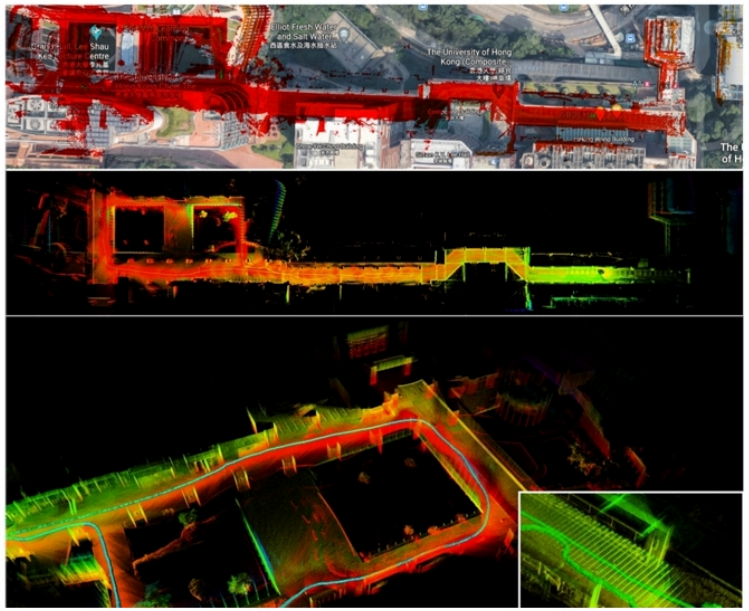
Language-Based Scene Understanding for Mobile Robot

