

Embodied AI: Language and Perception

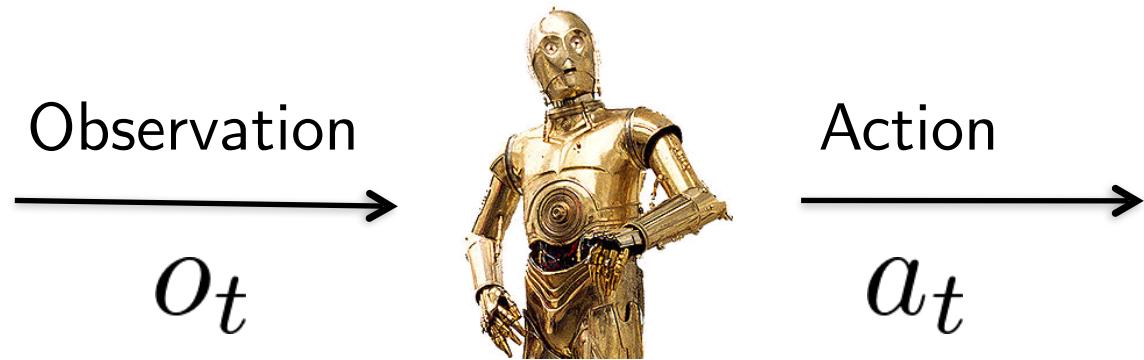
Russ Salakhutdinov

Machine Learning Department
Carnegie Mellon University

Carnegie
Mellon
University

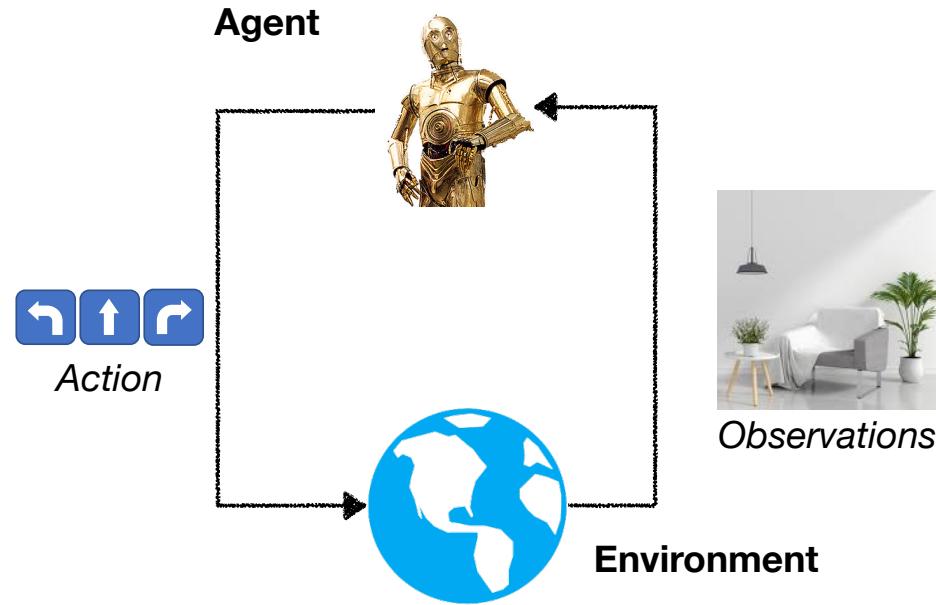


