Egocentric Open-World 3D Scene Understanding

JULIAN STRAUB

ICCV 2025 – 5th Workshop on Open-World 3D Scene Understanding with Foundation Models

Reality Labs Research (RL-R)

Introducing the first AI glasses with a private in-lens display and on-wrist control

Project Aria Gen 2

www.projectaria.com

Privacy Switch and Volume control

7x Spatial mics

Ambient Light Sensor

2x IMU, Mag, Barometer

USB-C Port

LED

GNSS

4x CV Cameras

12 MP RGB Camera

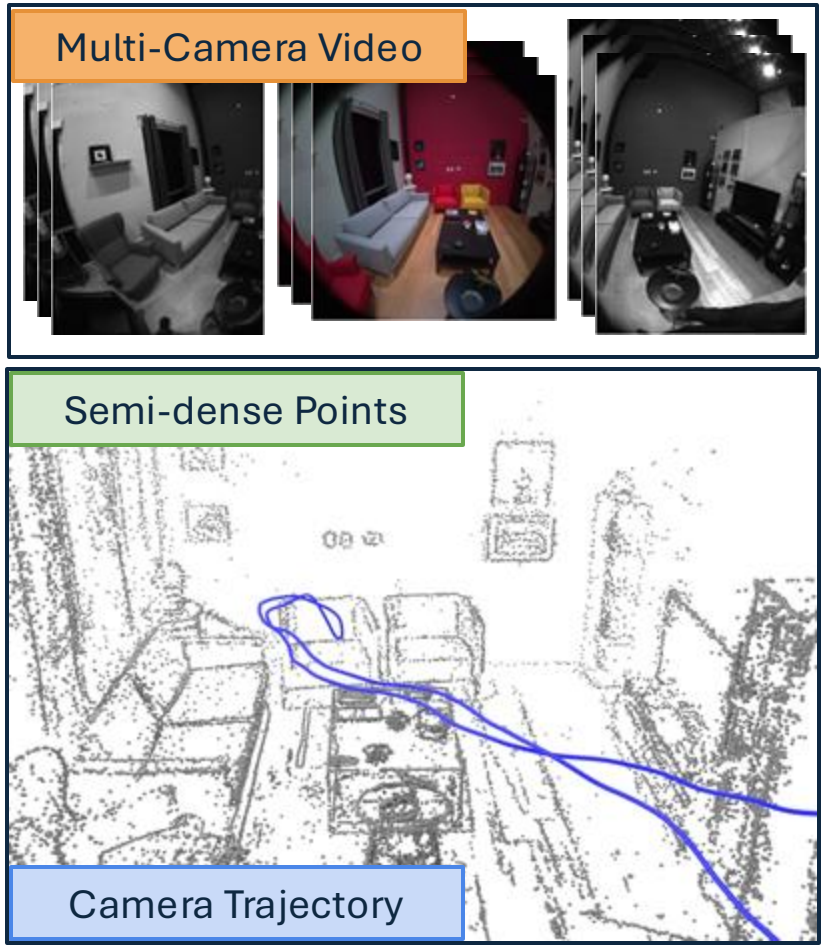
Eye tracking cameras

Health (PPG) and Contact mic

Stereo speakers

www.projectaria.com

Egocentric Data is a New Category of Data

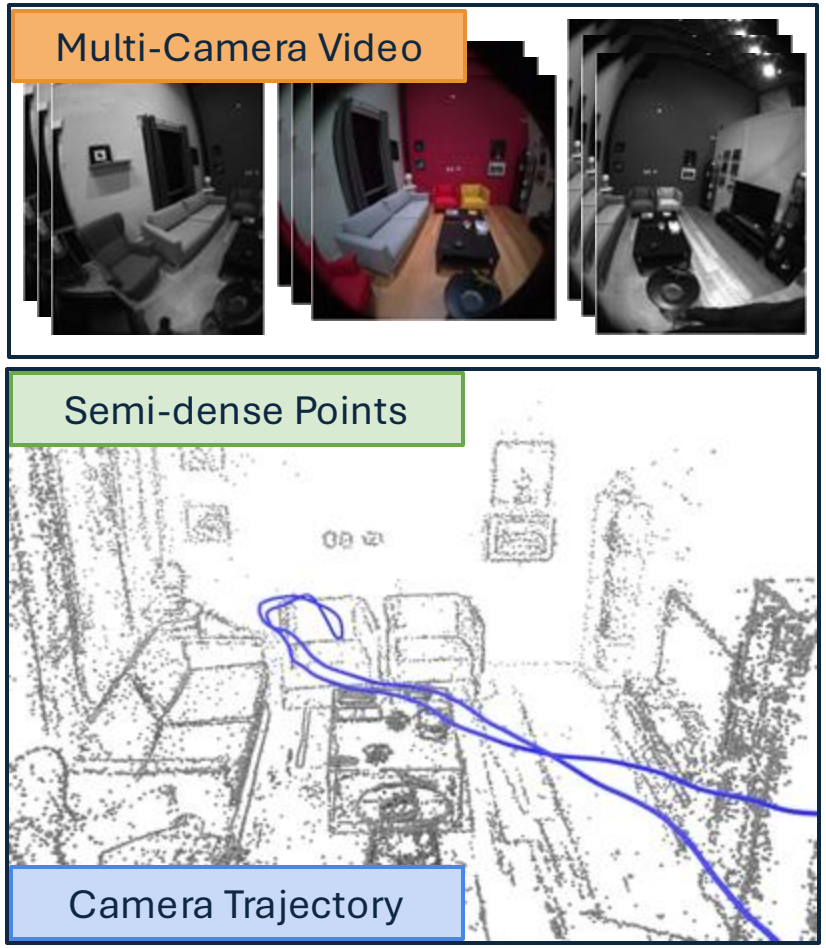


Egocentric Data

Project Aria Gen 1 Device

Project Aria Gen 2 Device

Egocentric Data is a New Category of Data

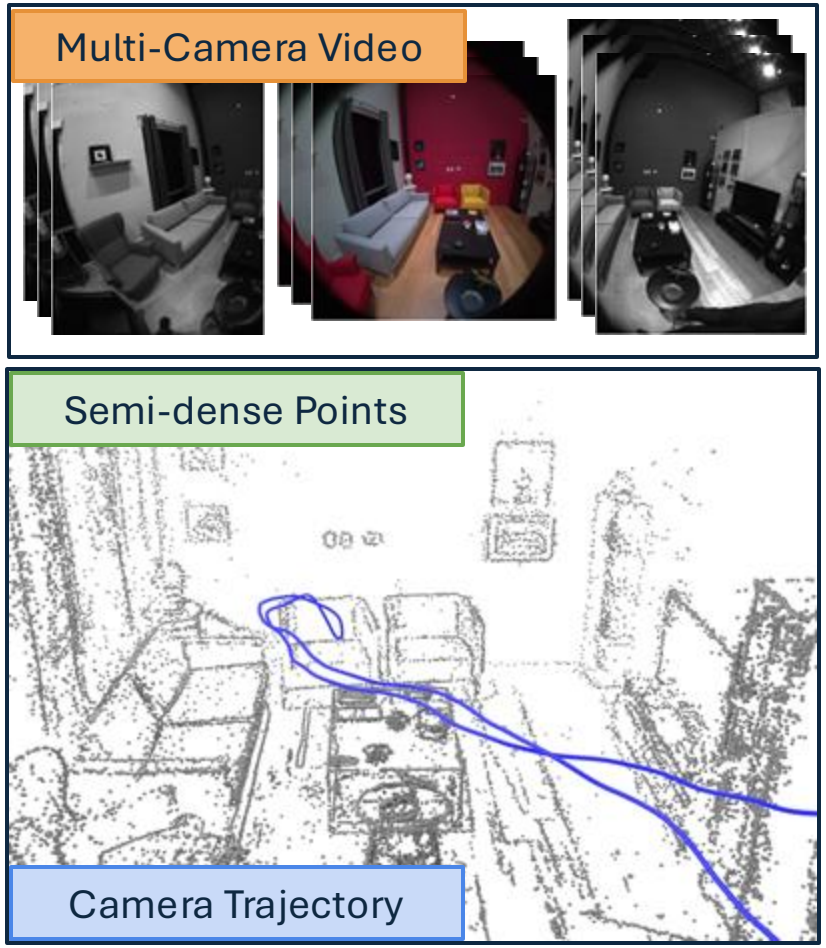


Egocentric Data

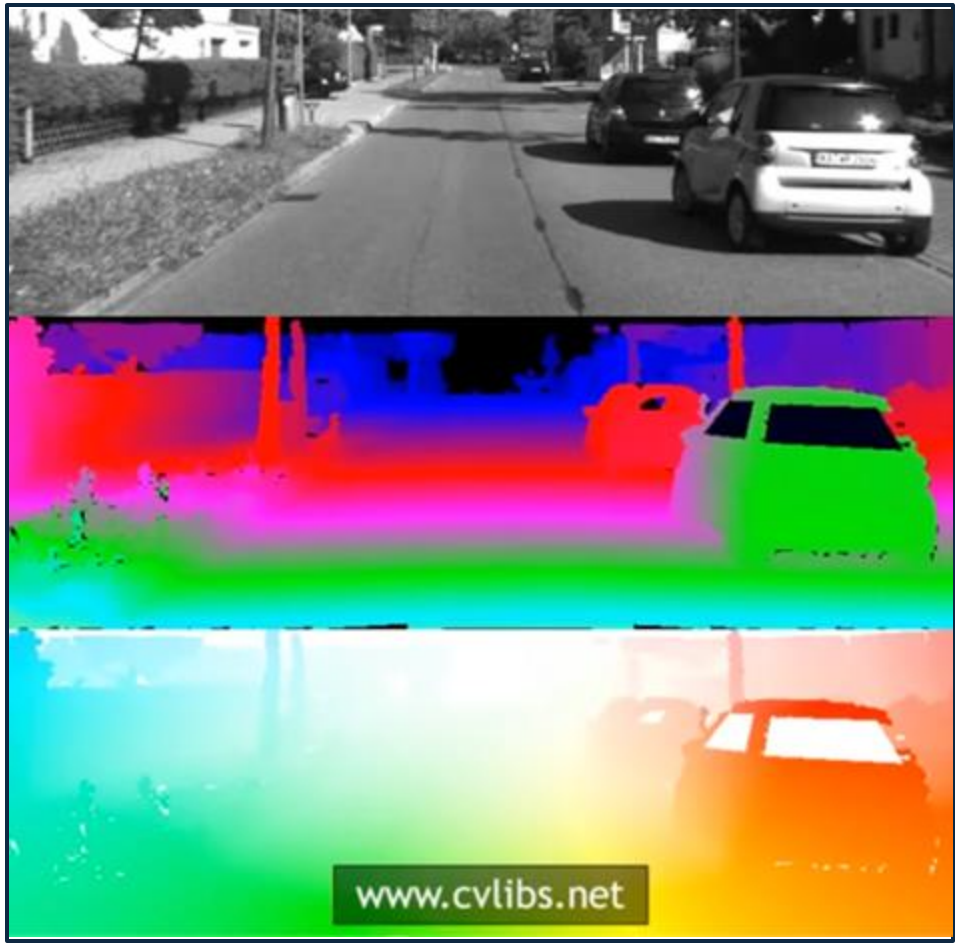
Key Properties of Egocentric data:

- Always-On Casual Capture
- Head-worn Natural Human Motion
- Partial Observations
- No Active Depth Camera
- Dynamic, Cluttered Scenes

Egocentric Data is a New Category of Data



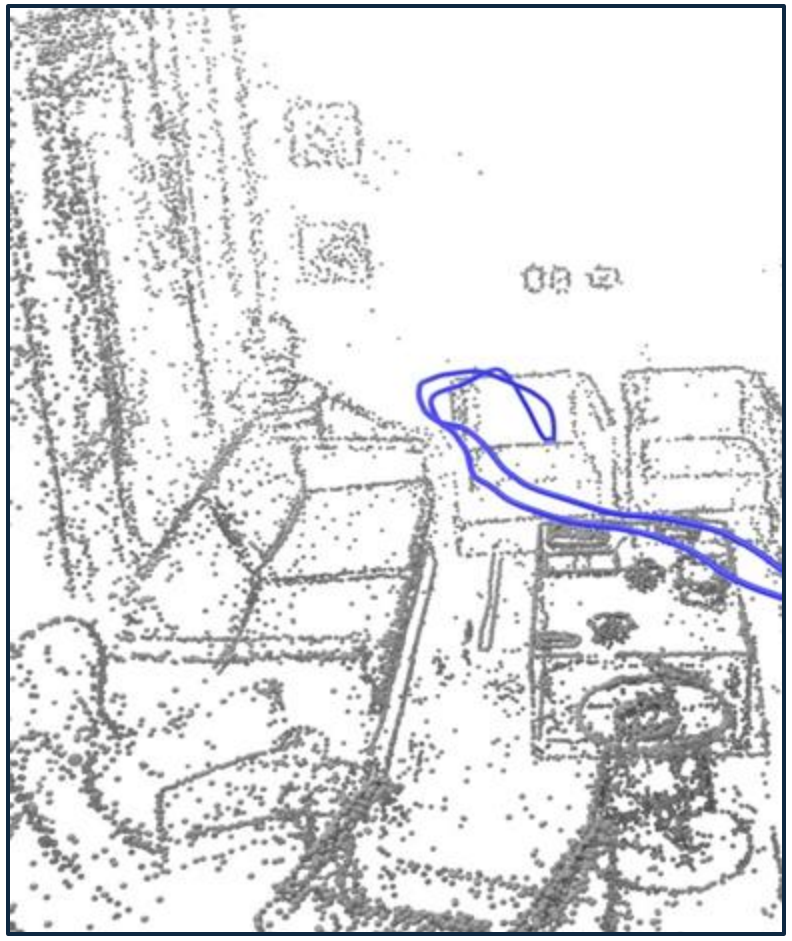
Egocentric Data



Car-centric Data
(Autonomous Car Data)

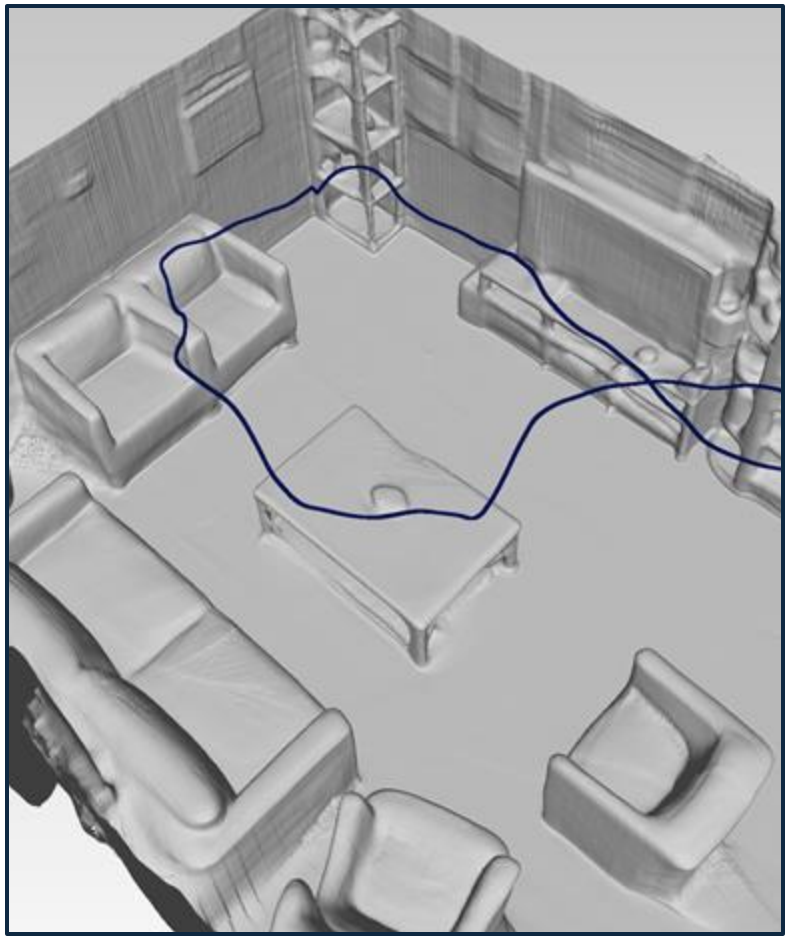
Room-centric Data
(Indoor Scanning Data)

3D & 4D Scene Information from Egocentric Devices



Semi-dense

Engel, Jakob, et al. *"Direct sparse odometry."* TPAMI



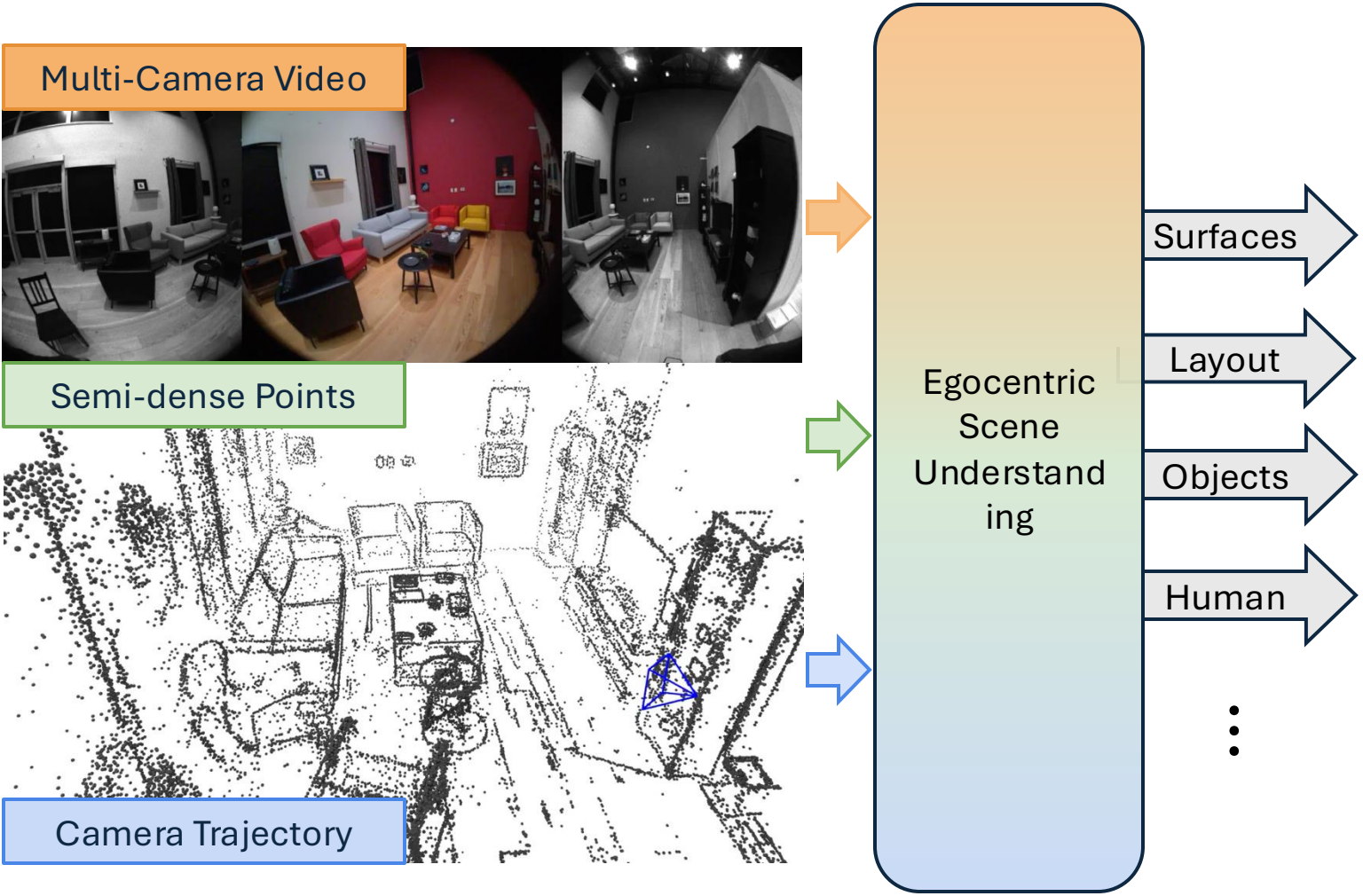
Dense

Straub, Julian, et al. *"EFM3D: A Benchmark for Measuring Progress Towards 3D Egocentric Foundation Models",* arXiv:2406.10224

Dynamic

Wen, Bowen, et al. *"FoundationStereo: Zero-shot stereo matching." CVPR* 2025.

