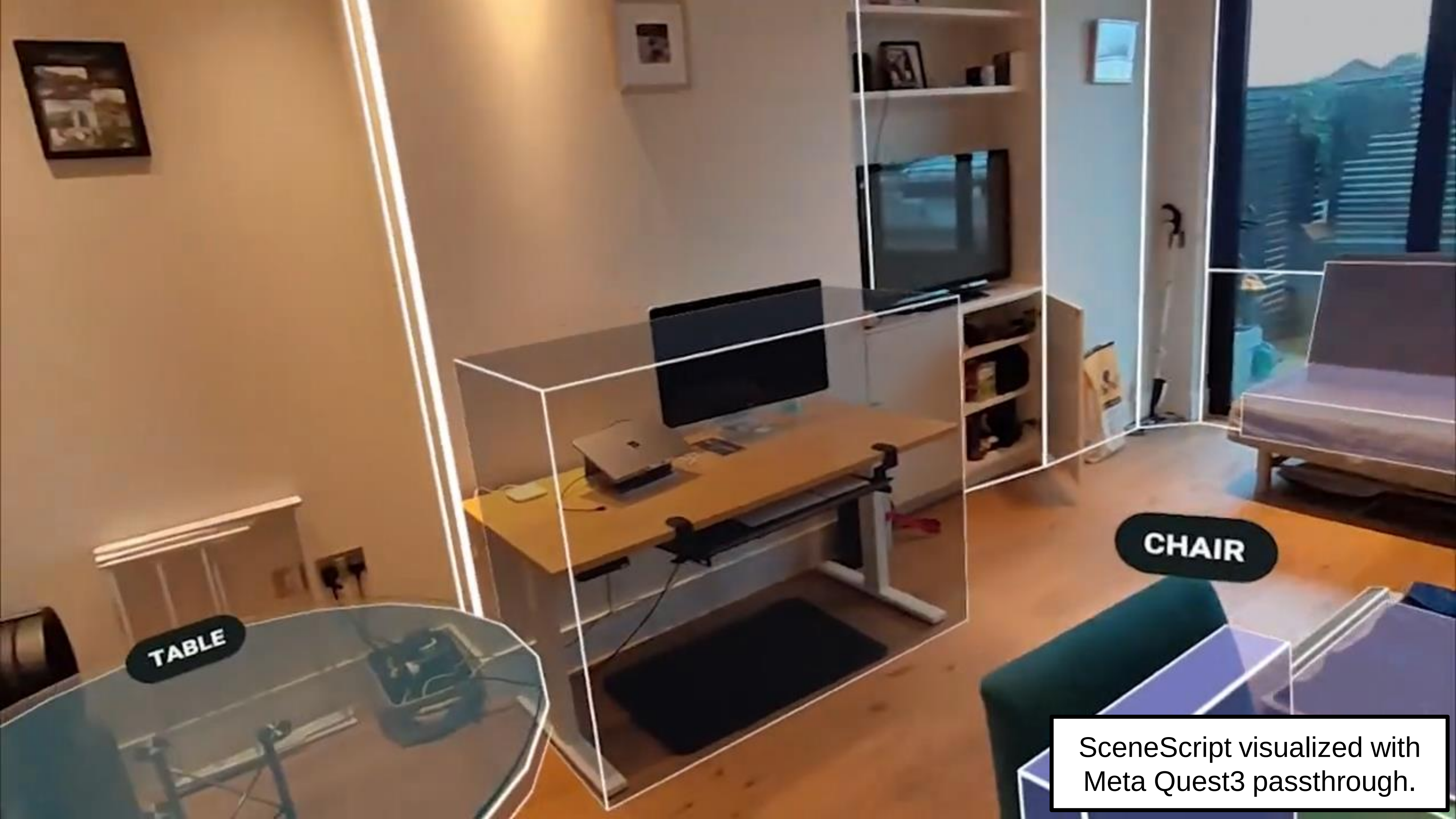
Input Scene:
SLAM Pointcloud;
TSDF volume; etc