Learning Behaviors



Learning to map sequences of observations to actions,
for a particular goal

Physical Intelligence

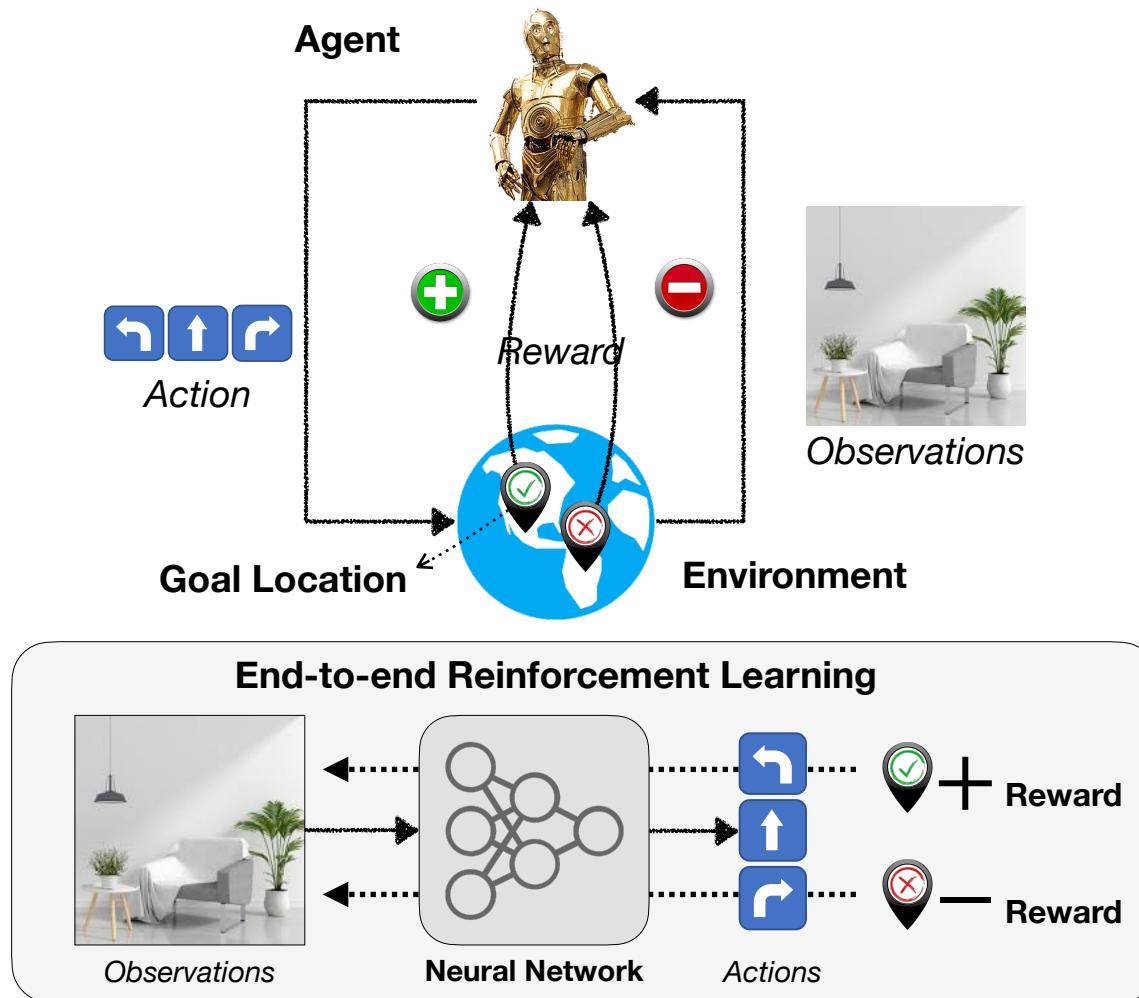


Agent needs to move in the world physically.