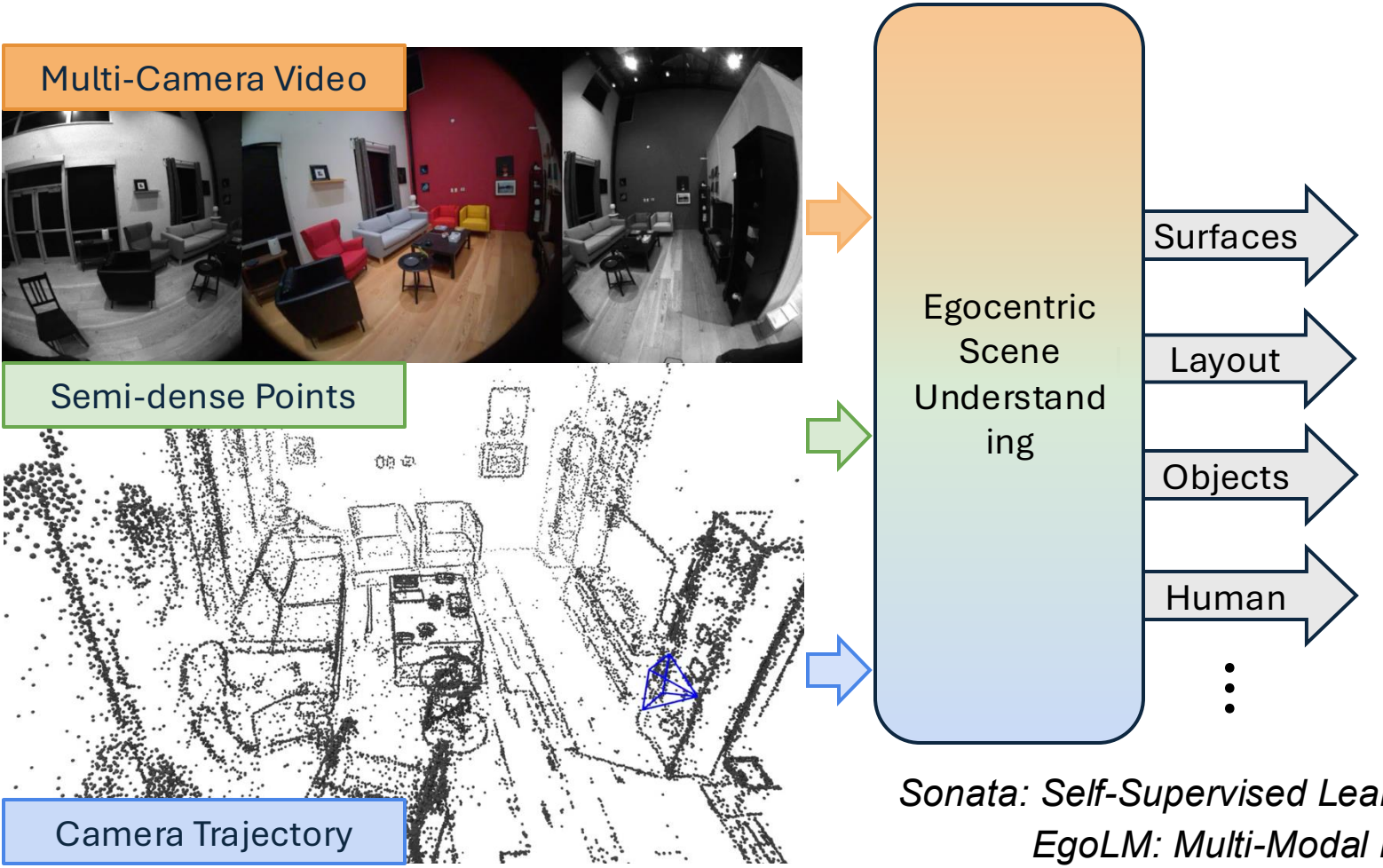
Egocentric Open-World Scene Understanding



Desired Properties

- Strong Learned Priors
- Scalable Incremental Inference
- Persistent Representation
- Coverage: Explain Observations Fully
- Predictivity

Egocentric Open-World Scene Understanding



Desired Properties

- Strong Learned Priors
- Scalable Incremental Inference
- Persistent Representation
- Coverage: Explain Observations Fully
- Predictivity

Sonata: Self-Supervised Learning of Reliable Point Representations, CVPR 2025

EgoLM: Multi-Modal Language Model of Egocentric Motions, CVPR 2025

HMD²: Environment-aware Motion Generation from Single Egocentric Head-Mounted Device, 3DV 2025

SceneScript: Reconstructing Scenes With An Autoregressive Structured Language Model, ECCV 2024

EFM3D: A Benchmark for Measuring Progress Towards 3D Egocentric Foundation Models, arXiv:2406.10224

Egocentric Open-World Scene Understanding

- Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).
- Lifting 2D open-world foundation models to 3D.
  - 2D Segmentation “Point Painting” on Sparse Point Cloud.
  - 2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).
  - 2D to 3D Bounding Box Lifting. (Sneak Peak).
- Self-supervised 3D foundation models (Sonata CVPR 2025).

Egocentric Open-World Scene Understanding

- **Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).**
- Lifting 2D open-world foundation models to 3D.
  - 2D Segmentation “Point Painting” on Sparse Point Cloud.
  - 2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).
  - 2D to 3D Bounding Box Lifting. (Sneak Peak).
- Self-supervised 3D foundation models (Sonata CVPR 2025).

Egocentric Closed-Set 3D Object Detection is working decently

Straub, Julian, et al. *"EFM3D: A Benchmark for Measuring Progress Towards 3D Egocentric Foundation Models"*, arXiv:2406.10224

Egocentric Open-World Scene Understanding

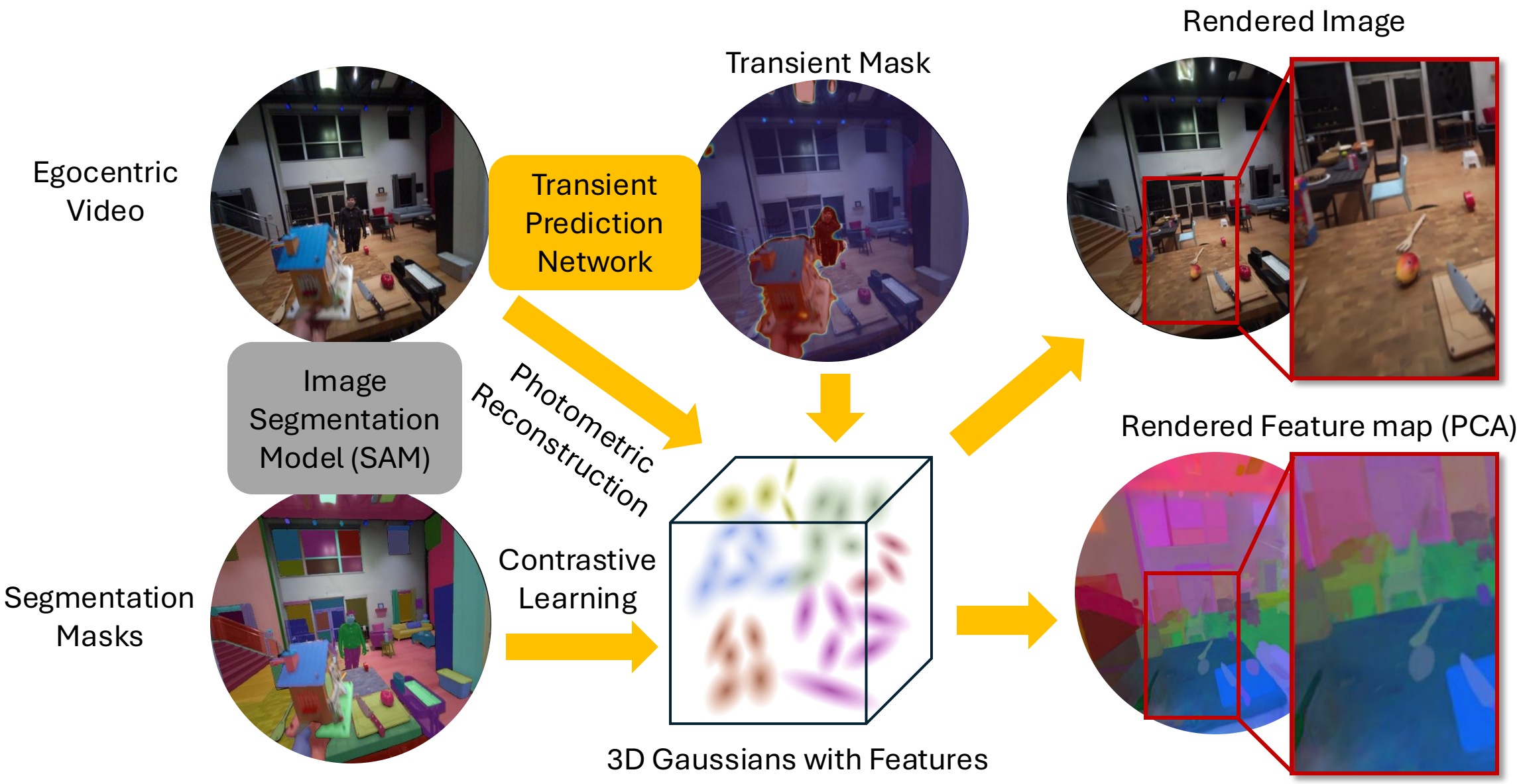
- Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).
- Lifting 2D open-world foundation models to 3D.
  - **2D Segmentation “Point Painting” on Sparse Point Cloud.**
  - 2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).
  - 2D to 3D Bounding Box Lifting. (Sneak Peak).
- Self-supervised 3D foundation models (Sonata CVPR 2025).


Plant
vase
vase
Screen
fireplace
candle
book
Table
Couch
Plant

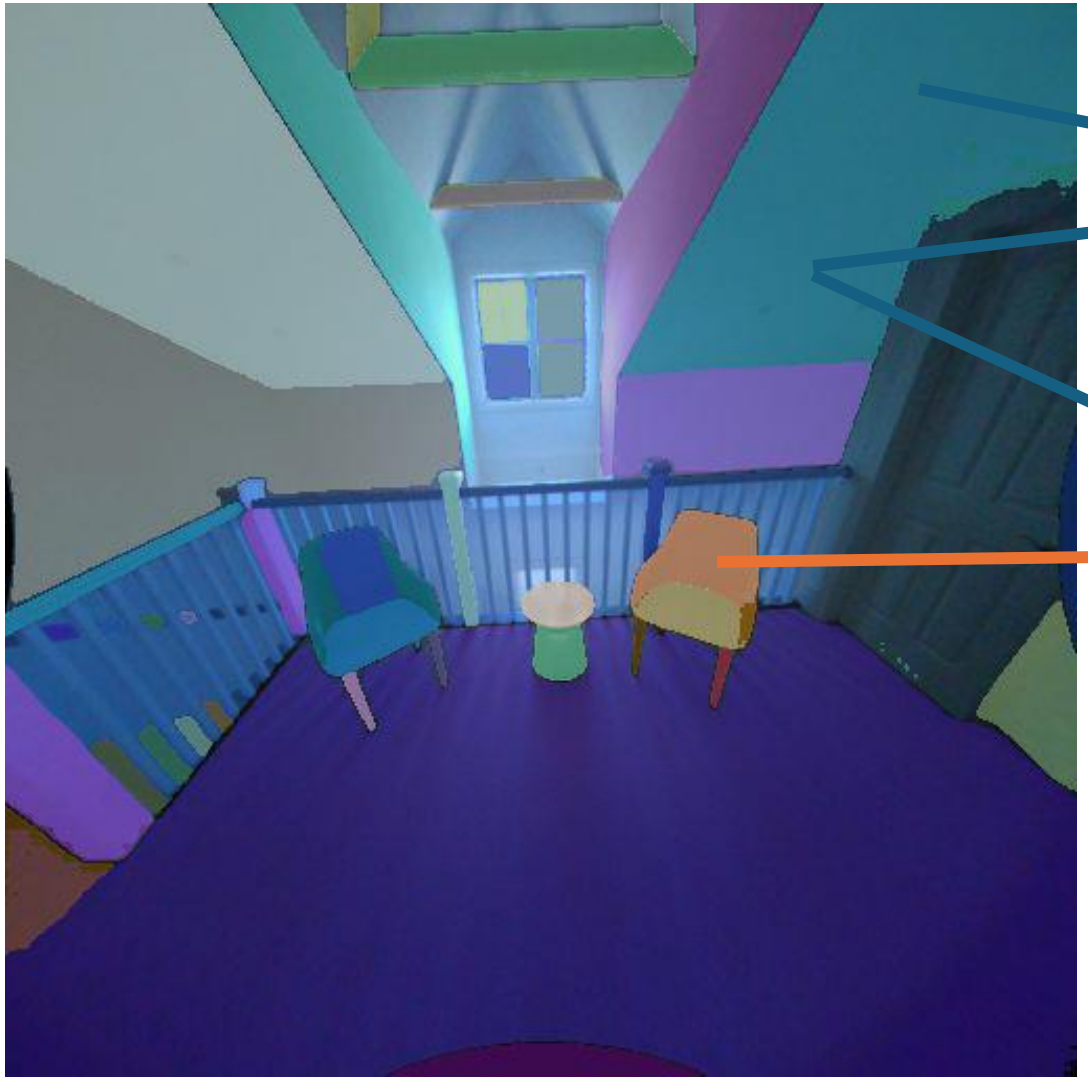
Egocentric Open-World Scene Understanding

- Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).
- Lifting 2D open-world foundation models to 3D.
  - 2D Segmentation “Point Painting” on Sparse Point Cloud.
  - **2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).**
  - 2D to 3D Bounding Box Lifting. (Sneak Peak).
- Self-supervised 3D foundation models (Sonata CVPR 2025).

EgoLifter (ECCV 2024)



EgoLifter (ECCV 2024)



Masks from Segment-Anything (SAM)

Positive Pair

Contrastive learning on rendered features at these pixels from 3D GS.

Negative Pair

Lift SAM to 3D by implicitly solving multi-view association problem of 2D segmentation masks.

EgoLifter reconstructs a clean 3D feature field anchored to GS in the presence of typical egocentric dynamics in the scene.

Clustering Features into Open-World Objects

Egocentric Open-World Scene Understanding

- Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).
- Lifting 2D open-world foundation models to 3D.
  - 2D Segmentation “Point Painting” on Sparse Point Cloud.
  - 2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).
  - **2D to 3D Bounding Box Lifting. (Sneak Peak).**
- Self-supervised 3D foundation models (Sonata CVPR 2025).

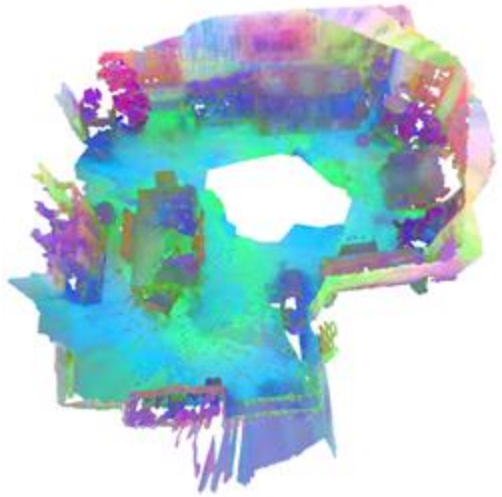
Sneak Peak: 2D to 3D Bounding Box Lifting Model

Egocentric Open-World Scene Understanding

- Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).
- Lifting 2D open-world foundation models to 3D.
  - 2D Segmentation “Point Painting” on Sparse Point Cloud.
  - 2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).
  - 2D to 3D Bounding Box Lifting. (Sneak Peak).
- **Self-supervised 3D foundation models (Sonata CVPR 2025).**

Why not simply run 2D backbones and lift them to 3D?