Autoregressive Transformer

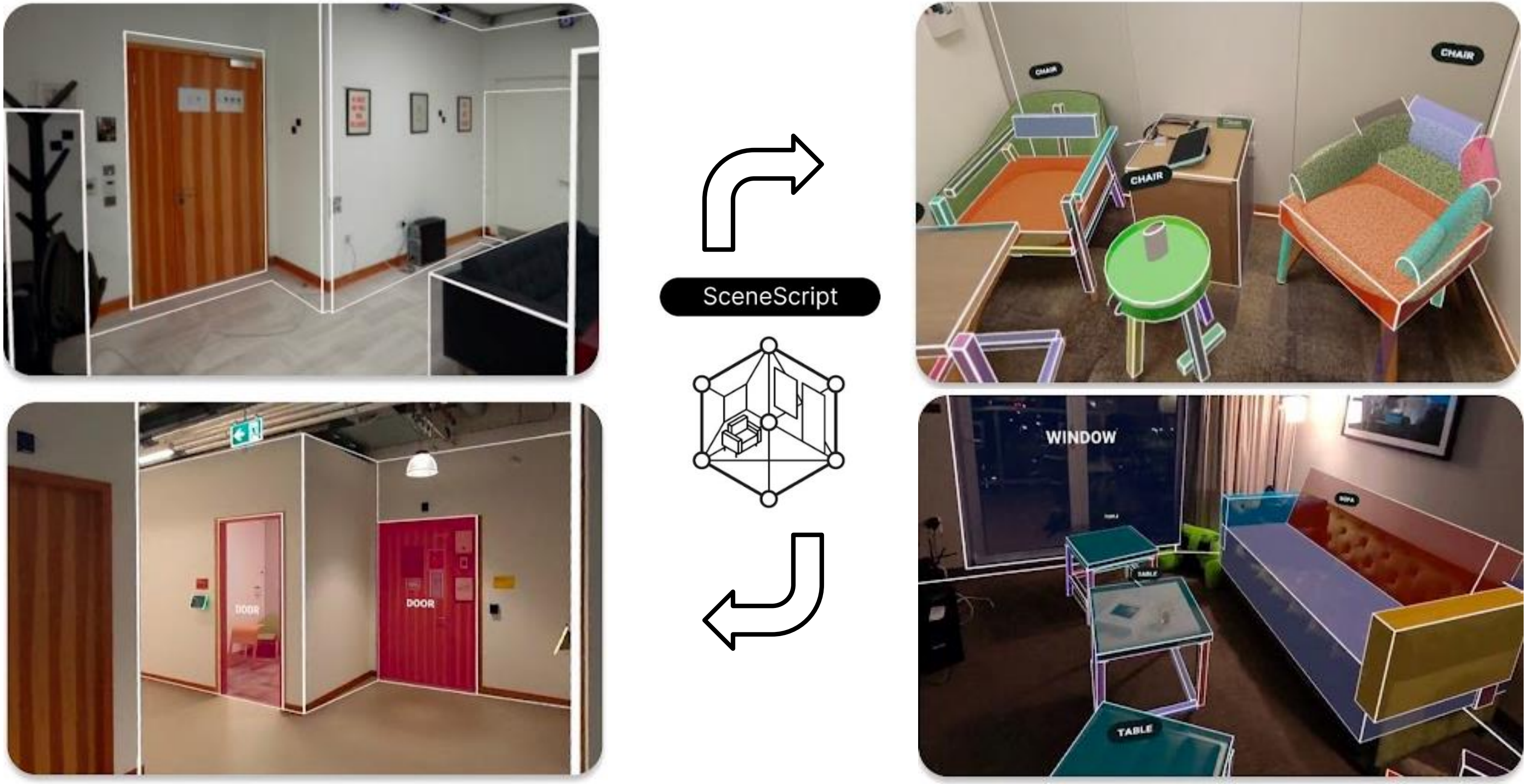
SceneScript:
commands to
generate scene

Tokenized Scene

SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model; Avetisyan et.al., ArXiv March 2024



SceneScript visualized with Meta Quest3 passthrough.



Rooms => walls, doors, windows, etc.

Objects => cuboids, cylinders, etc.