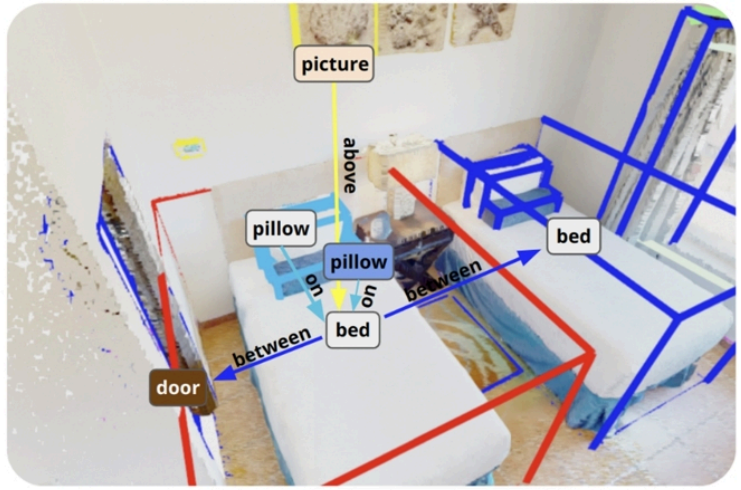
ROAS6800H Seminar

Yiming ZHU, MPhil's Student
ROAS, Systen Hub
2025/04/01

Contents

Background



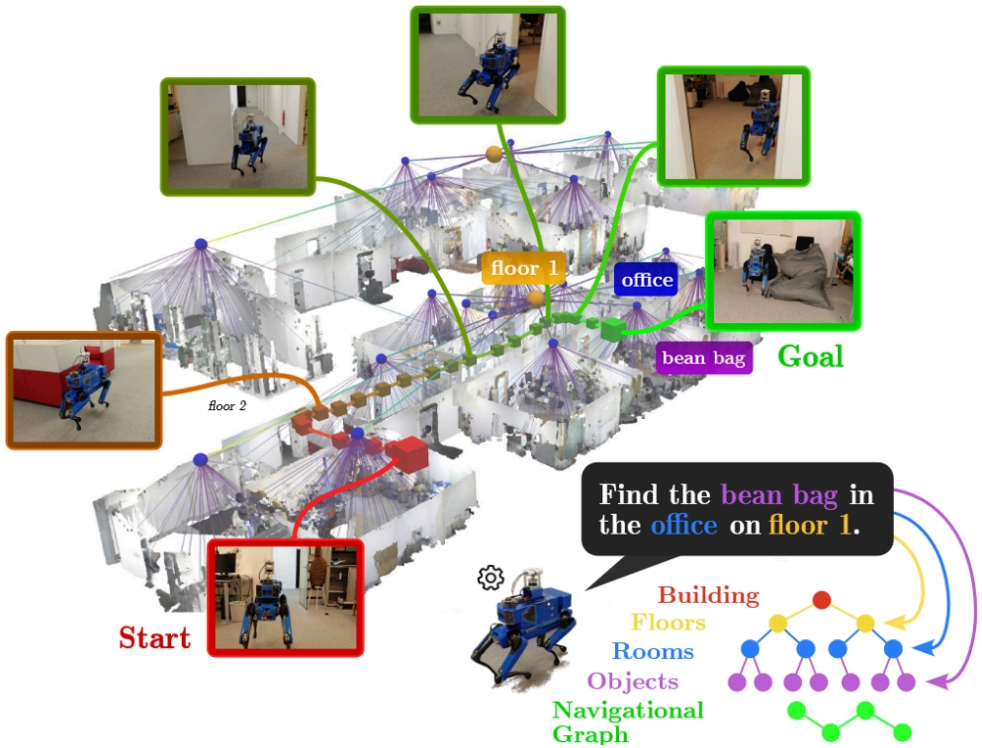


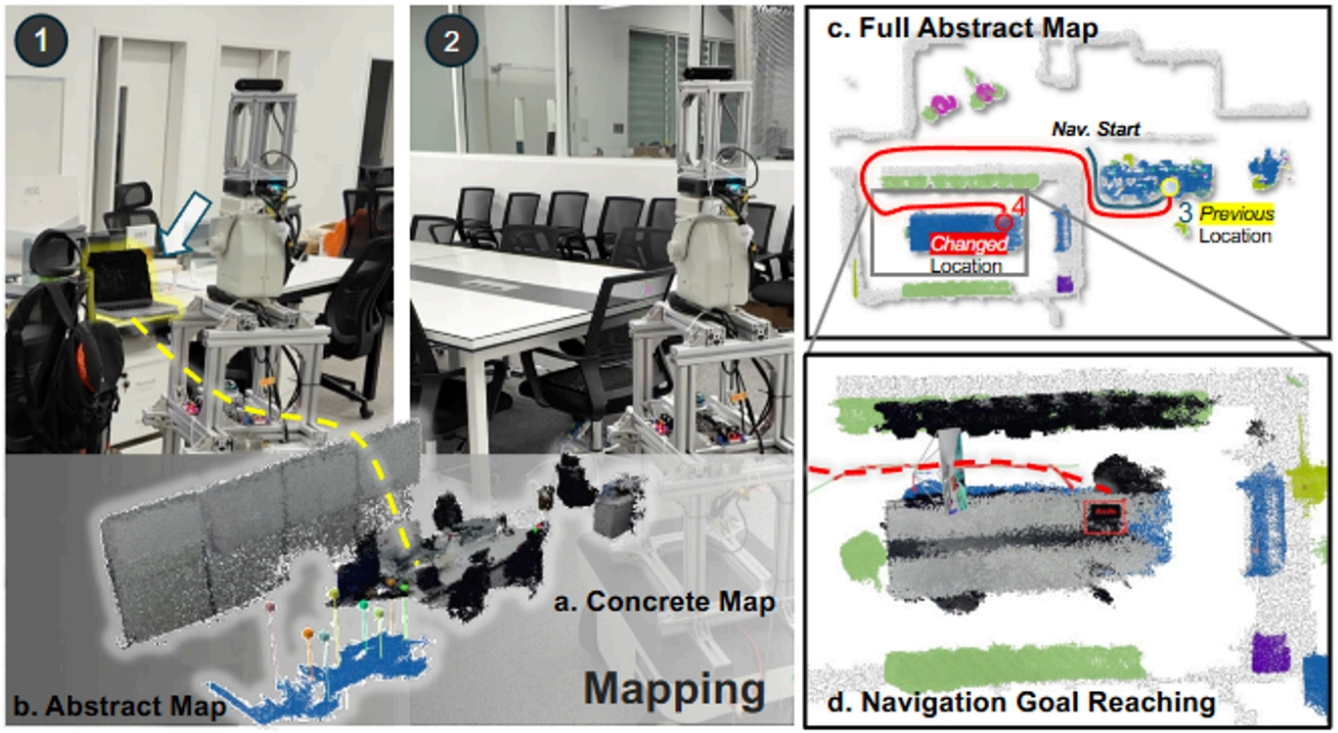
(a) Robust Odometry[1] (b) Complex Scenes[2] (c) Rich Annotations[3]