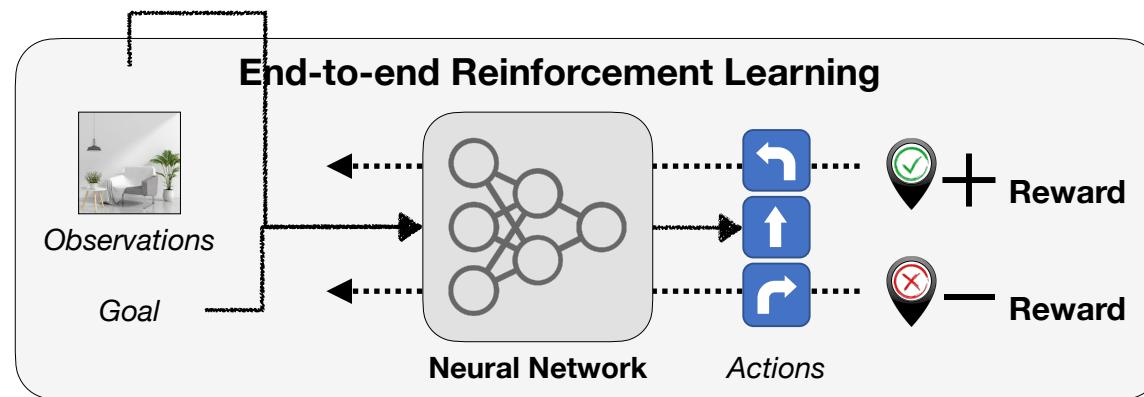
Current actions affect future observations.

Require Spatial and Semantic Understanding.