Aria Synthetic Environments

Dataset

100k unique apartments (procedurally generated)

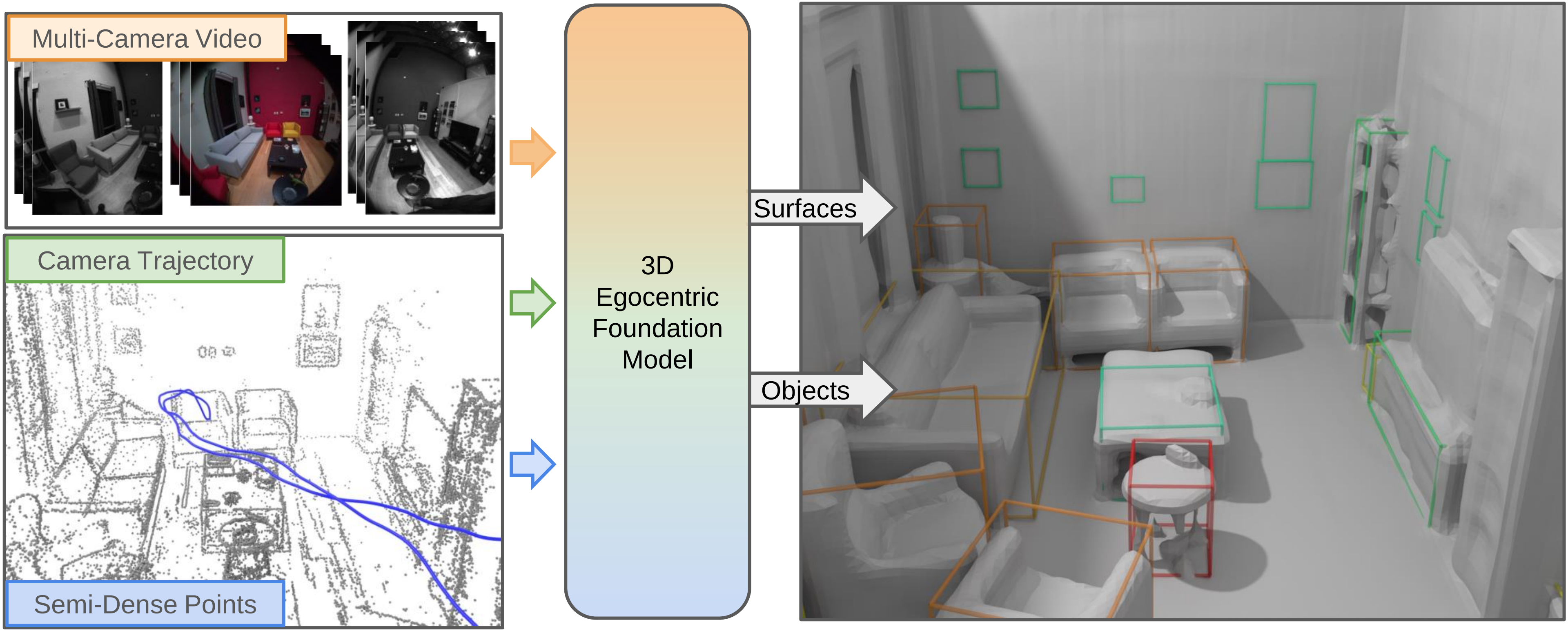
Dataset contains

- GT scene language commands
- 2 minutes simulated walk-through recording per scene
- RGB, depth, segmentation, point-clouds, etc.