1: Fast-lio https://ieeexplore.ieee.org/abstract/document/9372856/
2: Habitat3.0 https://arxiv.org/abs/2310.13724
3: VLA-3D https://arxiv.org/abs/2411.03540

Background

- Mainstream methods[1] use scene graph to represent a scene.
- Our work DualMap[2] aims at real-time language navigation.





1: HOV-SG https://arxiv.org/abs/2403.17846
2: under review

What is SceneScript?

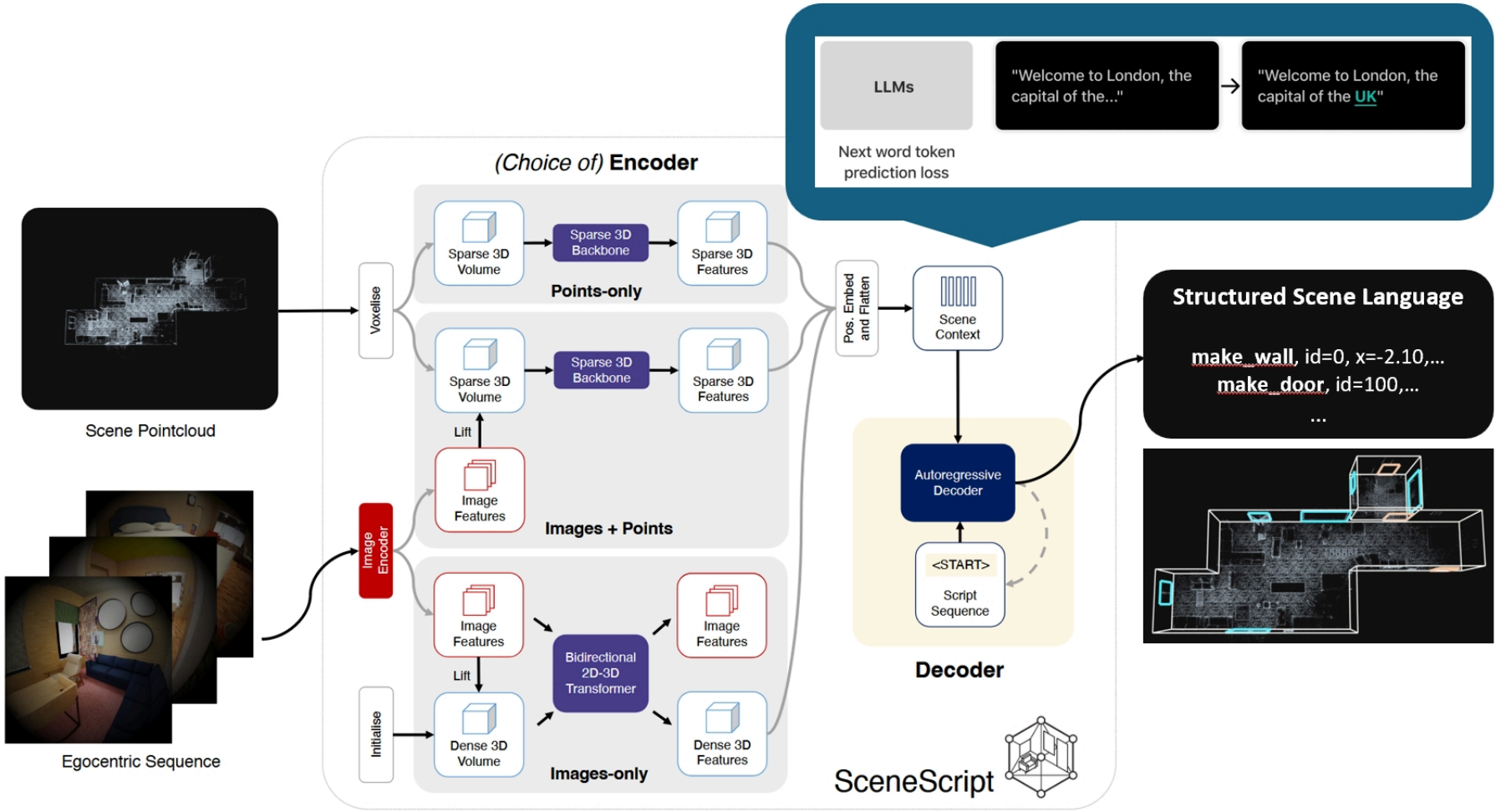
SceneScript (Meta Reality Labs, ECCV24') is a novel method of **end-to-end** representing scenes using **language**.

```
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)
```