… × N



It requires **significant inference time** to process a scene video with thousands of frames.
It **lacks geometric awareness**, as it is not natively encoded in 3D space.

Linear-probed Self-Supervised 2D Backbones ~= Supervised ones.

Method	ViT	ImageNet Val	ImageNet V2	ImageNet ReaL
Supervised backbones				
Zhai et al. (2022a)*	G/14	89.0	81.3	90.6
Chen et al. (2023)*	e/14	89.3	82.5	90.7
Dehghani et al. (2023)*	22B/14	89.5	83.2	90.9
Weakly-supervised backbones				
PEcore	G/14	**89.3**	**81.6**	90.4
SigLIP 2	g/16	89.1	81.6	**90.5**
AIMv2	3B/14	87.9	79.5	89.7
EVA-CLIP	18B/14	87.9	79.3	89.5
Self-supervised backbones				
Web-DINO	7B/14	85.9	77.1	88.6
Franca	g/14	84.8	75.3	89.2
DINOv2	g/14	87.3	79.5	89.9
DINOv3	7B/16	88.4	81.4	90.4

< 2% points difference between supervised and unsupervised 2D backbones

In 3D there is a 50-60% point gap!

Method	ViT	ImageNet Val	ImageNet V2	ImageNet ReaL
Supervised backbones				
Zhai et al. (2022a)*	G/14	89.0	81.3	90.6
Chen et al. (2023)*	e/14	89.3	82.5	90.7
Dehghani et al. (2023)*	22B/14	89.5	83.2	90.9
Weakly-supervised backbones				
PEcore	G/14	**89.3**	**81.6**	90.4
SigLIP 2	g/16	89.1	81.6	**90.5**
AIMv2	3B/14	87.9	79.5	89.7
EVA-CLIP	18B/14	87.9	79.3	89.5
Self-supervised backbones				
Web-DINO	7B/14	85.9	77.1	88.6
Franca	g/14	84.8	75.3	89.2
DINOv2	g/14	87.3	79.5	89.9
DINOv3	7B/16	88.4	81.4	90.4

Param. Effciency	Params		ScanNet Val [23]		
Methods	Learn.	Pct.	mIoU	mAcc	allAcc
○ SparseUNet [17]	39.2M	100%	72.3	80.2	90.0
● PC [93] (lin.)	<0.2M	<0.1%	5.6	9.7	50.0
● CSC [38] (lin.)	<0.2M	<0.1%	12.6	18.1	64.2
● MSC [88] (lin.)	<0.2M	<0.1%	14.1	20.3	62.9
○ PTv3 [89]	124.8M	100%	77.6	85.0	92.0
● MSC [88] (lin.)	<0.2M	<0.2%	21.8	32.2	65.5

< 2% points difference between supervised and unsupervised 2D backbones.

50-60% points difference between supervised and unsupervised 3D backbones.

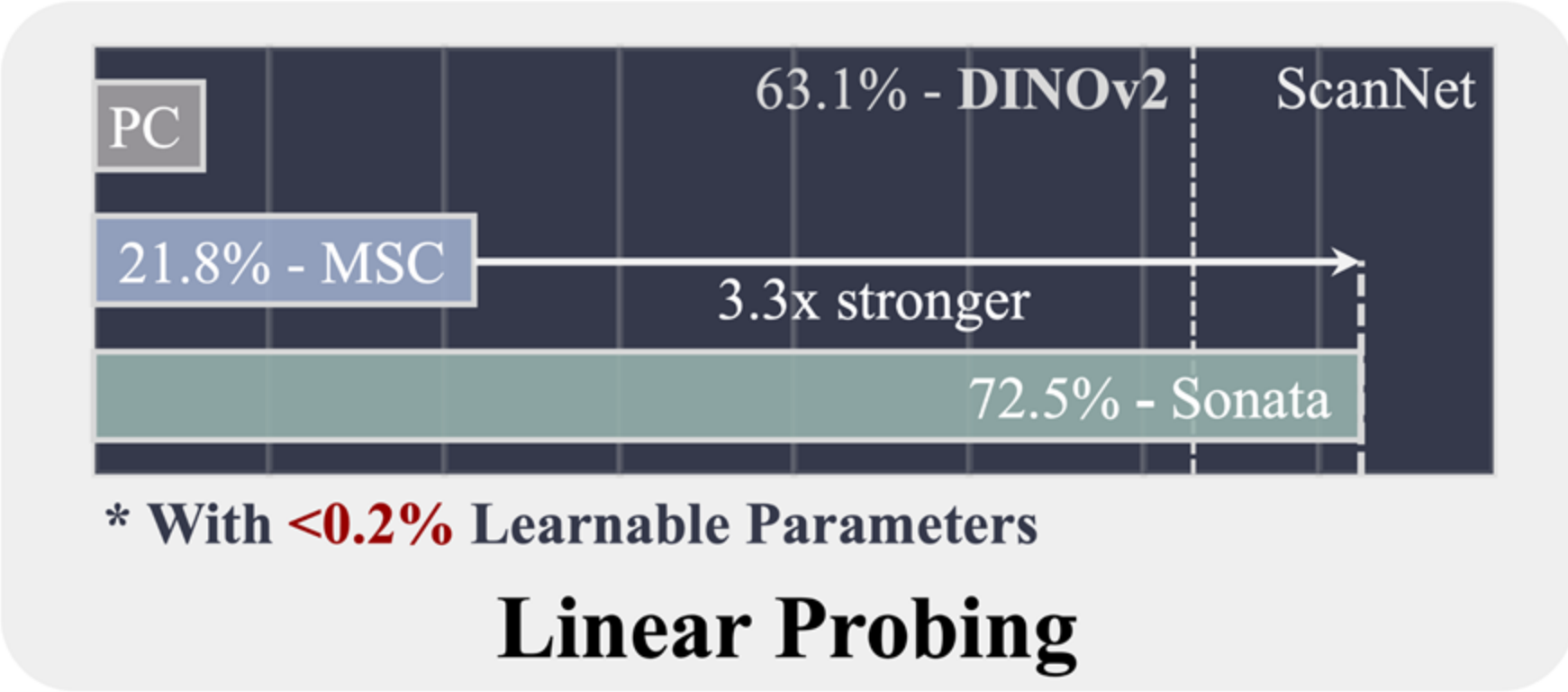
Sonata pretraining improves linear probing by 3.3x down to a gap of only 5%.

Method	ViT	ImageNet Val	ImageNet V2	ImageNet ReaL
Supervised backbones				
Zhai et al. (2022a)*	G/14	89.0	81.3	90.6
Chen et al. (2023)*	e/14	89.3	82.5	90.7
Dehghani et al. (2023)*	22B/14	89.5	83.2	90.9
Weakly-supervised backbones				
PEcore	G/14	**89.3**	**81.6**	90.4
SigLIP 2	g/16	89.1	81.6	**90.5**
AIMv2	3B/14	87.9	79.5	89.7
EVA-CLIP	18B/14	87.9	79.3	89.5
Self-supervised backbones				
Web-DINO	7B/14	85.9	77.1	88.6
Franca	g/14	84.8	75.3	89.2
DINOv2	g/14	87.3	79.5	89.9
DINOv3	7B/16	88.4	81.4	90.4

Param. Effciency	Params		ScanNet Val [23]		
Methods	Learn.	Pct.	mIoU	mAcc	allAcc
○ SparseUNet [17]	39.2M	100%	72.3	80.2	90.0
● PC [93] (lin.)	<0.2M	<0.1%	5.6	9.7	50.0
● CSC [38] (lin.)	<0.2M	<0.1%	12.6	18.1	64.2
● MSC [88] (lin.)	<0.2M	<0.1%	14.1	20.3	62.9
○ PTv3 [89]	124.8M	100%	77.6	85.0	92.0
● MSC [88] (lin.)	<0.2M	<0.2%	21.8	32.2	65.5
● Sonata (lin.)	<0.2M	<0.2%	72.5	83.1	89.7
● Sonata (dec.)	16.3M	13%	**79.1**	**86.6**	**92.7**

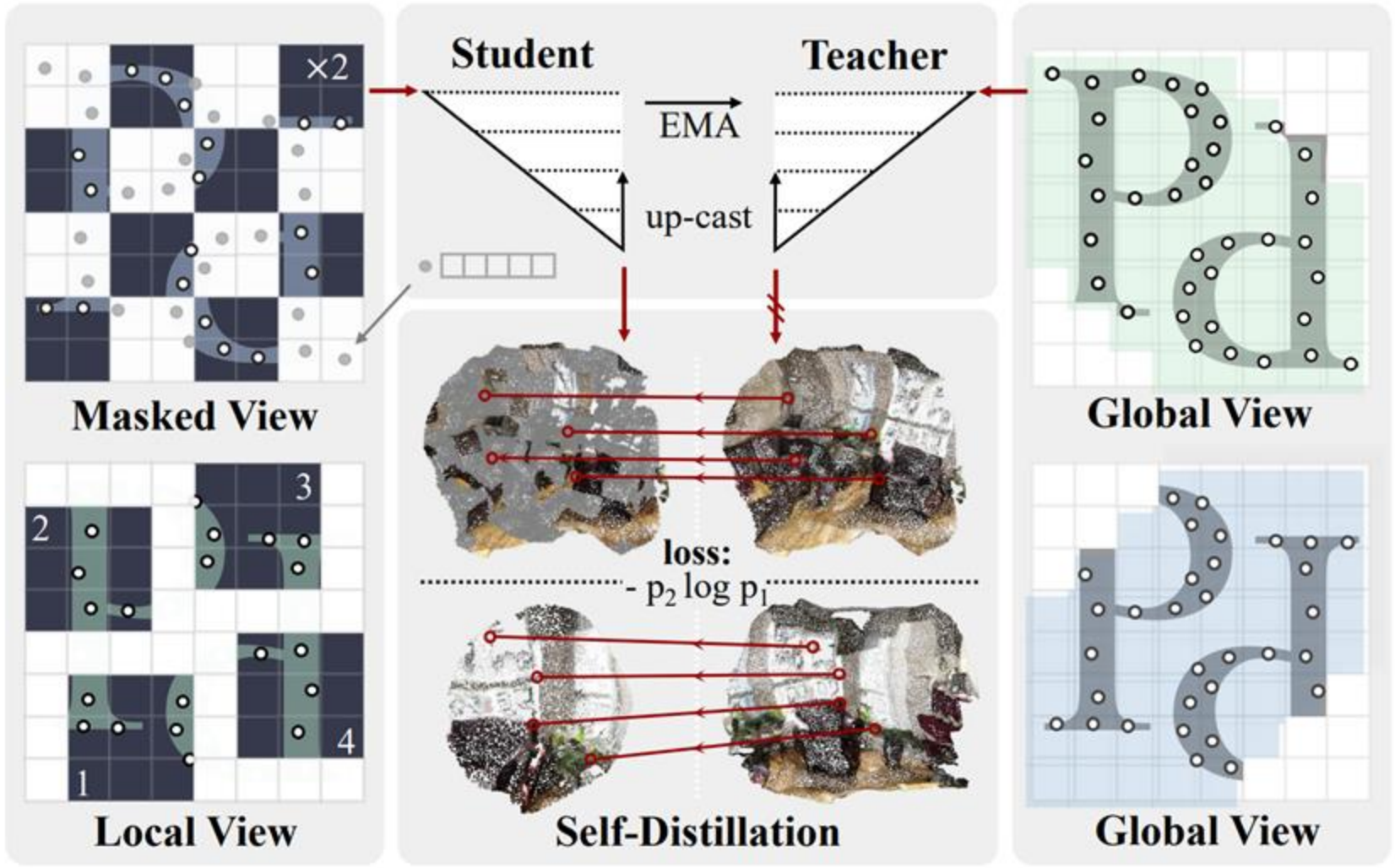
< 2% points difference between supervised and unsupervised 2D backbones.

Sonata improves to ~5% points difference between supervised and unsupervised training.



How did we do it?

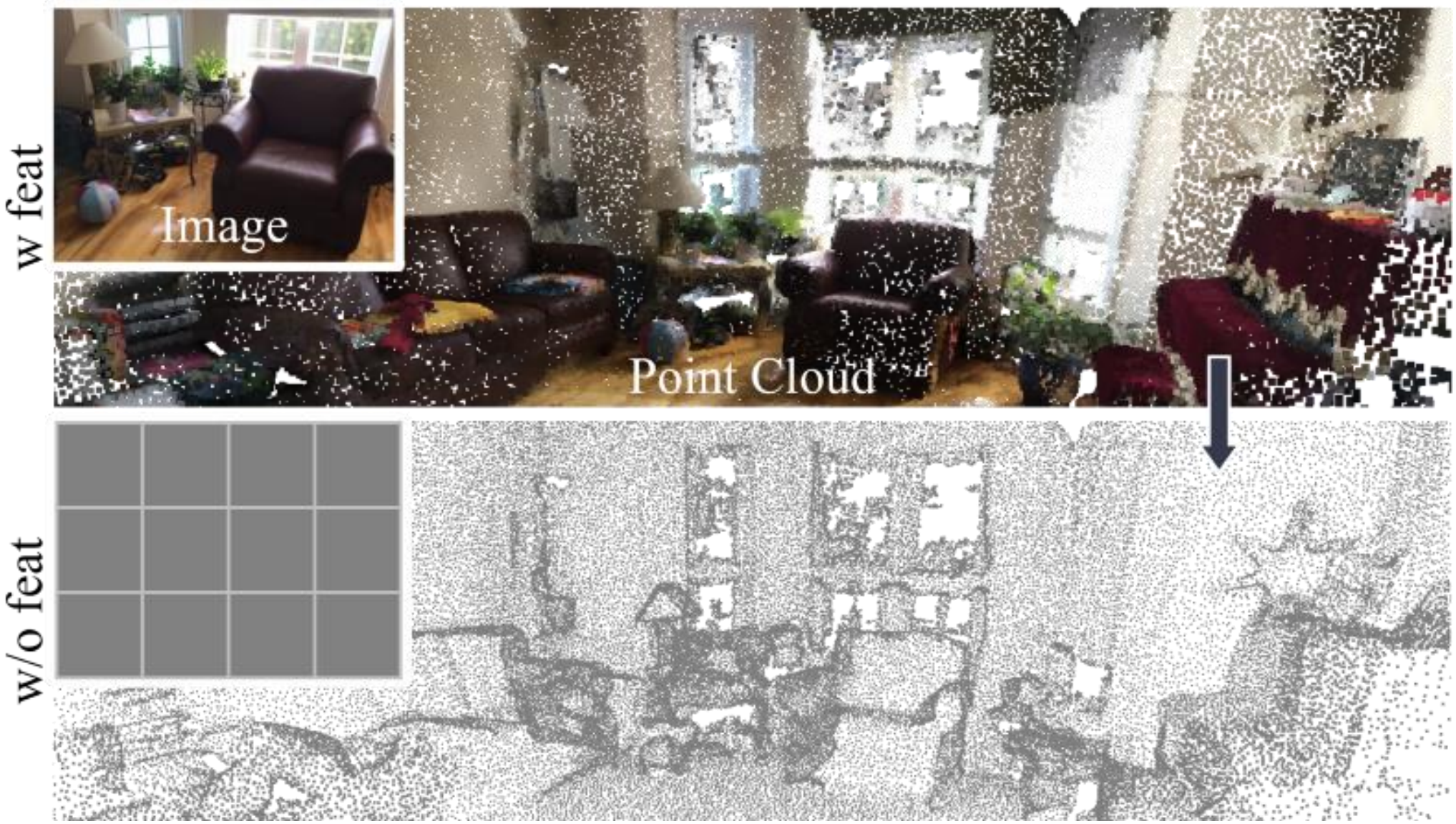
Sonata Pretraining largely follows 2D Image Self-Supervised Learning



- Local-Global Alignment
- Mask-Unmask Alignment
- Exponential Moving Average
- Online Clustering
- Sinkhorn-Knopp centering
- KoLeo regularizer
- Simple Model & Data Scaling