projectaria.com/datasets/ase

Egocentric Foundation Models

Foundational Features *in 3D*



Multi-Camera Video

Camera Trajectory

Semi-Dense Points

3D Egocentric Foundation Model

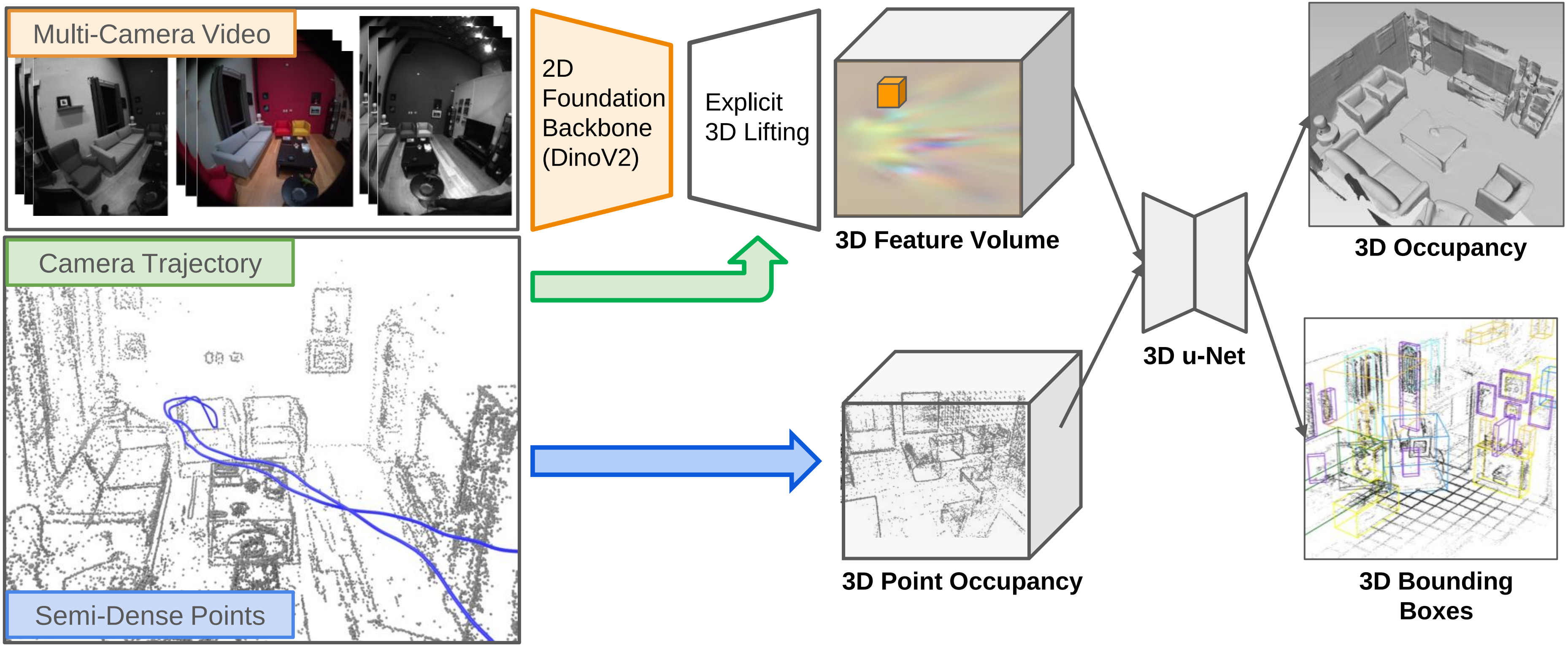
Surfaces

Objects

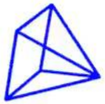
EFM3D: A Benchmark for measuring Progress Towards 3D Egocentric Foundation Models; Straub et.al.; ArXiv June 24