Navigation



Goal-conditioned Navigation



Point Goal

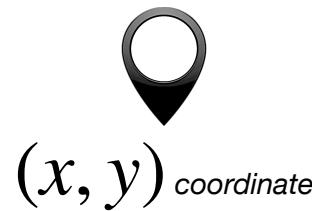


Image Goal



Object Goal

Chair
TV
Sofa

Language Goal

Blue Chair
Largest TV
White Sofa

- Convenient for humans
- Compositionality

Navigation Tasks

Point Goal

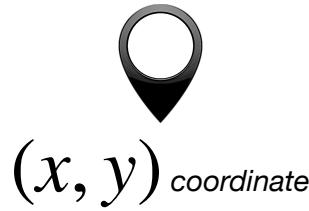


Image Goal

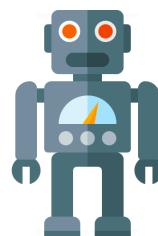


Object Goal

Chair
TV
Sofa

Language Goal

Blue Chair
Largest TV
White Sofa

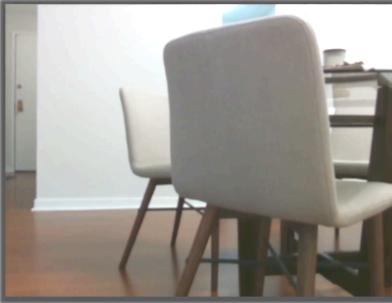


*Require exploring the environment
to find the goal*



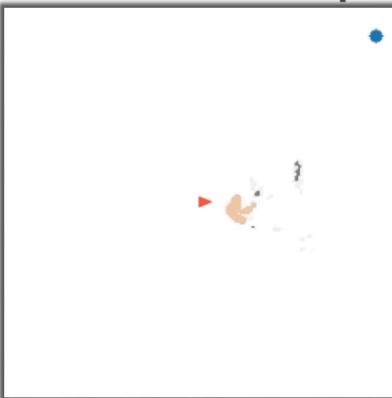