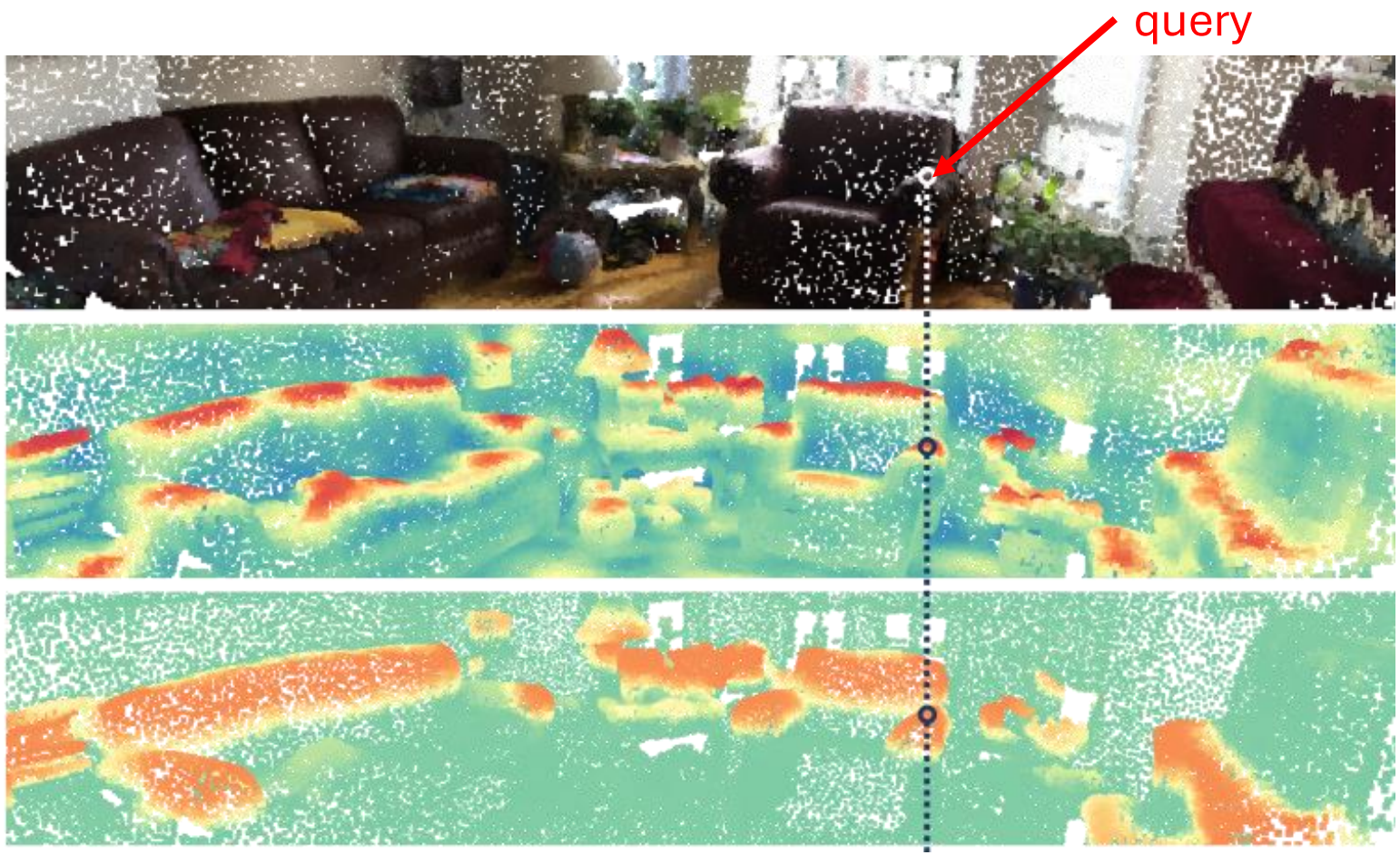
But there is one key difference to 2D—the “Geometric Shortcut”

But there is one key difference to 2D—the “Geometric Shortcut”

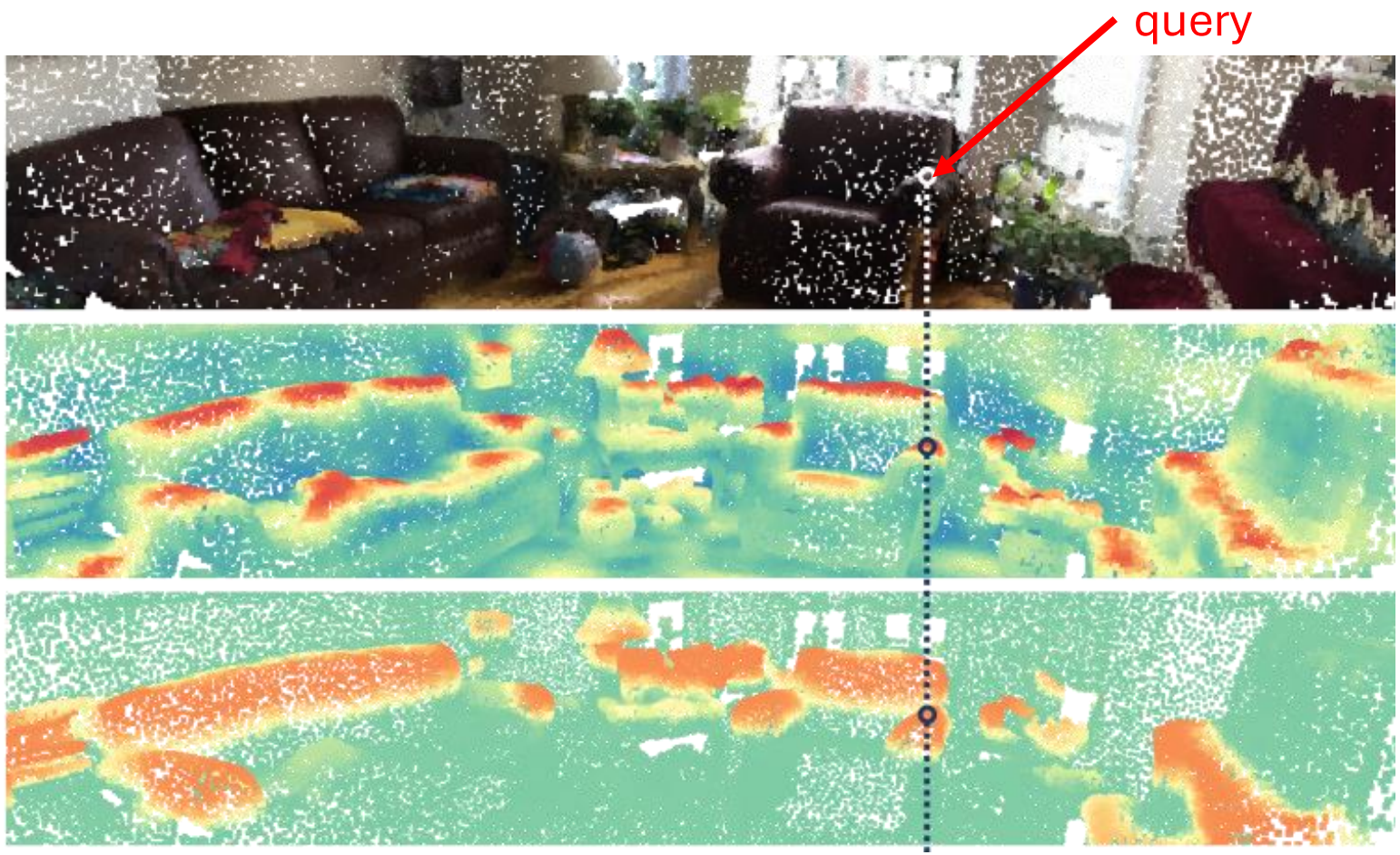


All images are organized in the same way. Point clouds are not.

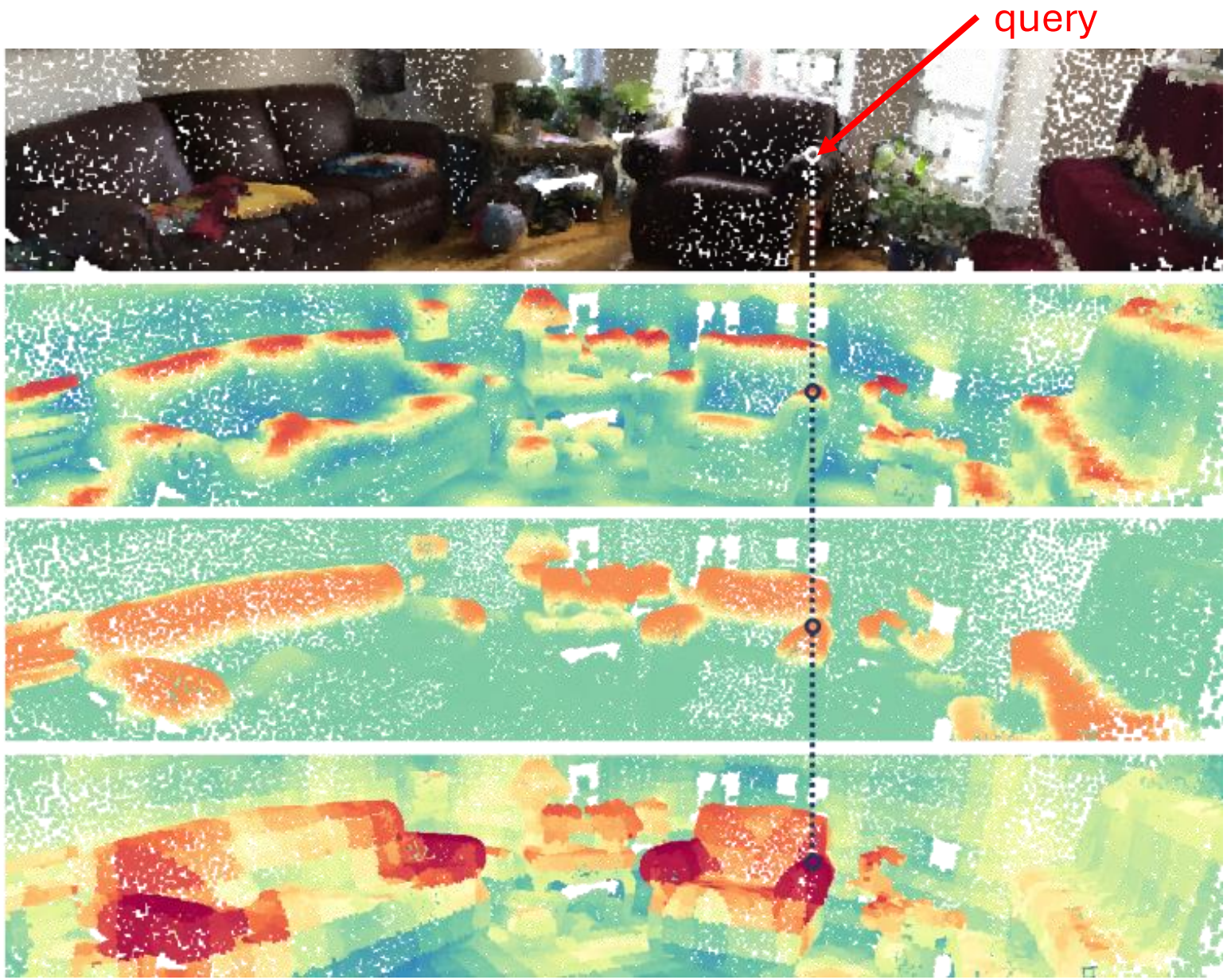
Geometric Shortcut Collapses Learned 3D Representations



Geometric Shortcut Collapses Learned 3D Representations

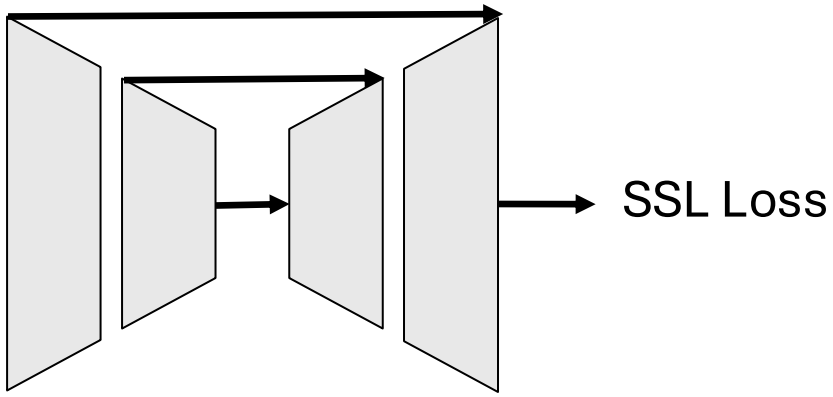


Geometric Shortcut Collapses Learned 3D Representations



Two Key Strategies to Break the Shortcut:

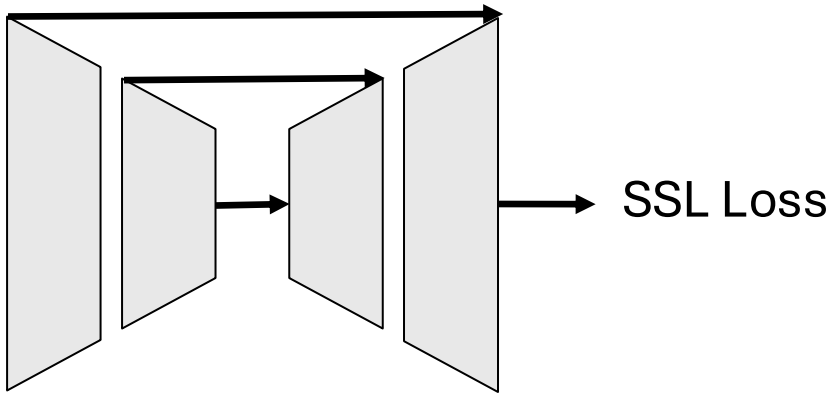
Decoder Free Design: Replacing Decoder with Learning-Free Upcasting



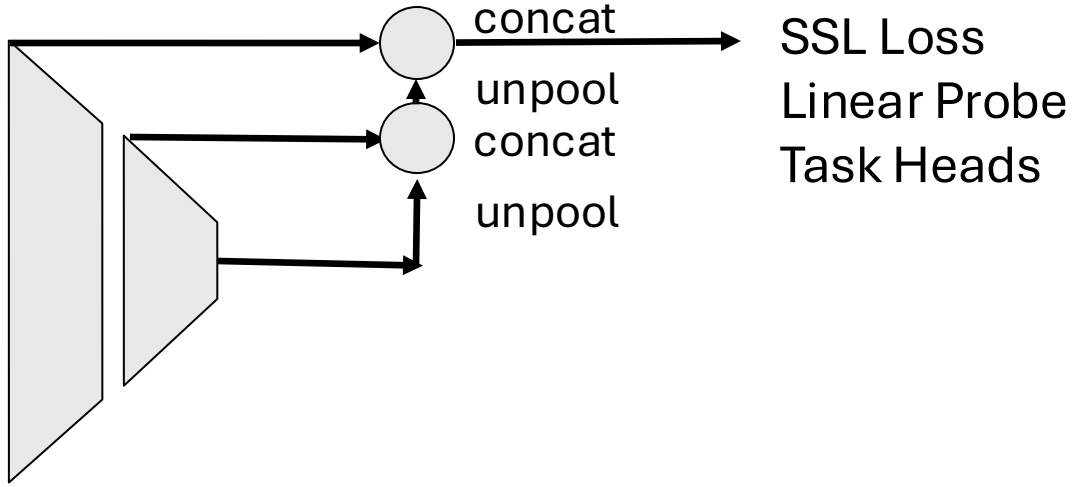
Standard Encoder-Decoder

Two Key Strategies to Break the Shortcut:

Decoder Free Design: Replacing Decoder with Learning-Free Upcasting



Standard Encoder-Decoder

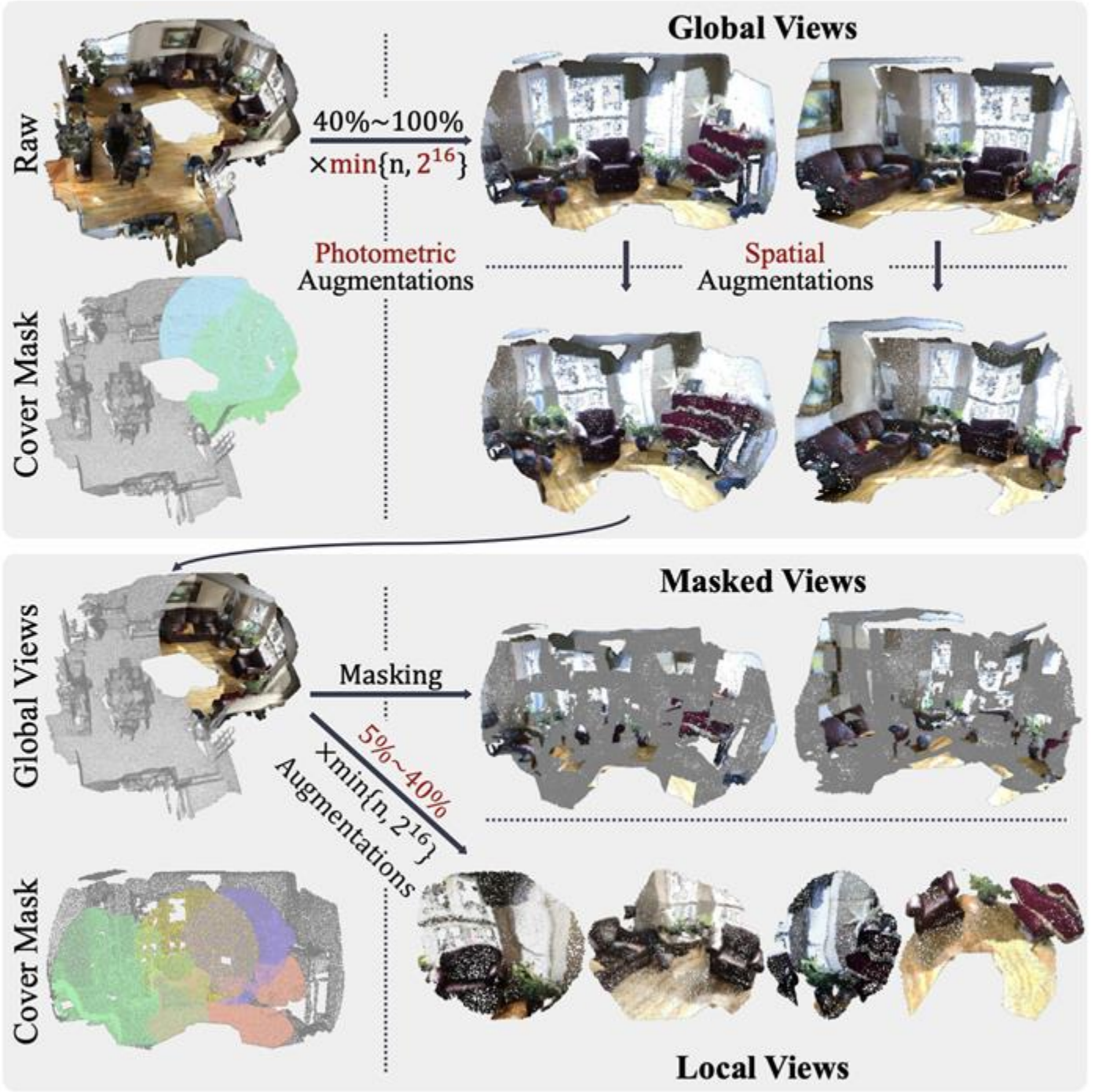


Upcasting of Output Features

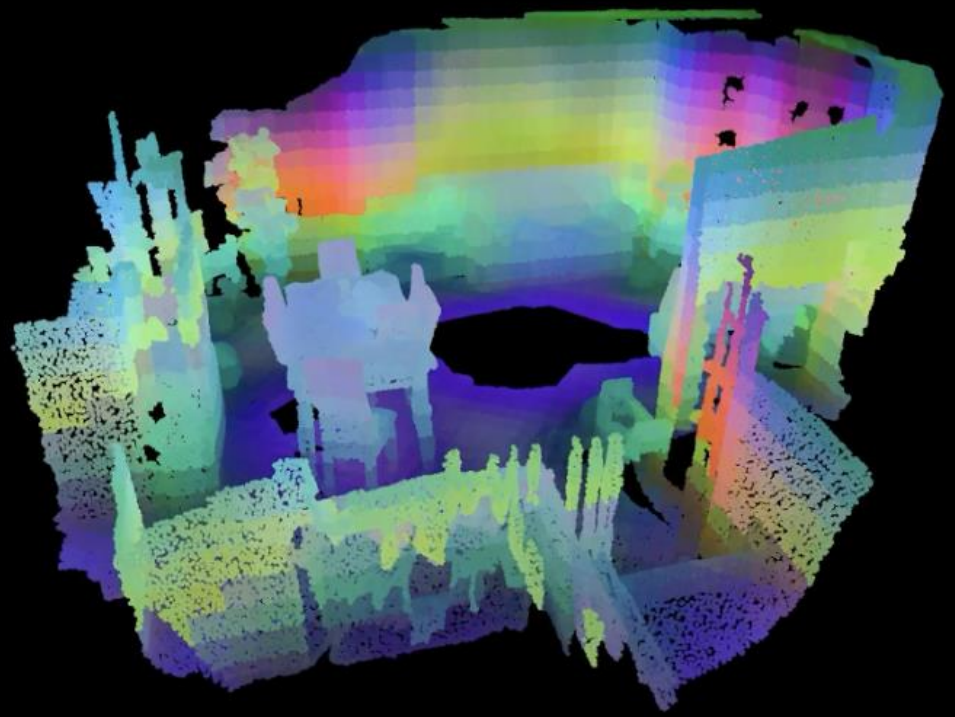
- no “wasted” parameters in decoder
- better gradient flow

Two Key Strategies to Break the Shortcut:

Self-distillation with strong spatial obscuration that gradually increases throughout training.