Egocentric Voxel Lifting

Foundational Features *in 3D*

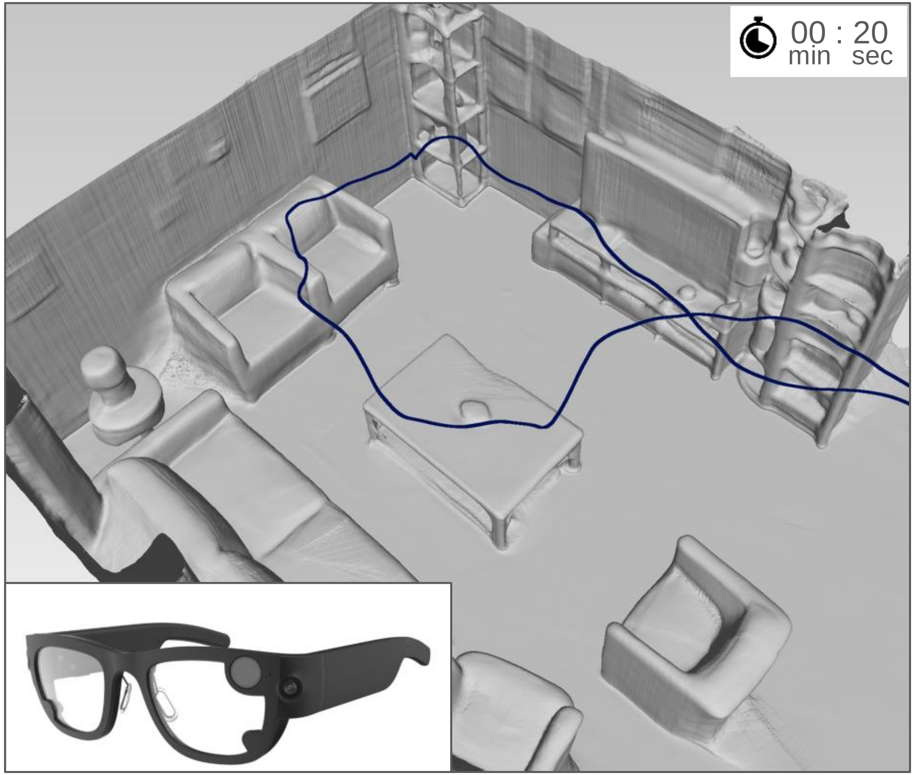


EFM3D: A Benchmark for measuring Progress Towards 3D Egocentric Foundation Models; Straub et.al.; ArXiv June 24