Real World: Object Goal Navigation

Observation



Goal: Potted Plant

Predicted
Semantic Map

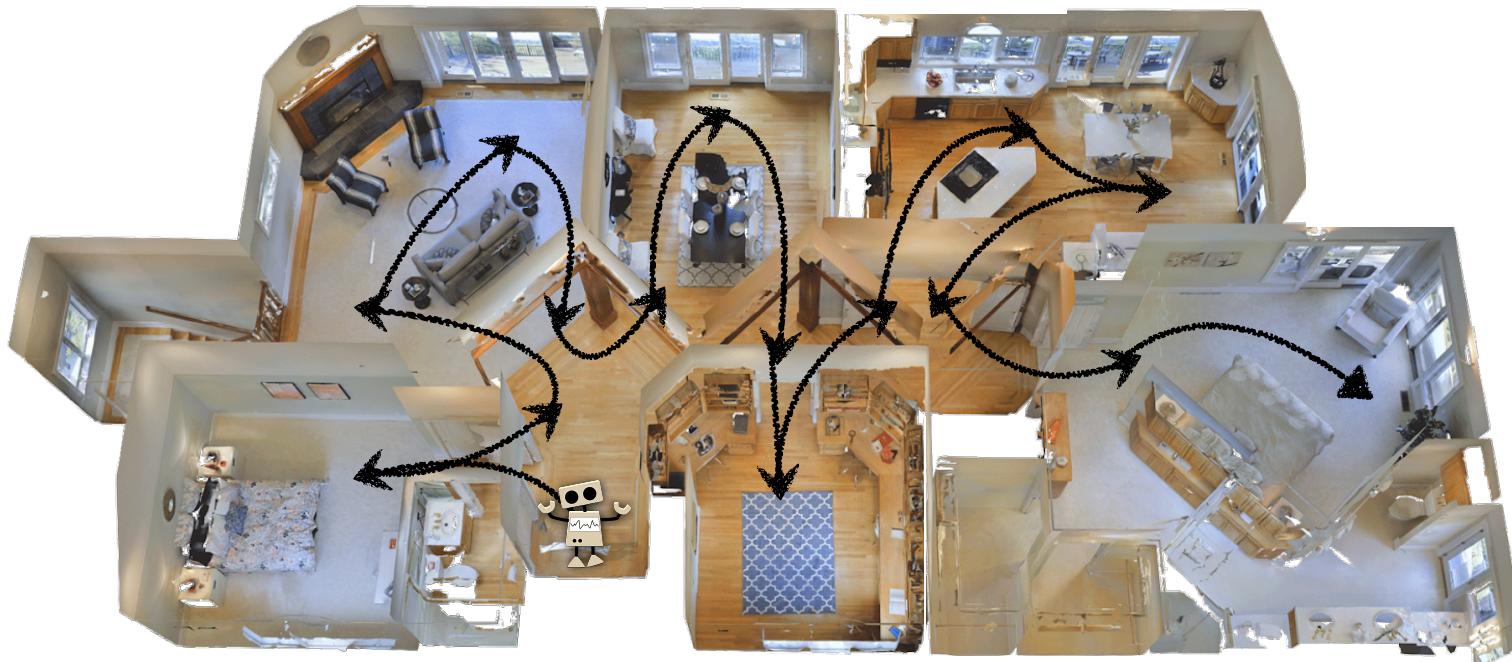


Third-person view



See video at: <https://devendrachaplot.github.io/projects/semantic-exploration>

Exploration



Exploration

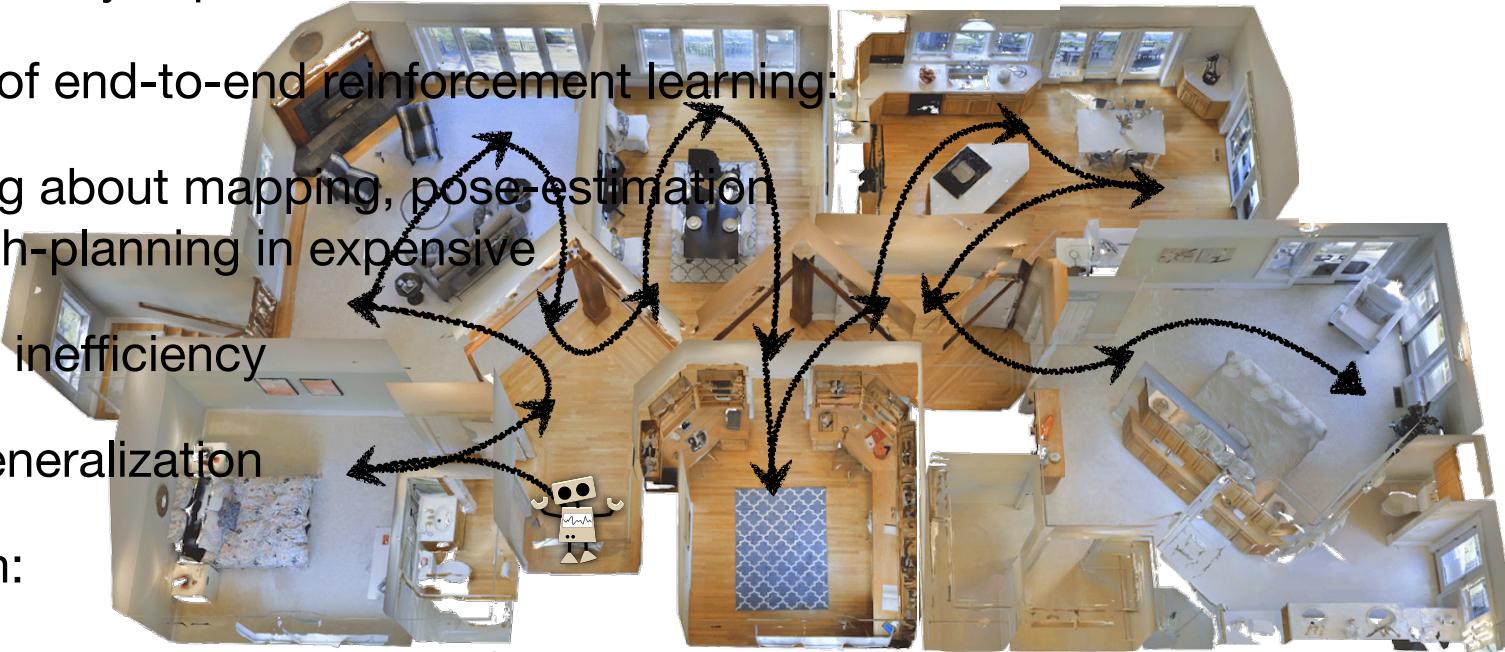
- How to efficiently explore an unseen environment?

- Limitations of end-to-end reinforcement learning:

- Learning about mapping, pose-estimation and path-planning is expensive
- Sample inefficiency
- Poor generalization

- Our solution:

- Incorporating the strengths of learning
- Modular and hierarchical system

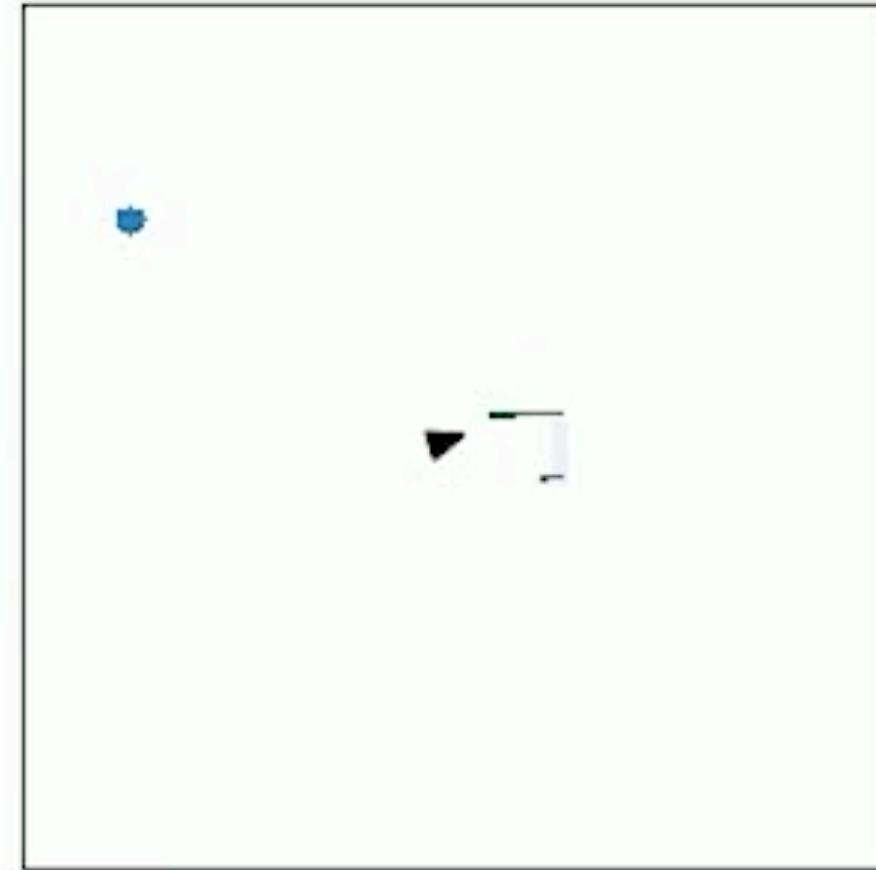


Preview: Visual Navigation in the Real World

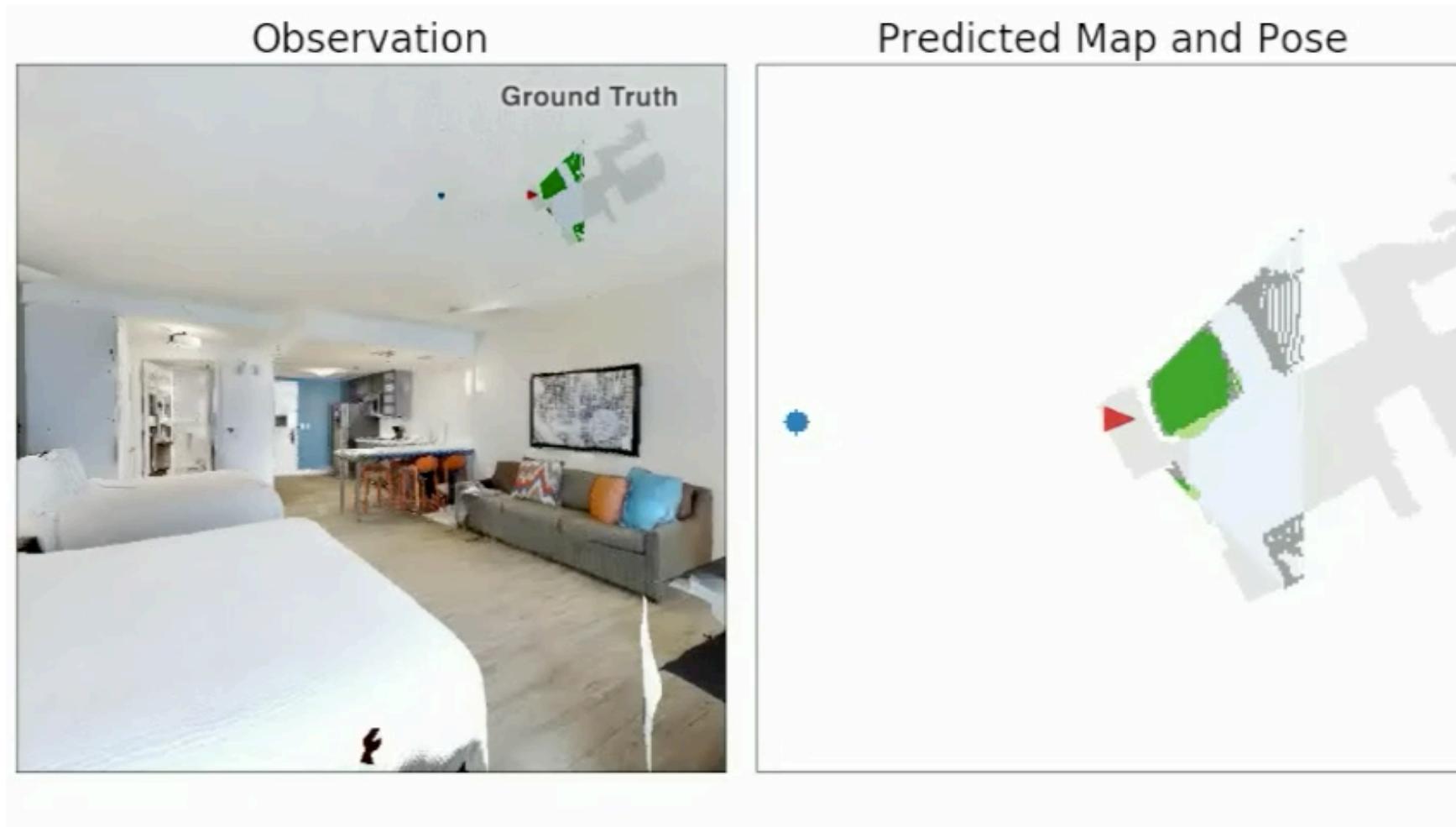