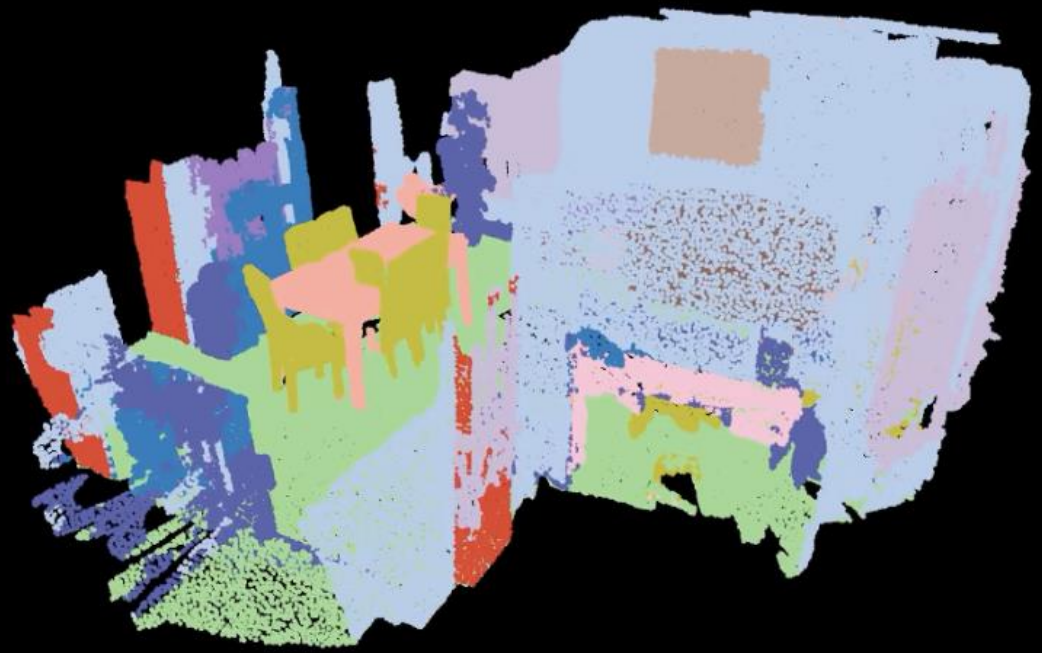


Input Pointcloud



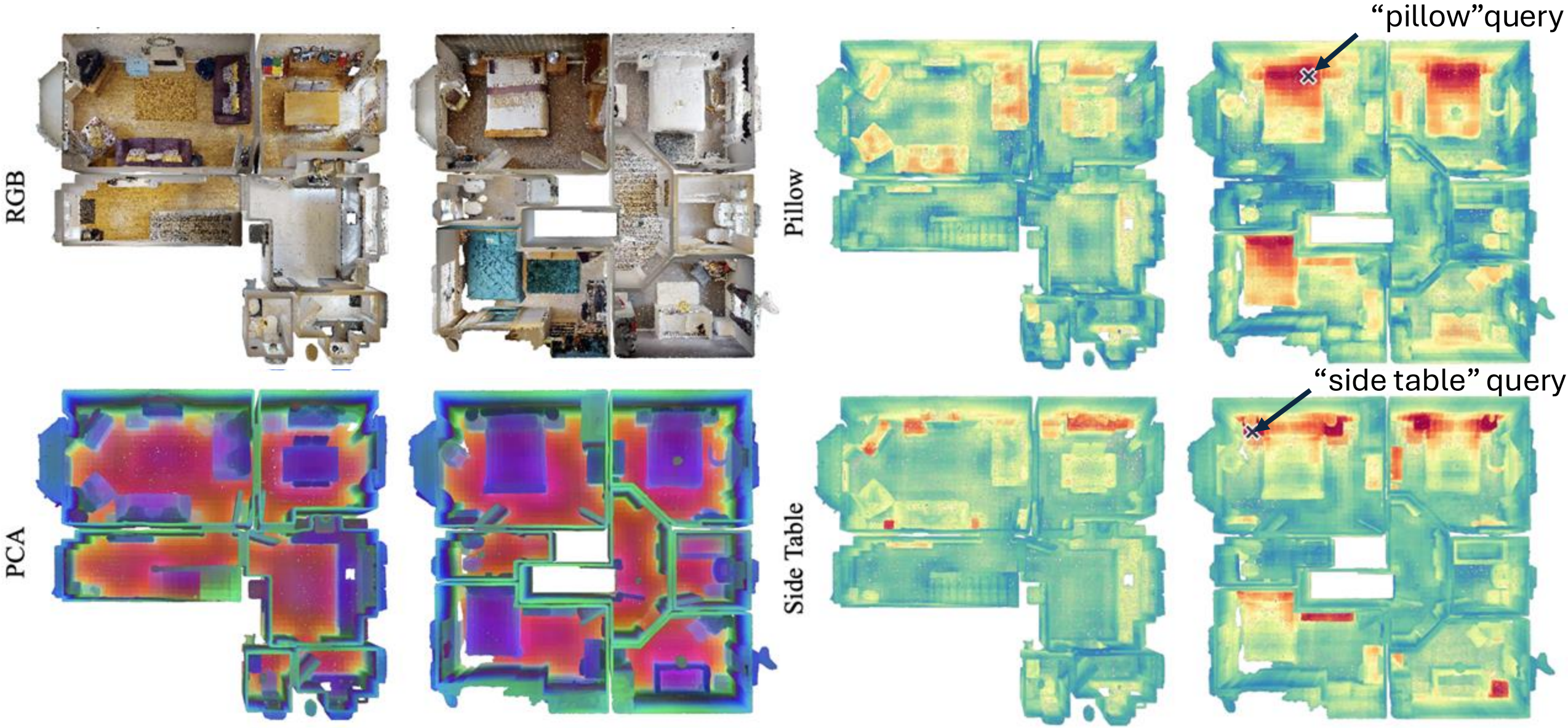
Sonata: Features (PCA)

Input Pointcloud

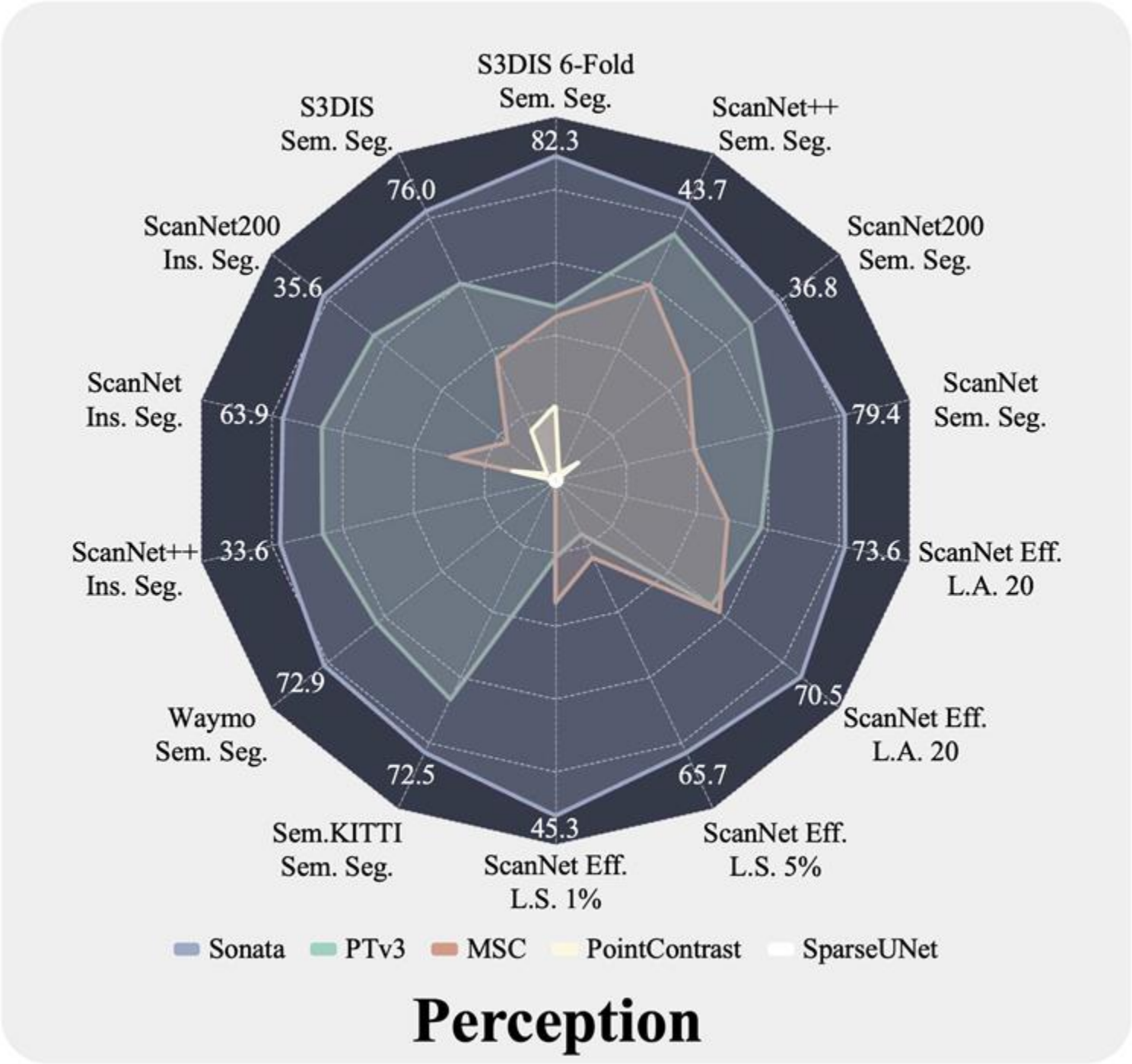


Sonata: Segmentation (Linear Probe)

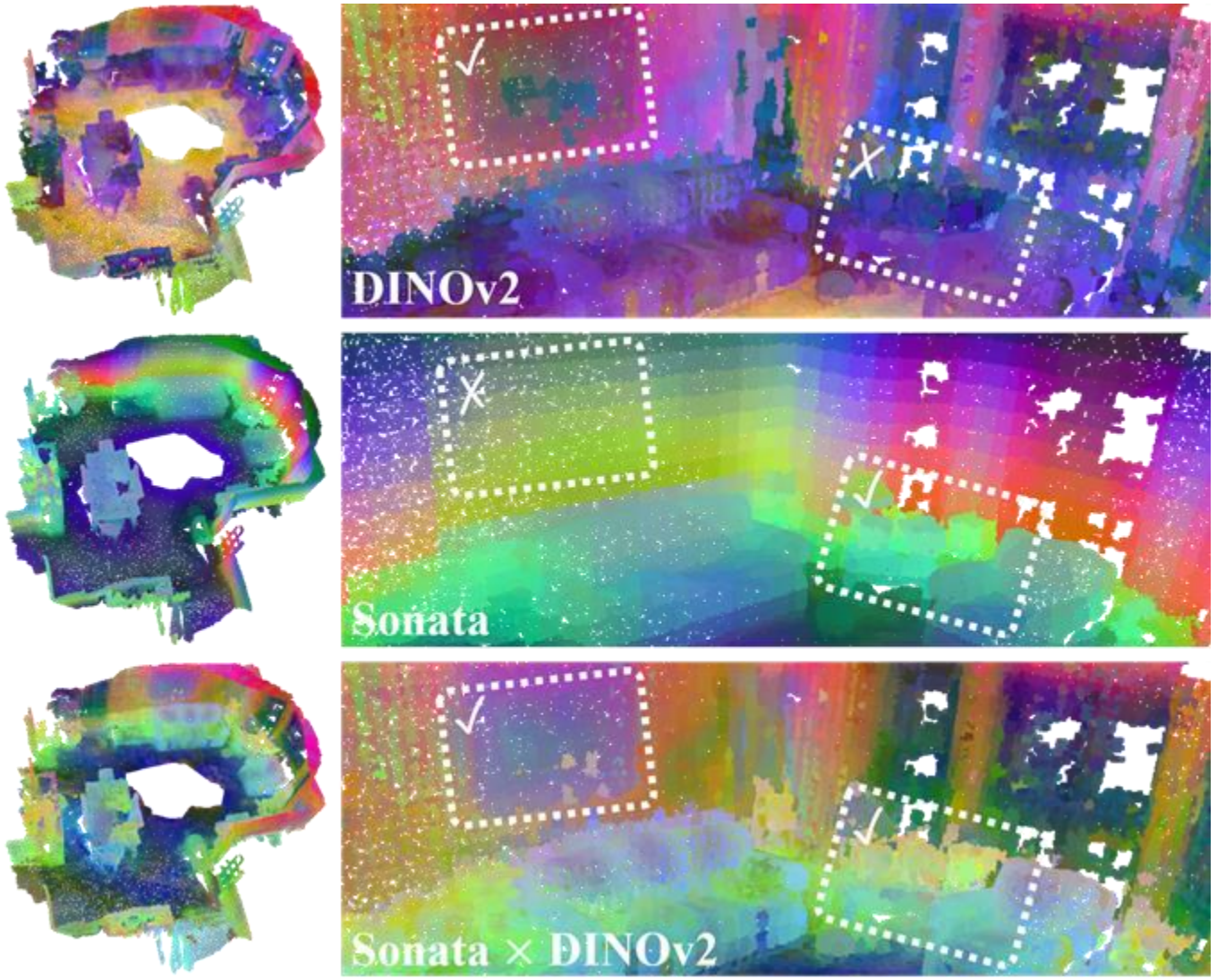
Similarity Heatmaps for Different Query Locations