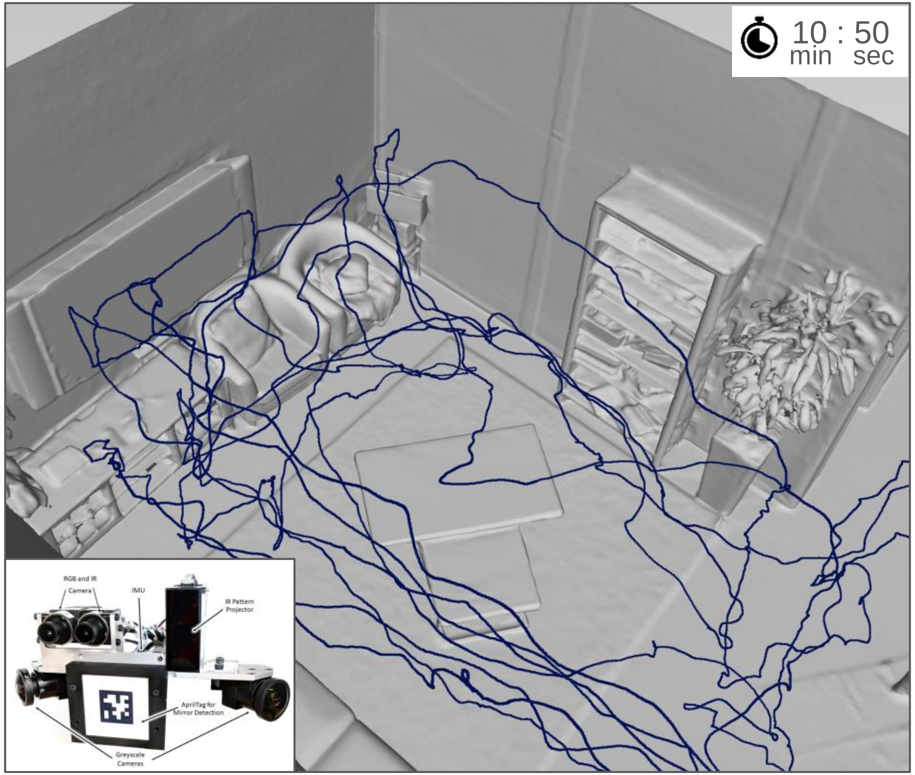
EVL + TSDF Fusion + 3DBB filtering

EVL

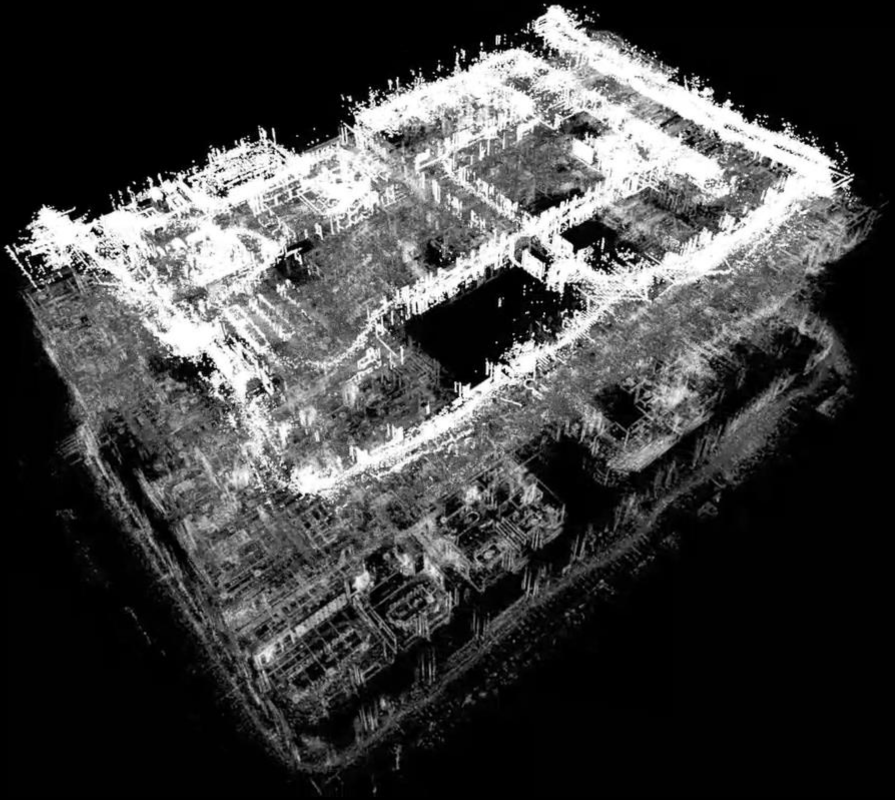


Aria + Casual Motion + EVL