Structured scene language

How is SceneScript trained?

Similar to next token prediction in LLMs.



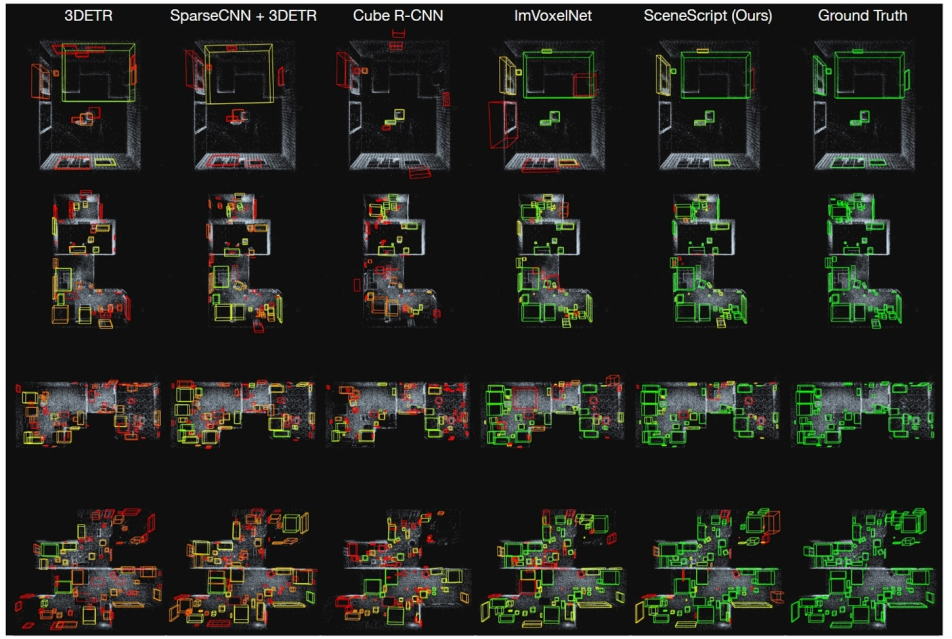
How is SceneScript trained?

- 3D scene data is not as accessible as web texts for LLM training.
- Synthetic dataset Aria Synthetic Environments with 100K scenes.

Experimental Results

The authors tested the performance of SceneScripts on layout estimation & object detection.