Results - Sonata is SOTA



Results - concat(Sonata, DINOv2.5) > Sonata > DINOv2.5

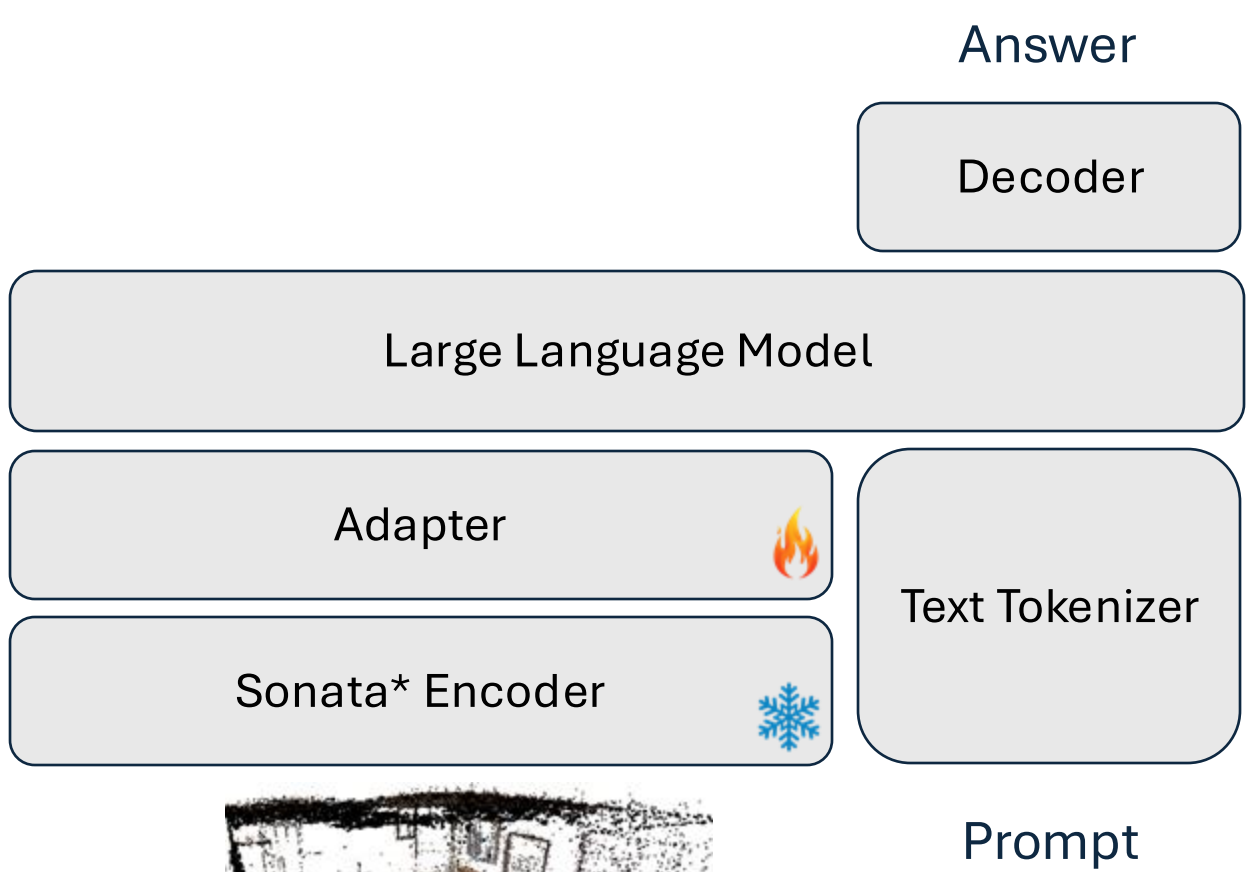


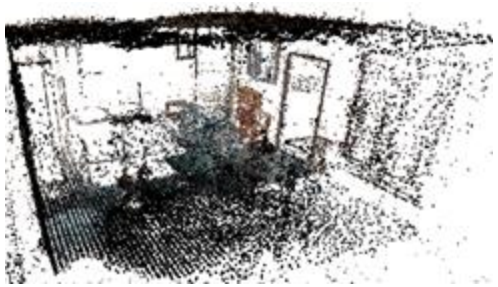
2D × 3D	ScanNet Val [23]			ScanNet200 Val [23]		
Methods	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
• DINOv2 (lin.) [60]	63.09	75.50	82.42	27.42	37.59	72.80
• DINOv2.5 (lin.) [24]	63.36	75.94	82.30	27.75	39.23	72.53
• Sonata (lin.)	72.52	83.11	89.74	29.25	41.61	81.15
+ DINOv2 (lin.)	75.91	85.36	91.25	36.67	46.98	**82.85**
+ DINOv2.5 (lin.)	**76.44**	**85.68**	**91.33**	**36.96**	**48.23**	82.77

Scene Understanding by Injecting Spatial Information into LLM via Sonata

Encoder	F1 - object IoU_{3D}@0.25	F1 - object IoU_{3D}@0.5
Voxelize (**P**+DINOv2)	0.1	0.0
Rand. sampling (**P**+DINOv2)	0.2	0.0
3DCNN enc. (**P**)	59.8	46.1
3DCNN enc. (**P**+DINOv2)	57.1	45.0
3DCNN enc. (**P**) + Voxelize (DINOv2	62.9	46.7
Sonata/PTv3 enc. (**P**)	**65.1**	**49.4**

Mao, Yongsen, *"SpatialLM: Training Large Language Models for Structured Indoor Modeling"*, Arxiv 2025



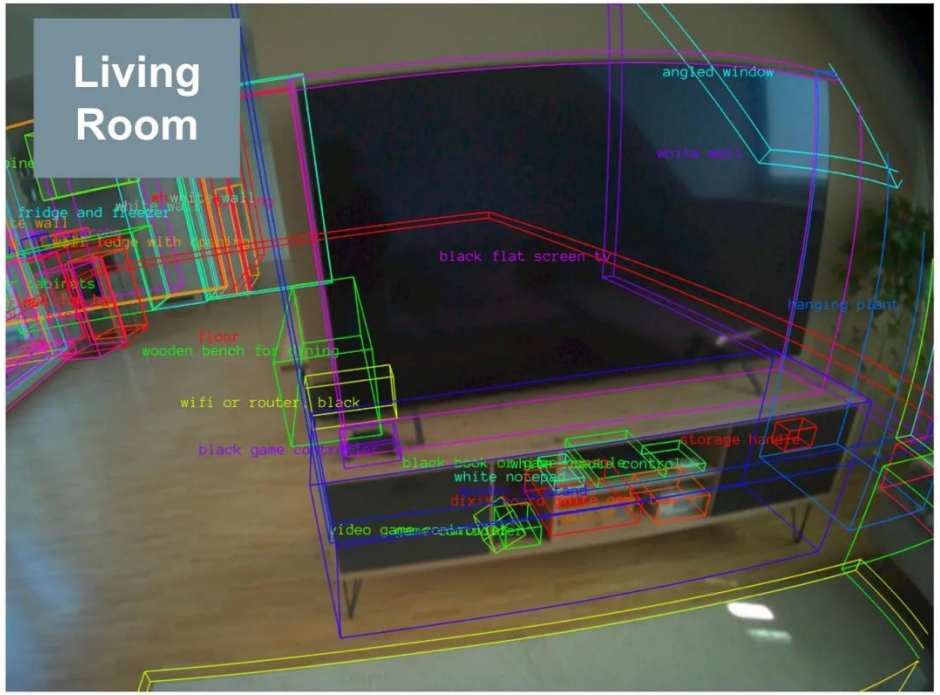


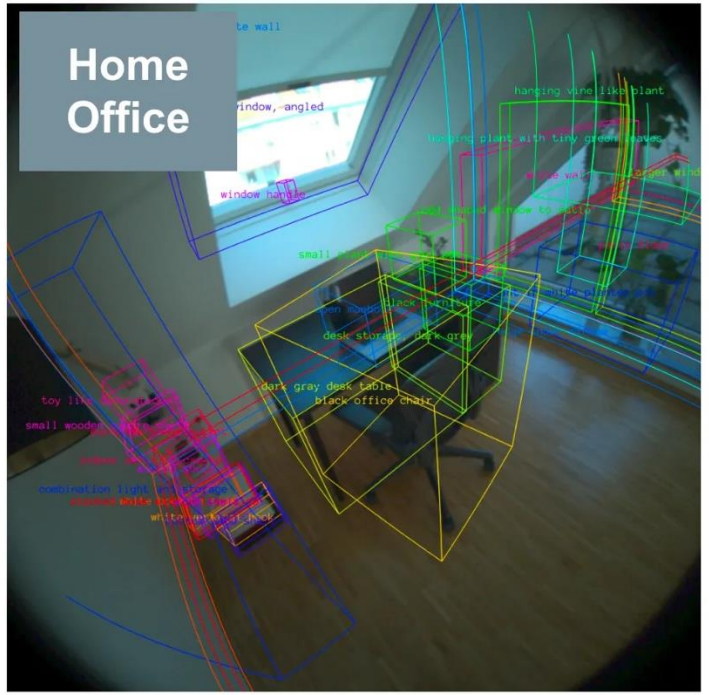
* it will still help to finetune Sonata on your data

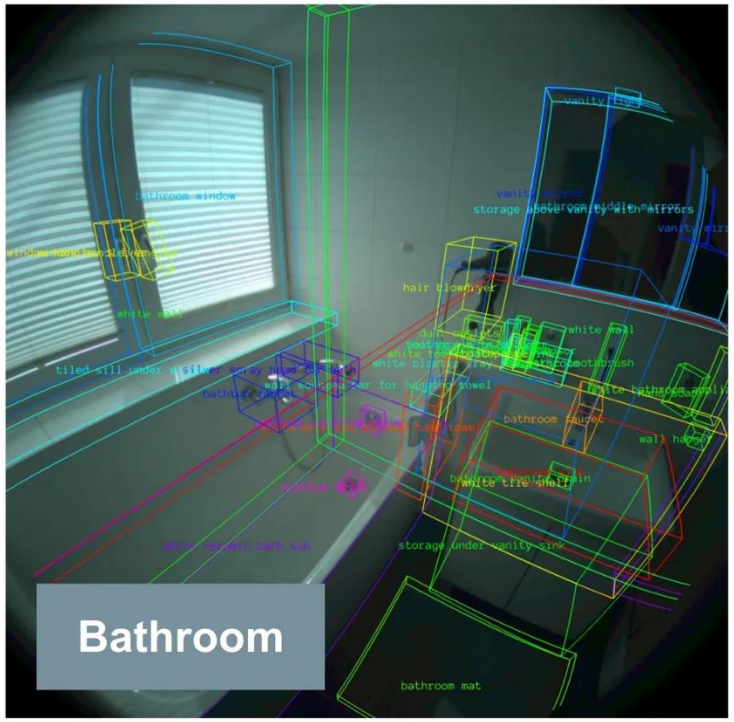
Egocentric Open-World Scene Understanding

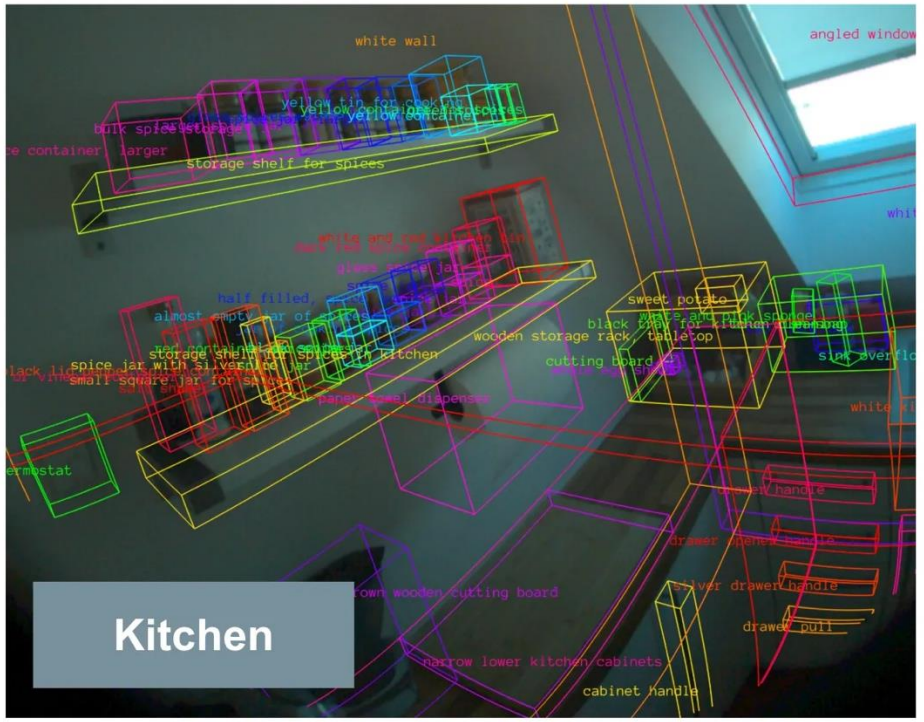
- Closed taxonomy 3D object detection (EFM3D, Arxiv 2024).
- Lifting 2D open-world foundation models to 3D.
  - 2D Segmentation “Point Painting” on Sparse Point Cloud.
  - 2D Segmentation lifting via Gaussian Splats (EgoLifter, ECCV 2024).
  - 2D to 3D Bounding Box Lifting. (Sneak Peak).
- Self-supervised 3D foundation models (Sonata CVPR 2025).