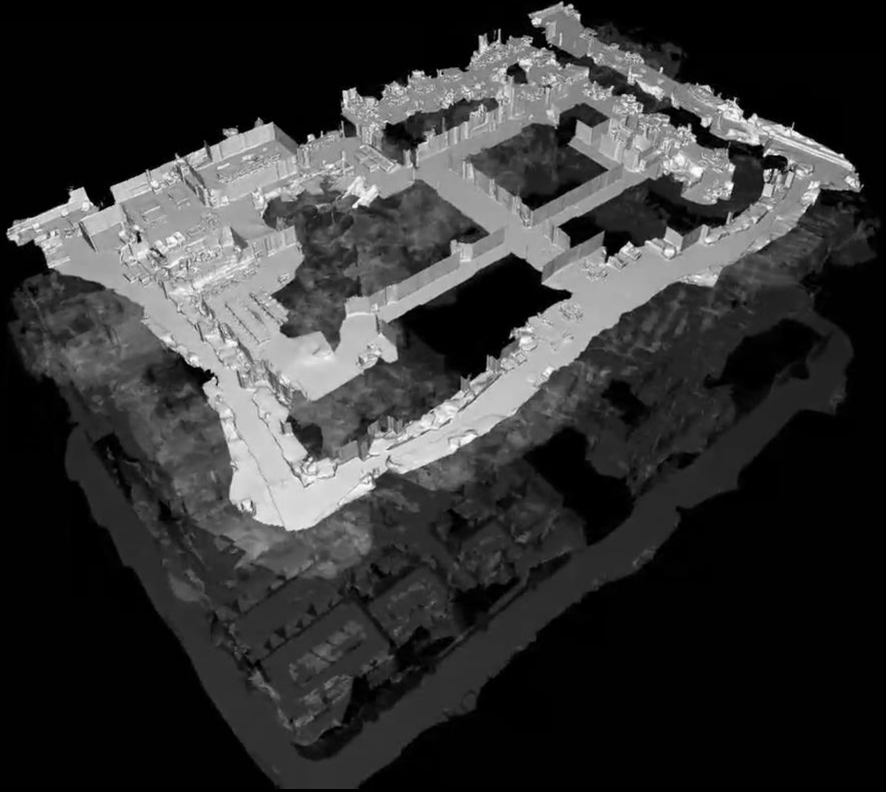
Replica



Camera + Depth Sensor +
Dedicated Scanning Motion



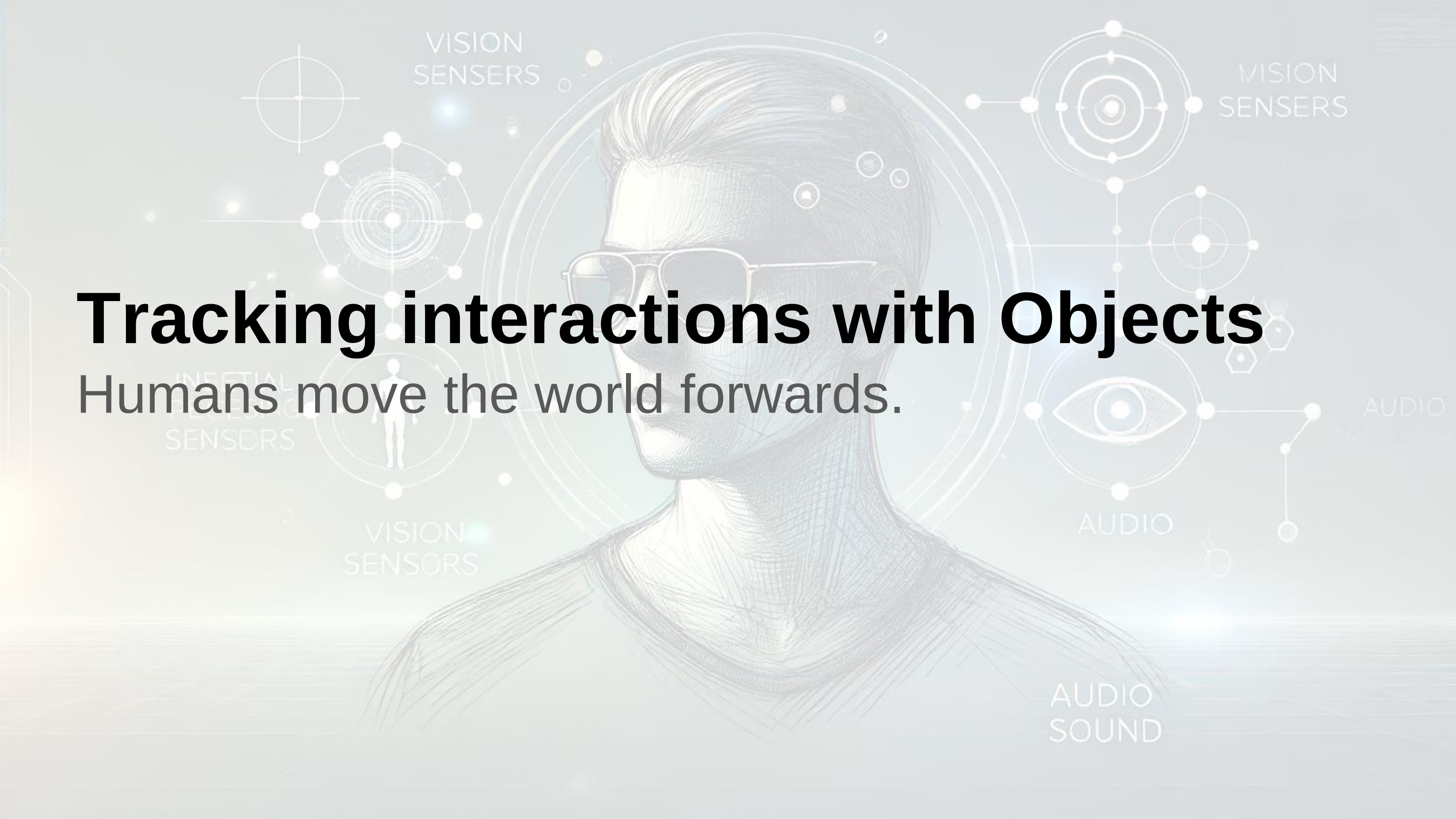
Semi-Dense Points (from Aria MPS)