Method	F1 @5cm mean	wall	door	window	Avg F1 mean	wall	door	window
SceneCAD '20 [2]	-	0.048	-	-	-	0.275	-	-
RoomFormer '23 [52]	0.139	0.159	0.148	0.110	0.464	0.505	0.481	0.407
Ours (Point cloud)	0.848	0.930	0.922	0.692	0.784	0.816	0.811	0.724
Ours (Lifted features)	**0.903**	**0.943**	**0.959**	**0.806**	**0.801**	**0.818**	**0.822**	**0.764**
Ours (Image-only)	0.661	0.687	0.798	0.497	0.719	0.727	0.772	0.658



(a) *Aria Synthetic Environments*

Method	Input	F1 @.25 IoU	@.50 IoU
3DETR '21 [26]	Points	0.201	0.078
Cube R-CNN '23 [4]	RGB	0.394	0.228
ImVoxelNet '22 [36]	RGB	0.584	0.516
Ours	Points	**0.620**	**0.577**

(b) ScanNet [10]

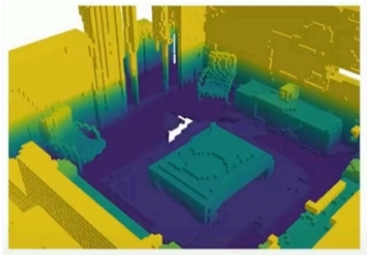
Method	Input	F1 @.25 IoU	@.50 IoU
3DETR '21 [26]	Points	0.480	0.349
3DETR-m '21 [26]	Points	0.536	0.407
SoftGroup '22 [46]	RGB Points	**0.622**	**0.573**
Ours	RGB Points	0.506	0.406

Why SceneScript?

Low memory, extensibility and integration with language models.

```
make_bbox: id, class, position_x, position_y,
    position_z, angle_z, scale_x, scale_y, scale_z
```

Mesh/PCD

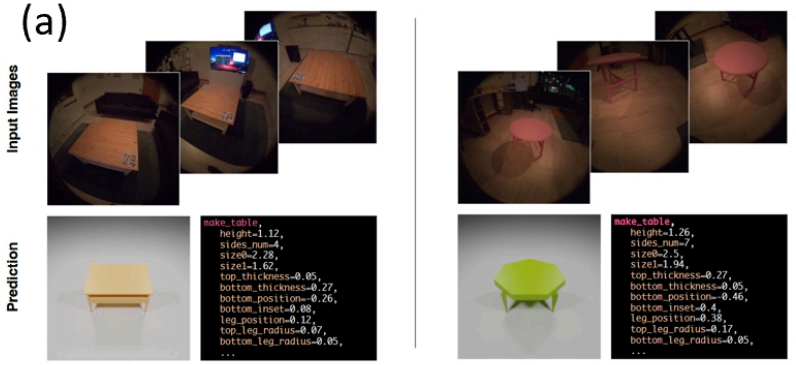


Voxel

Neural SDF

Compact

(a)



(b)

```
make_door: id, wall_id, pos_x, pos_y, pos_z, width,
    height, open_degree, hinge_side, open_direction
```

(c)

Original | Decomposed Primitives | Original | Decomposed Primitives

Extensible

"Hey Meta, how many pots of paint would it take to paint this room?"