Benchmarking Open-World 3D Object Detection

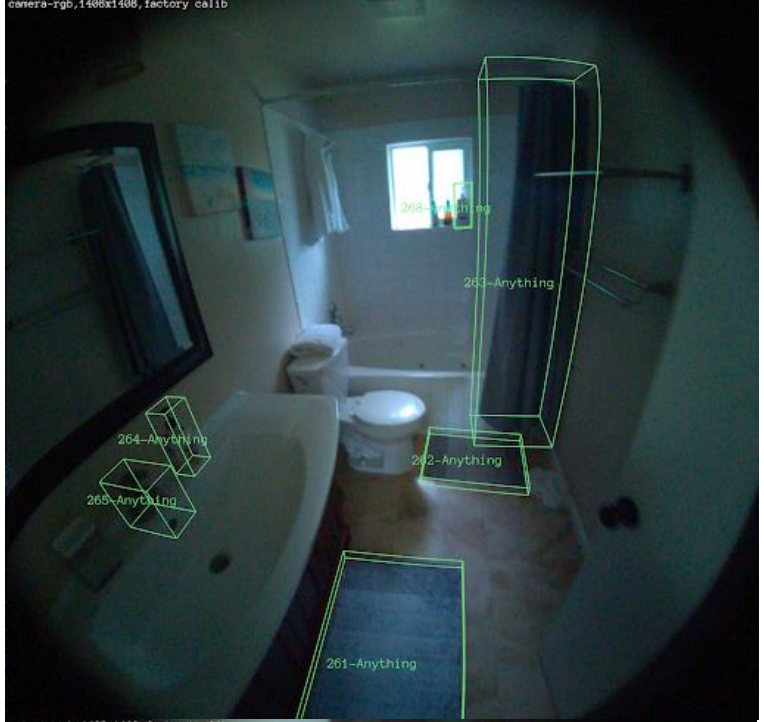




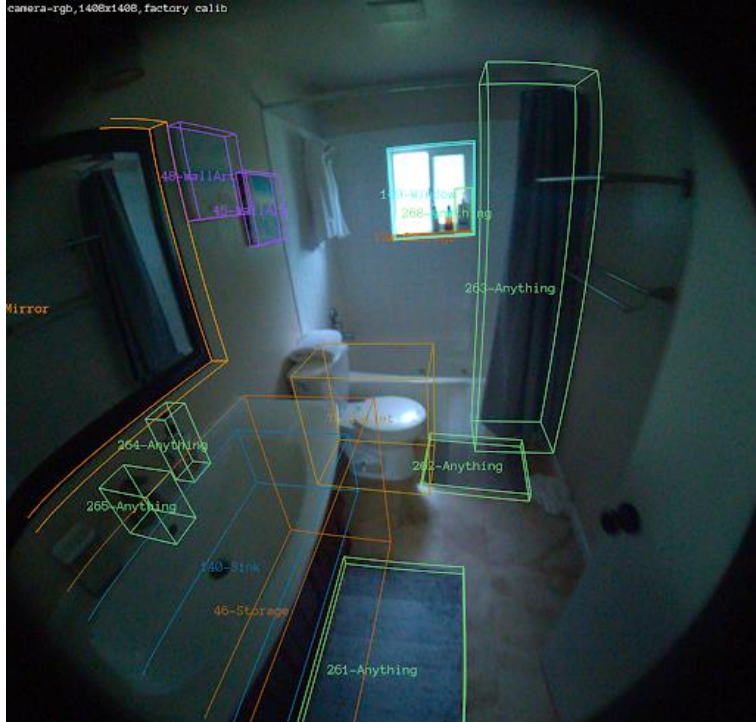


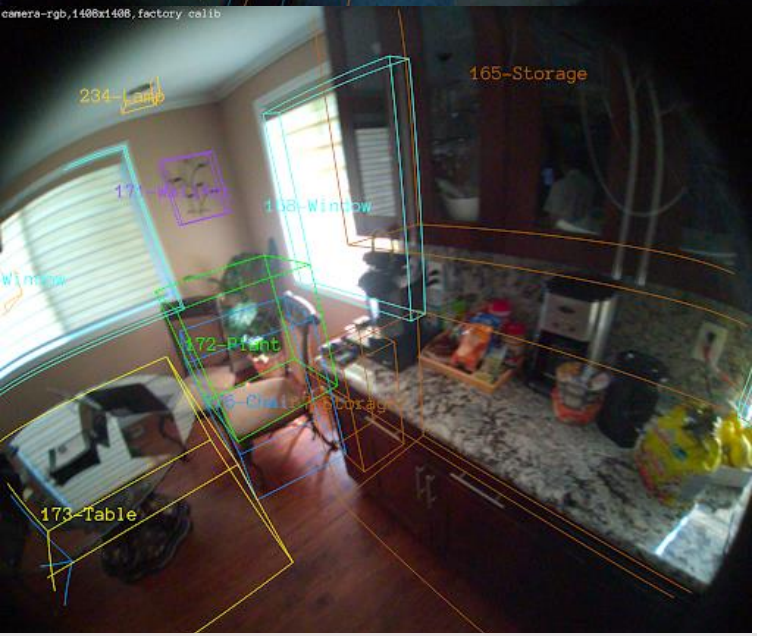


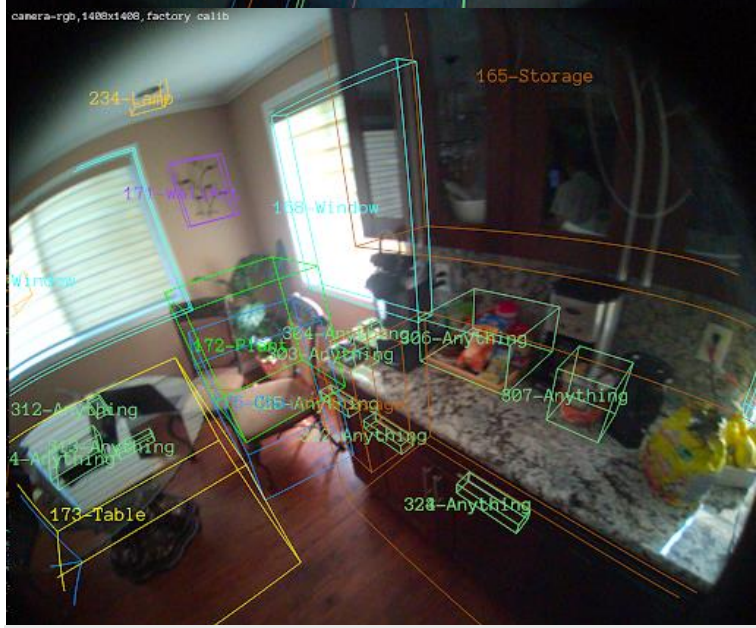
+



=



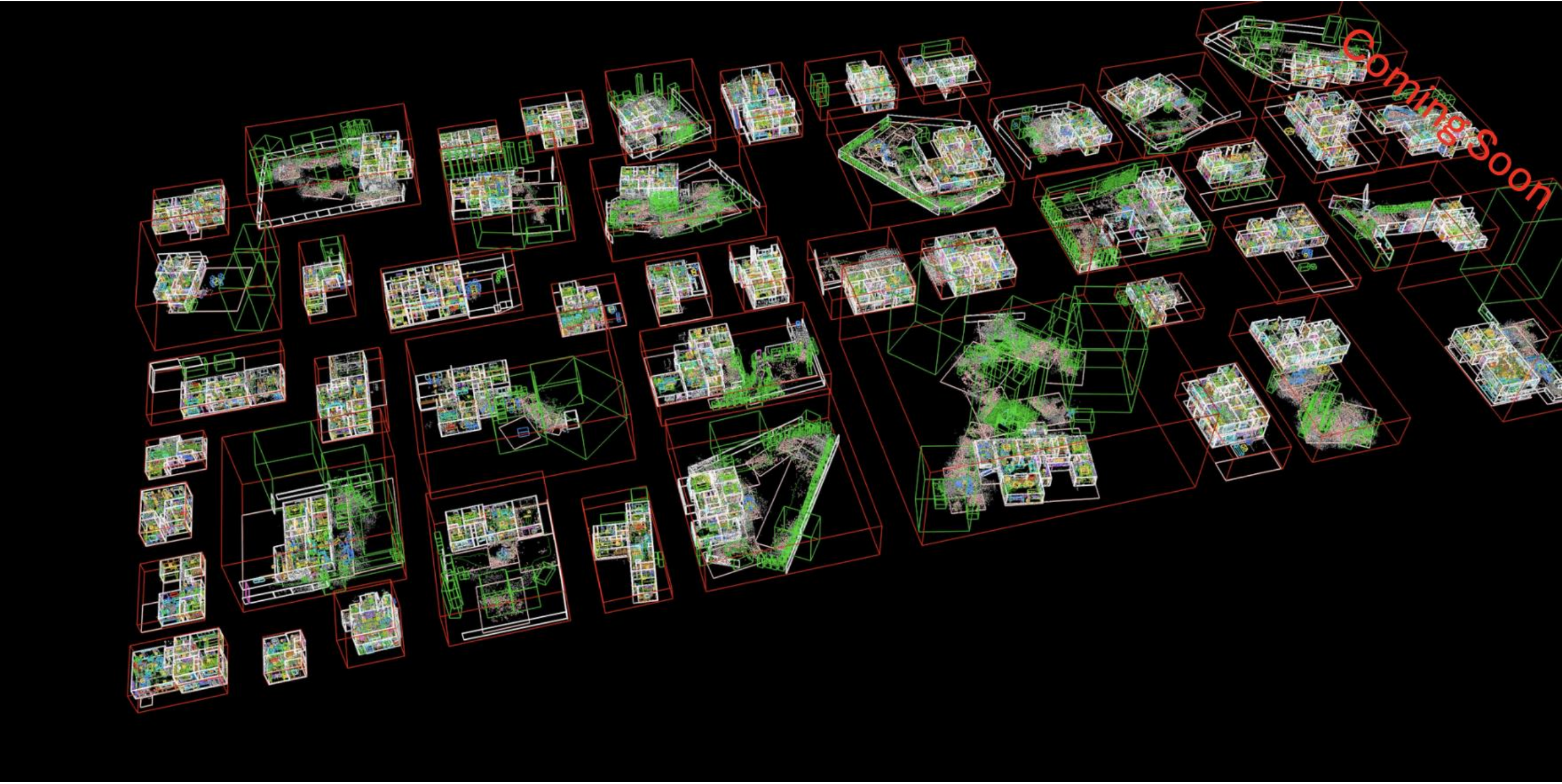




Closed-Taxonomy OBBs

Open-World OBBs

All OBBs

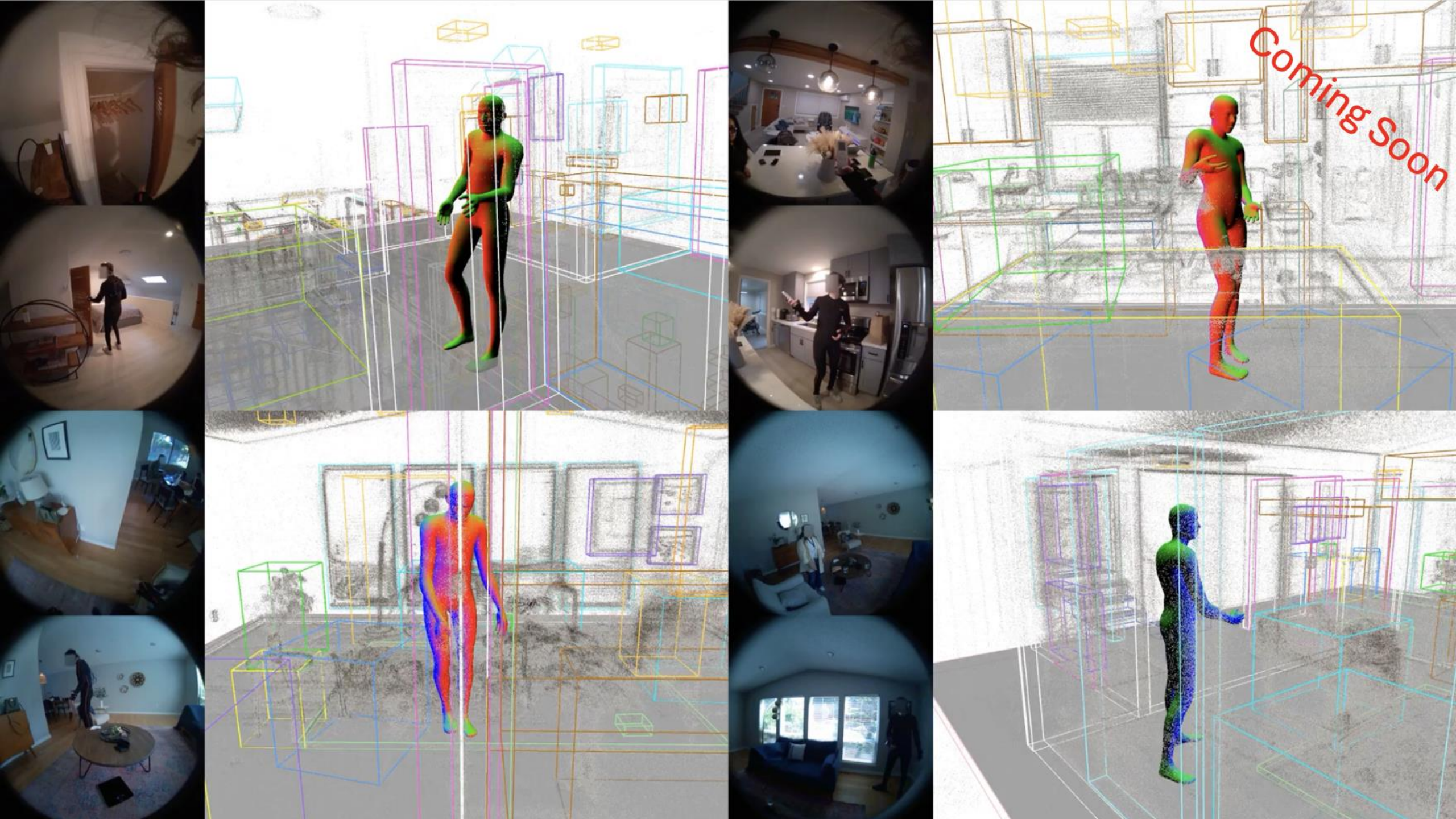


Nymeria++: Human Motion & 11k Closed Taxonomy + 10k Open-World Objects

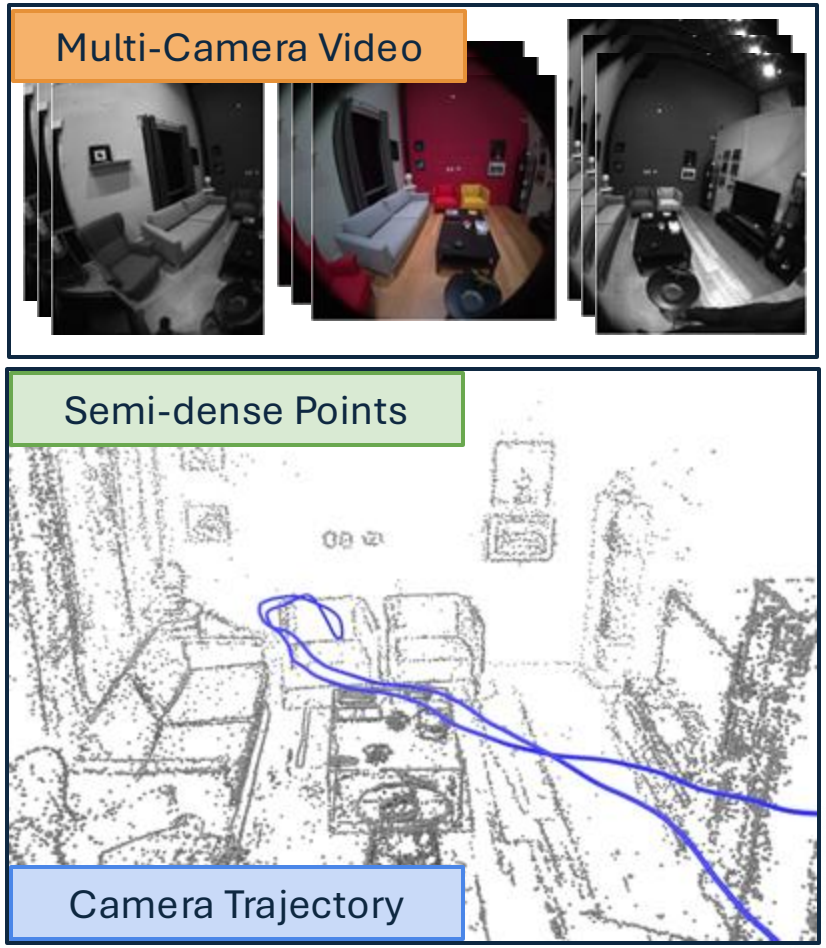
Nymeria++

Closed and Open-World OBBs for Egocentric Human Motion Dataset

Coming Soon


Coming Soon

Conclusion: Egocentric Data is a New Category of Data



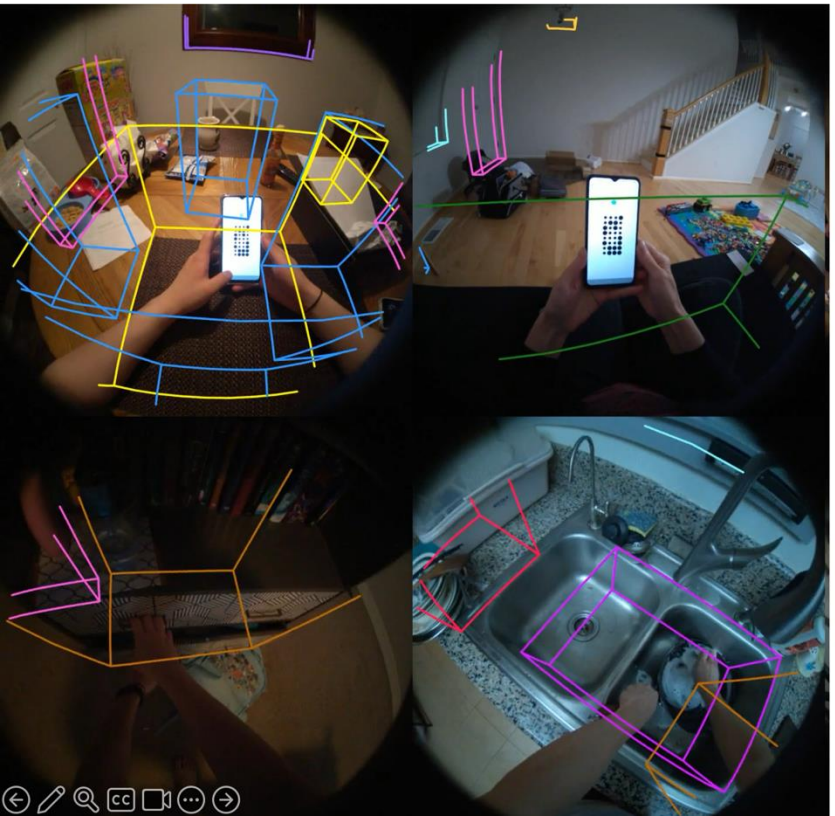
Egocentric Data

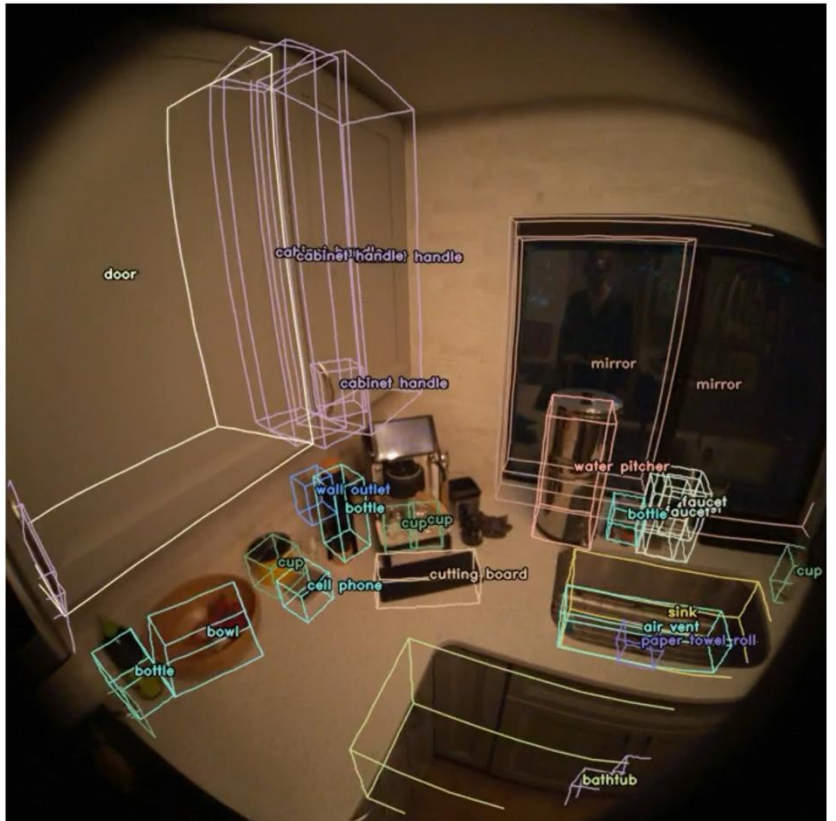
Project Aria Gen 1 Device

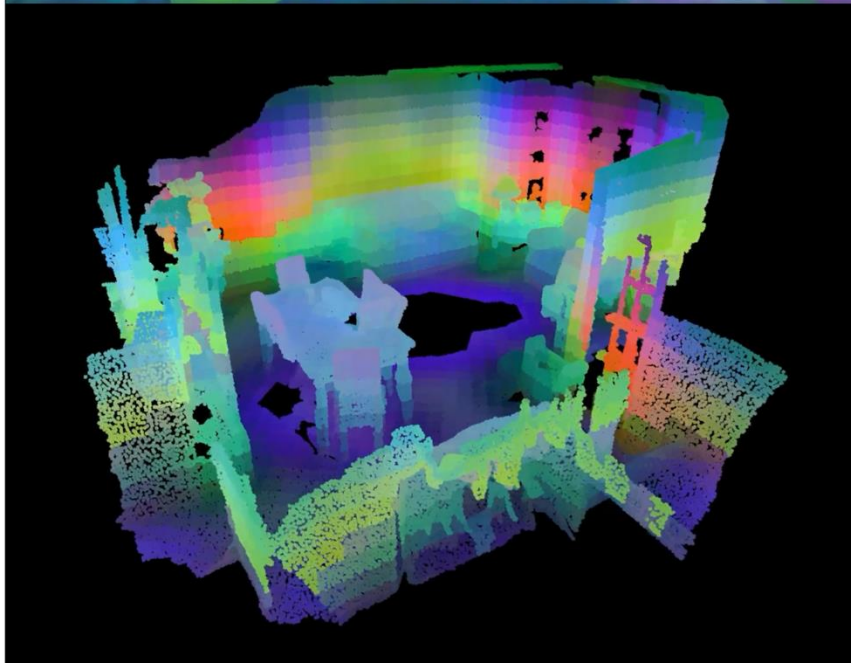
Project Aria Gen 2 Device

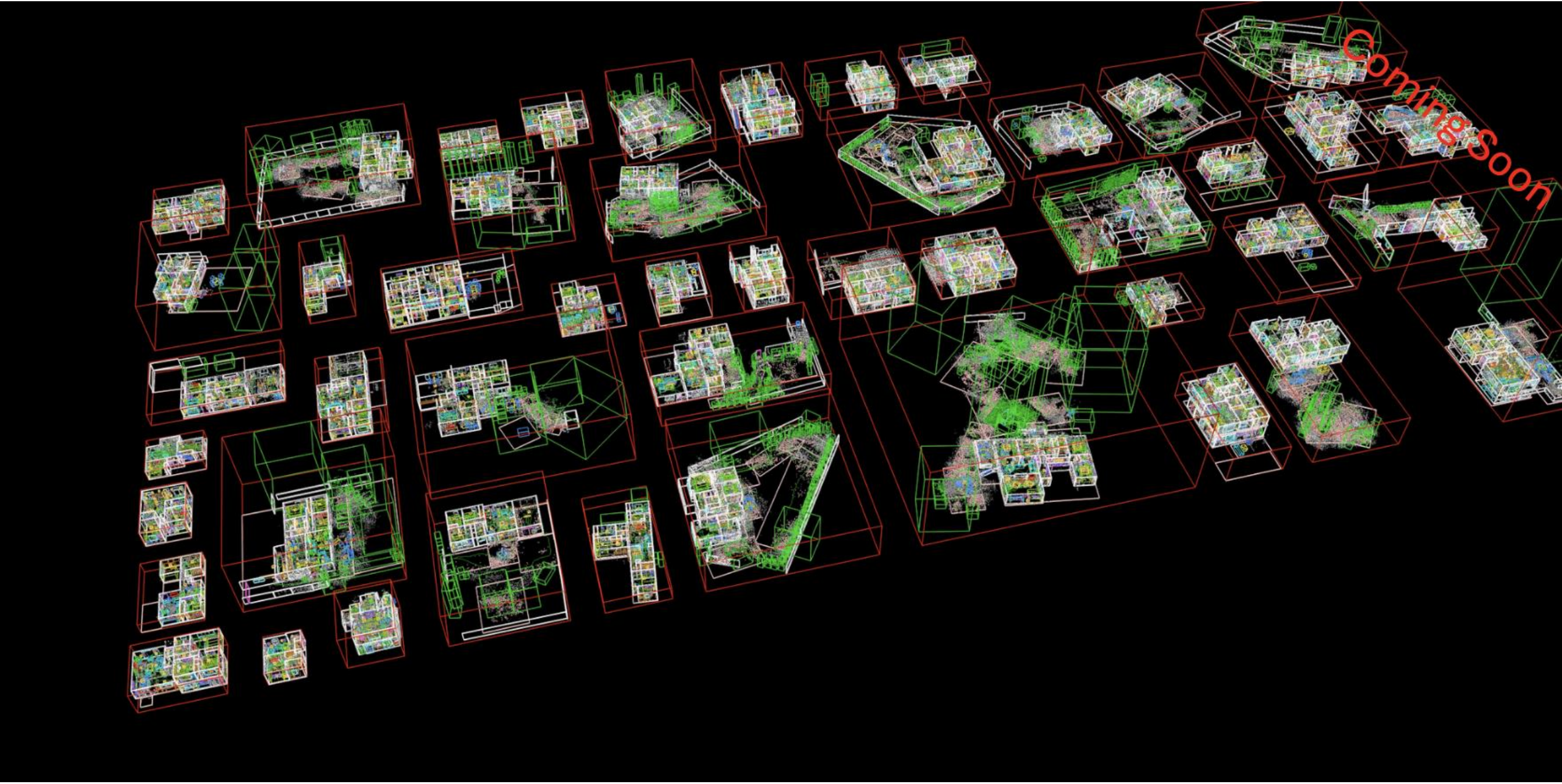
https://www.projectaria.com

- Robust Closed-Taxonomy Object Detection on Egocentric data. (EFM3D, Arxiv)