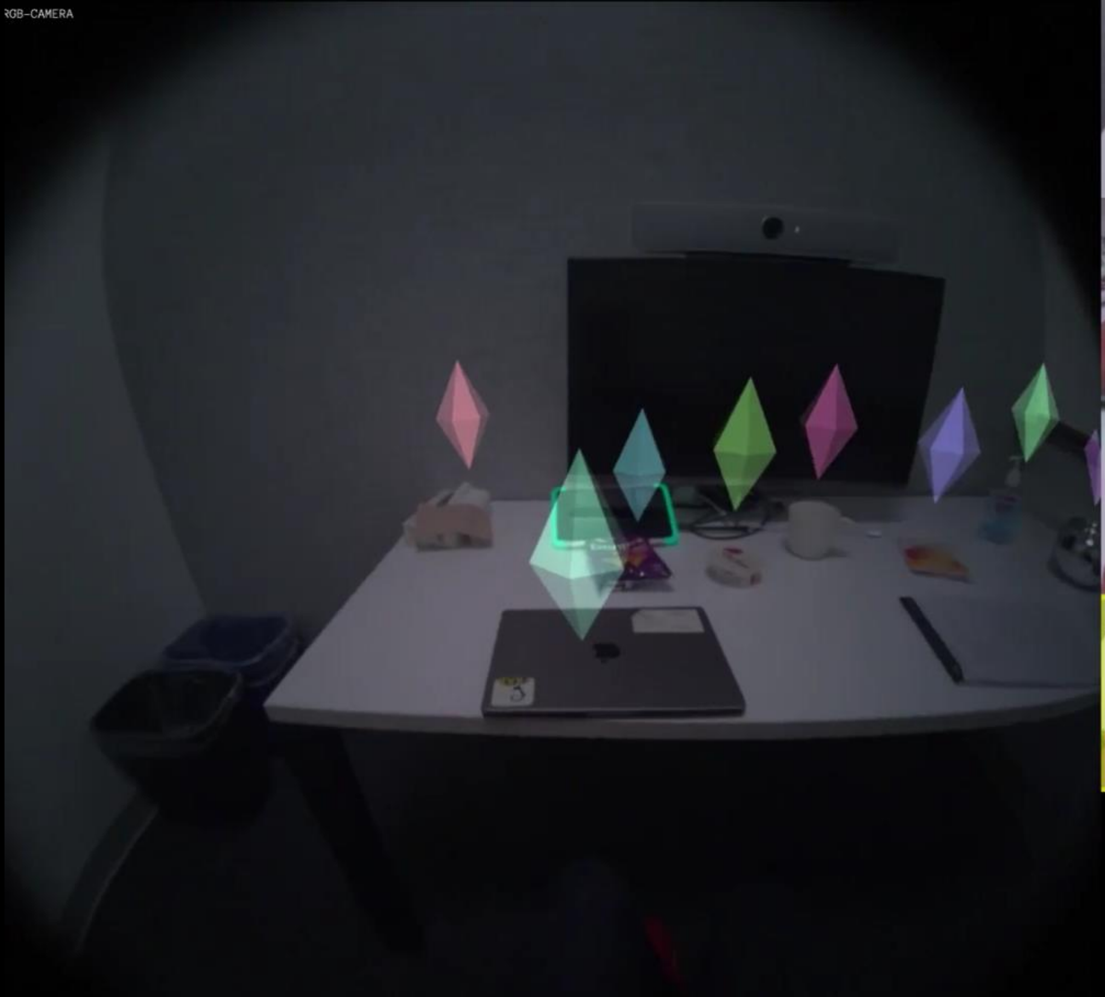


Fused EVL Reconstruction

Tracking interactions with Objects

Humans move the world forwards.





Egocentric View

Object Library

(ongoing work)

Questions