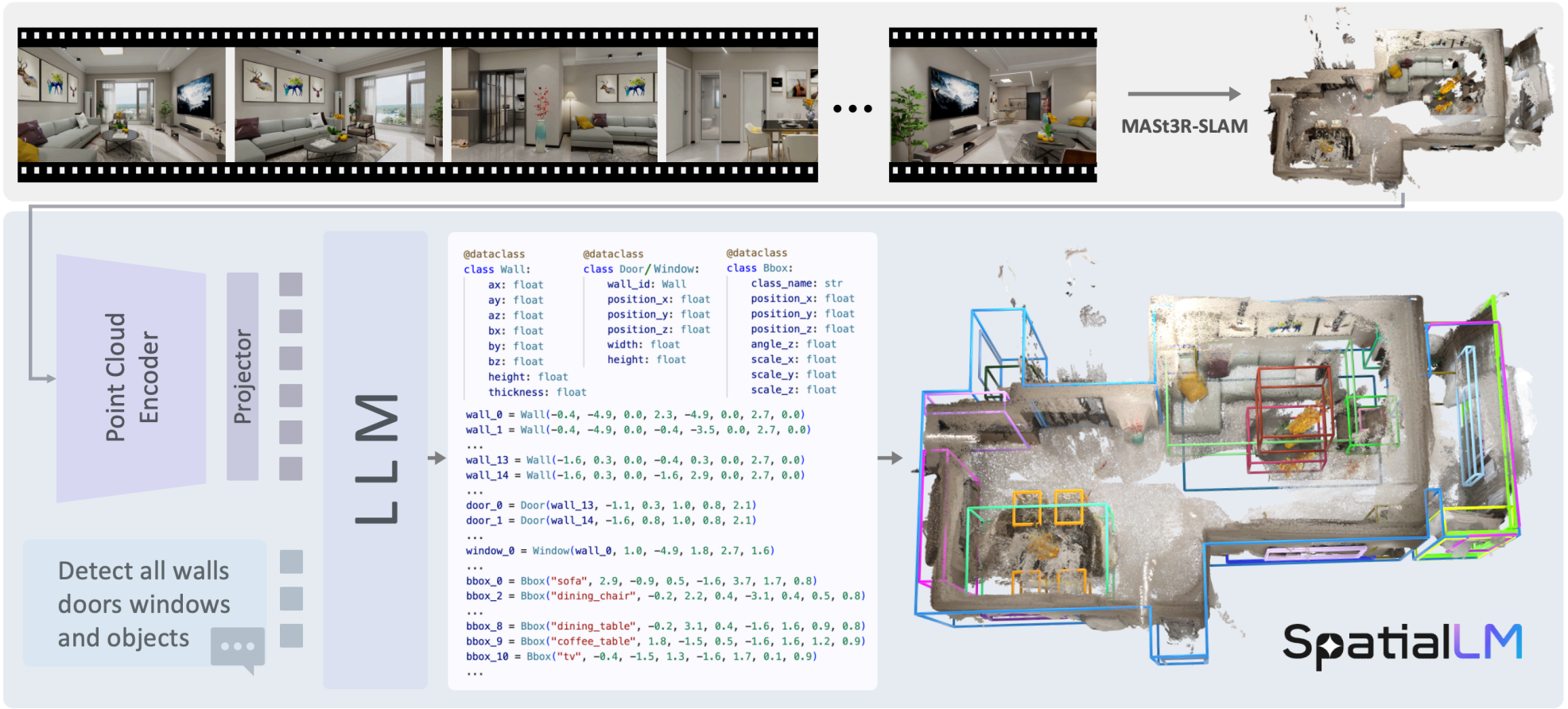
"Hey Meta, place an augment on the large table"

- make_wall, id=0, a_x=2.1, a_y=3.9, a_z=0.0, b_x=7.8, b_y=3.9, b_z=0.0, height=2.7
- make_wall, id=1, a_x=7.8, a_y=3.9, a_z=0.0, b_x=7.8, b_y=0.3, b_z=0.0, height=2.7
- make_wall, id=2, a_x=7.8, a_y=0.3, a_z=0.0, b_x=2.1, b_y=0.3, b_z=0.0, height=2.7
- make_wall, id=3, a_x=2.1, a_y=0.3, a_z=0.0, b_x=2.1, b_y=3.9, b_z=0.0, height=2.7
- make_door, id=4, wall0_id=1, wall1_id=-1, position_x=7.8, position_y=1.5, position_z=1.0,
- make_window, id=5, wall0_id=2, wall1_id=-1, position_x=5.3, position_y=0.3, position_z=1.4,
- make_window, id=6, wall0_id=3, wall1_id=-1, position_x=2.1, position_y=2.1, position_z=1.4,

LLM

Application

SpatialLM uses the pointcloud encoder of SceneScript. It validates the SceneScript's compatibility with LLM.



Application

- Spatia"lLM" means Large Language Model, which is **over-claimed**.
- But it inspires us about future directions.

Reconstruction & Interaction

Reasoning

Block Test

This is a test for blocks.

Key Point

The next generation of robots will need to build **metric-semantic** representations of environments, enriched with **inter-object relations** and **human-in-the-loop** capabilities.

They looks all set.

Thank you!

* Slides made with Marp and kaisugi.