- 2D Foundation Model lifting using sparse points or Gaussian Splats. (EgoLifter 2024).
- 3D self supervised foundation model Sonata (CVPR 2025)



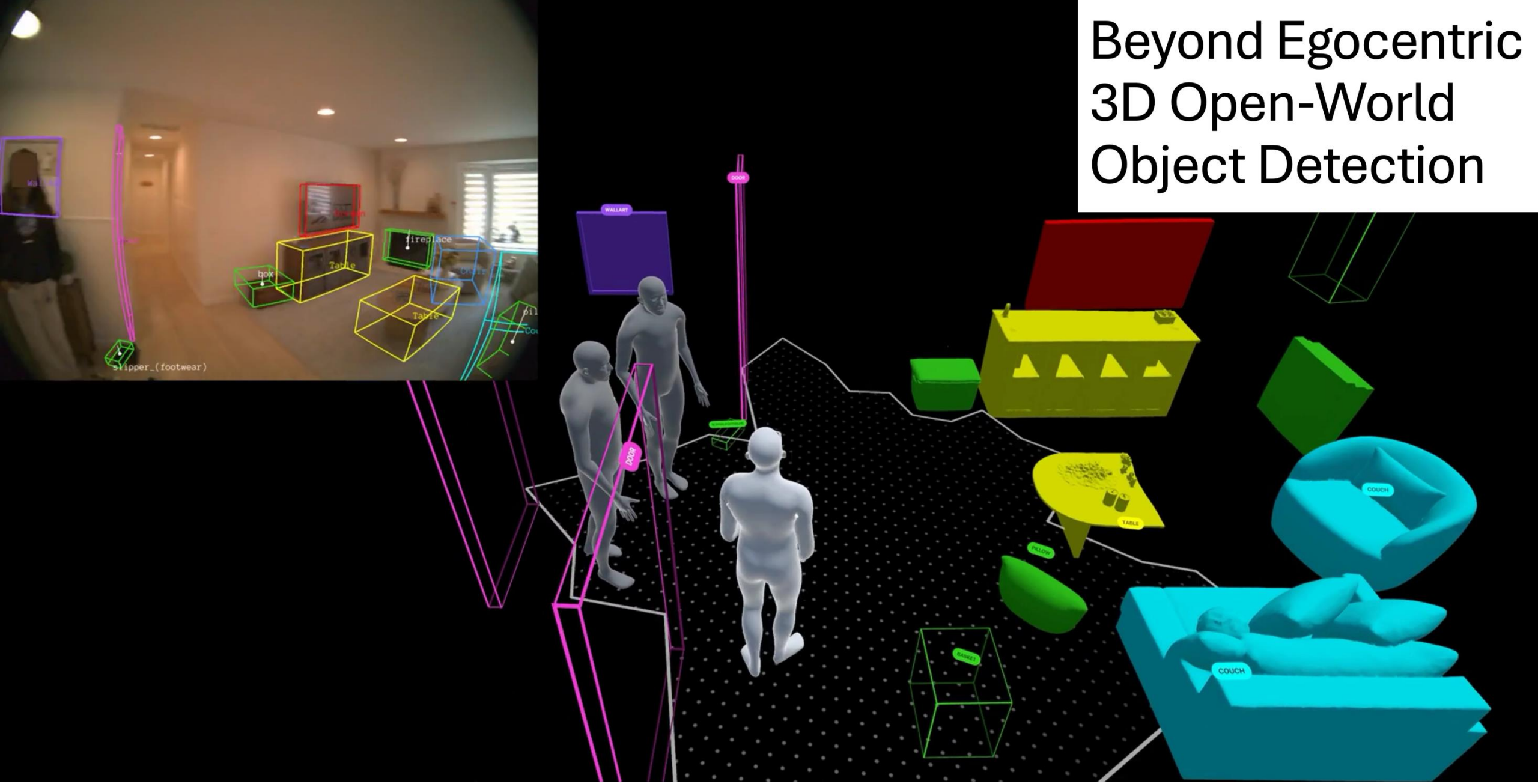




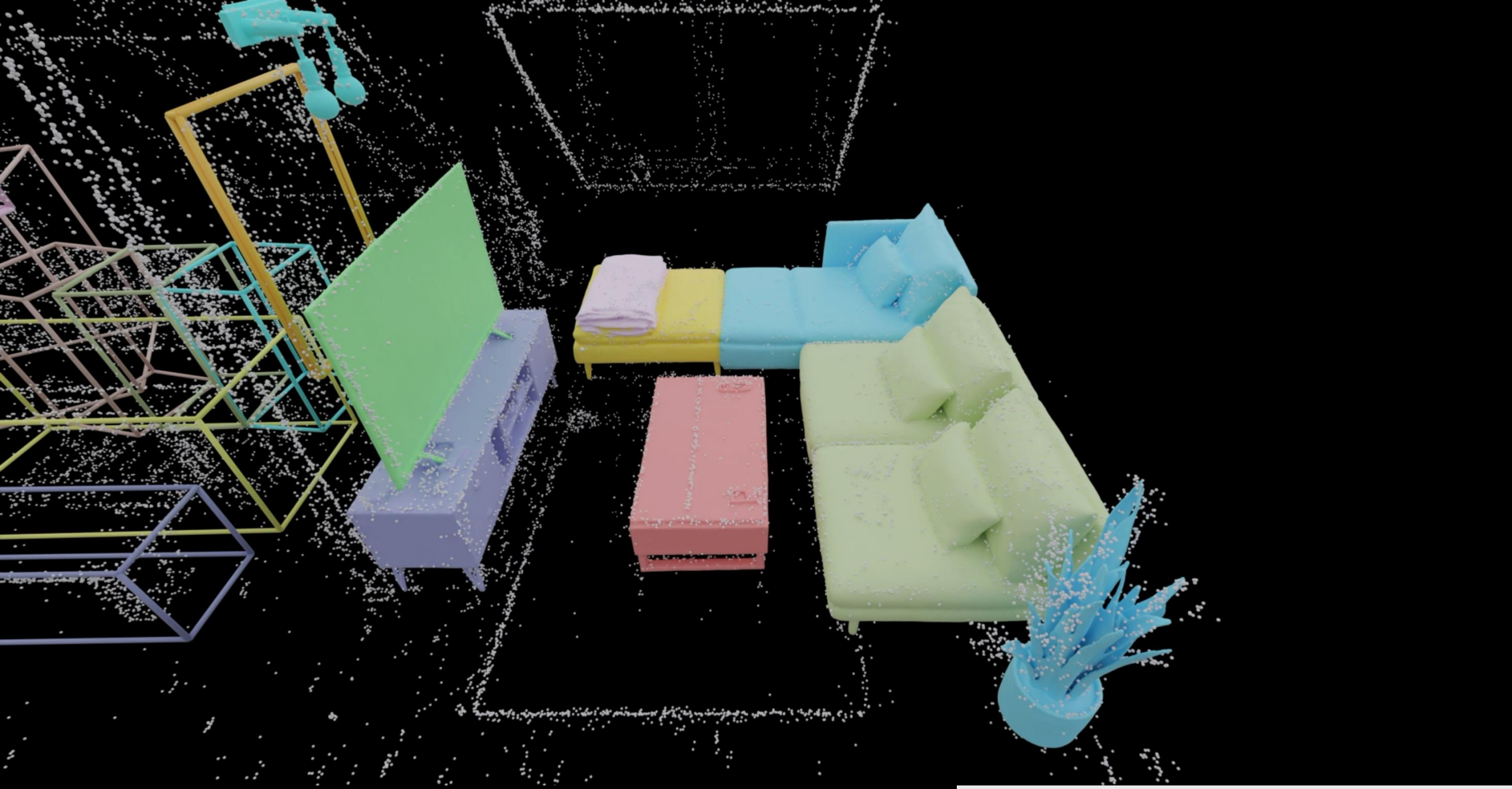


Nymeria++: Human Motion & 11k Closed Taxonomy + 10k Open-World Objects

Beyond Egocentric 3D Open-World Object Detection

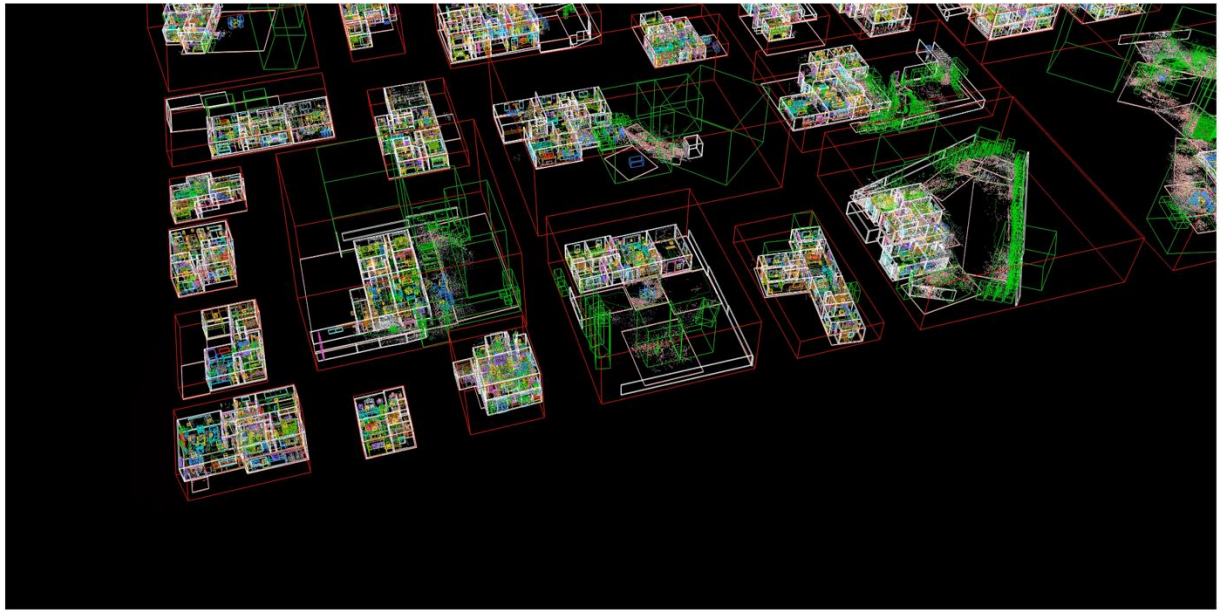


4th Hands on Egocentric Research Tutorial with Project Aria (8am-12pm on Monday)
EgoMotion Workshop (1pm-5pm on Monday)

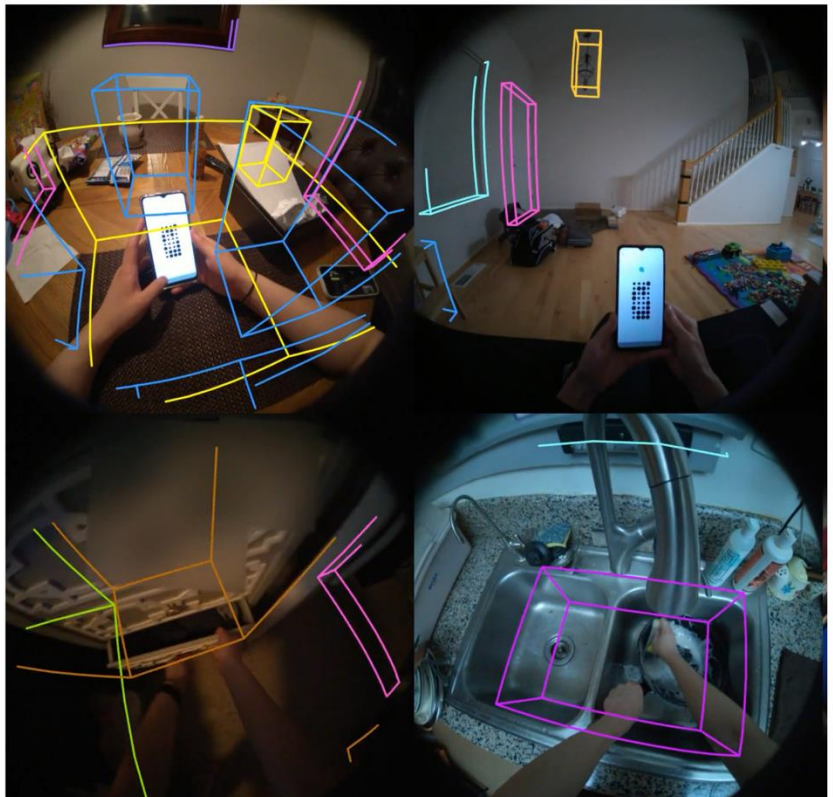


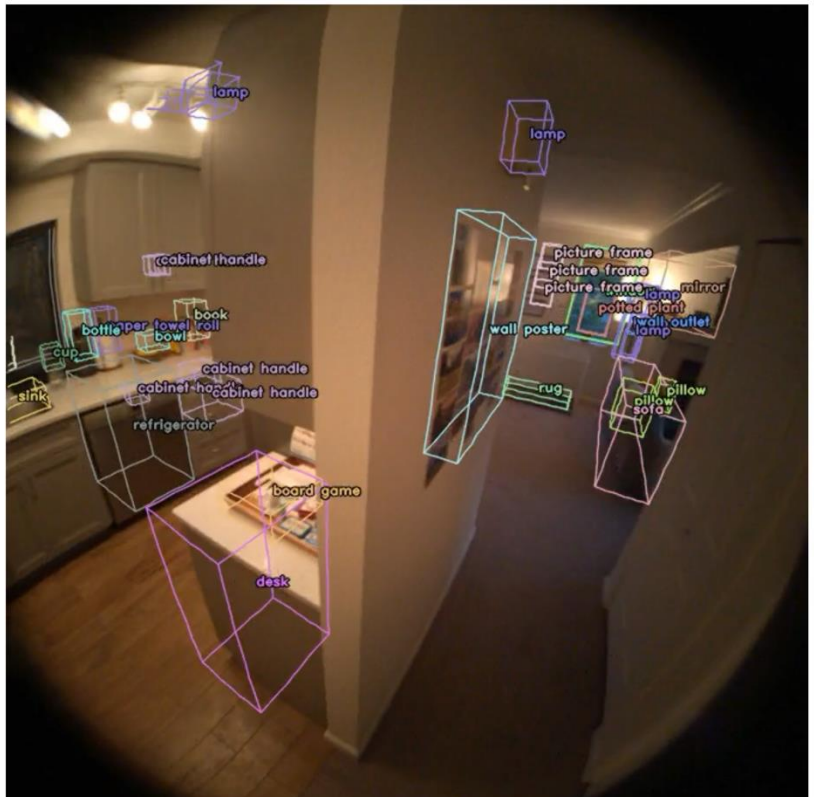
Siddiqui, Yawar, et al. *“ShapeR: Robust Conditional Shape Generation from Casual Captures”,* Coming Soon

Questions?




Feature PCA by EgoLifter




lamp
lamp
cabinet handle
picture frame
picture frame
picture frame
mirror
lamp
potted plant
wall outlet
lamp
wall poster
book
bottle
paper towel roll
bowl
cup
cabinet handle
cabinet handle
cabinet handle
rug
pillow
pillow
sofa
sink
refrigerator
board game
desk