Observation



Predicted Map and Pose



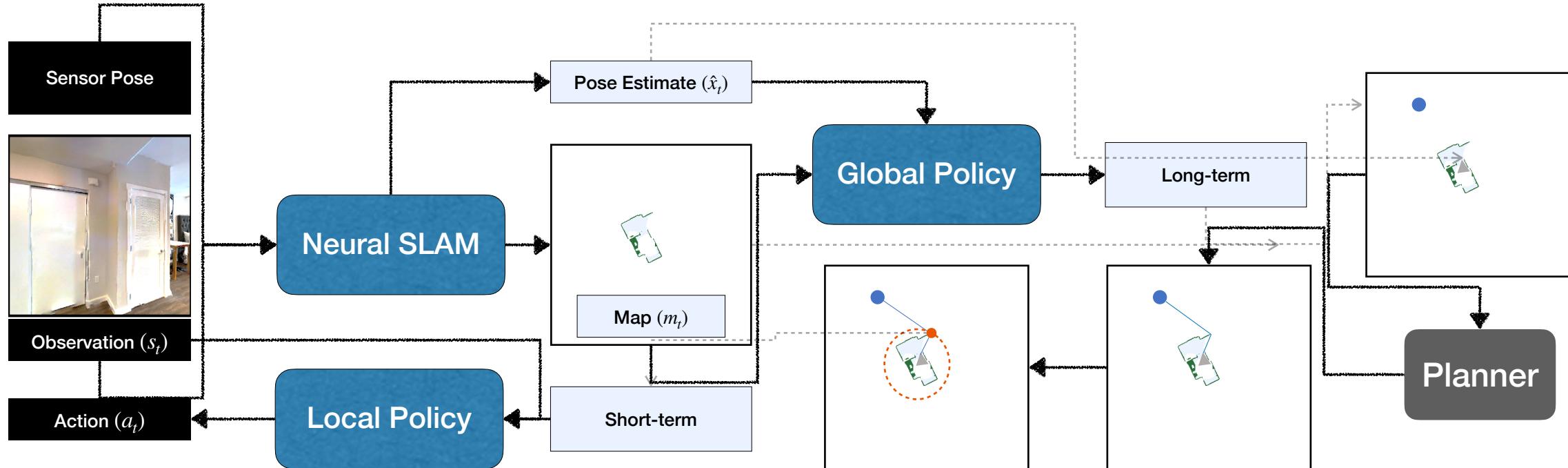
Exploration in Gibson Environment



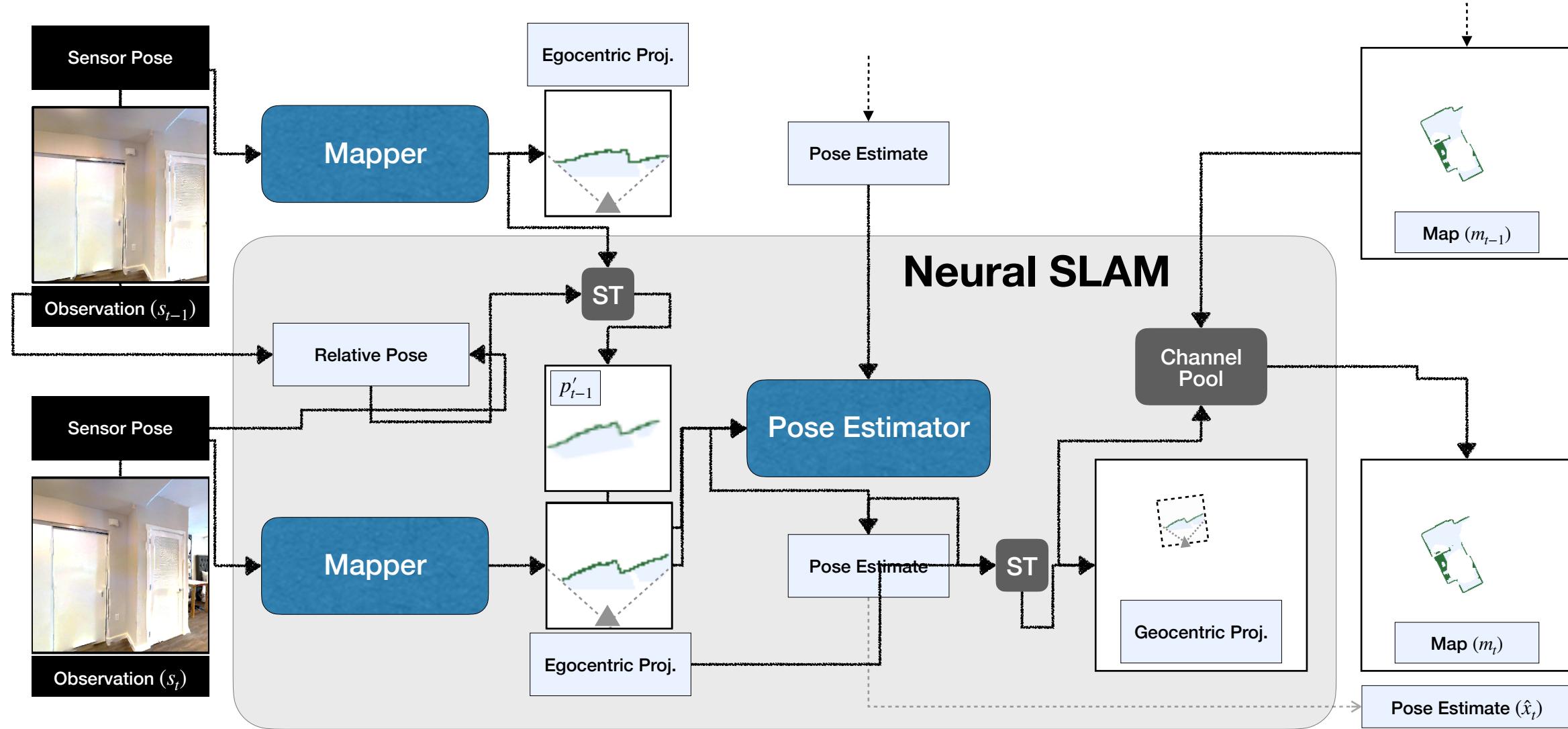
- Long-Term Goal
- Explored Area
- Correct Map prediction
- Incorrect Map prediction
- Agent Trajectory prediction
- ▶ Agent Pose prediction

- True Map
- ▶ Agent True Pose
- Agent True Trajectory

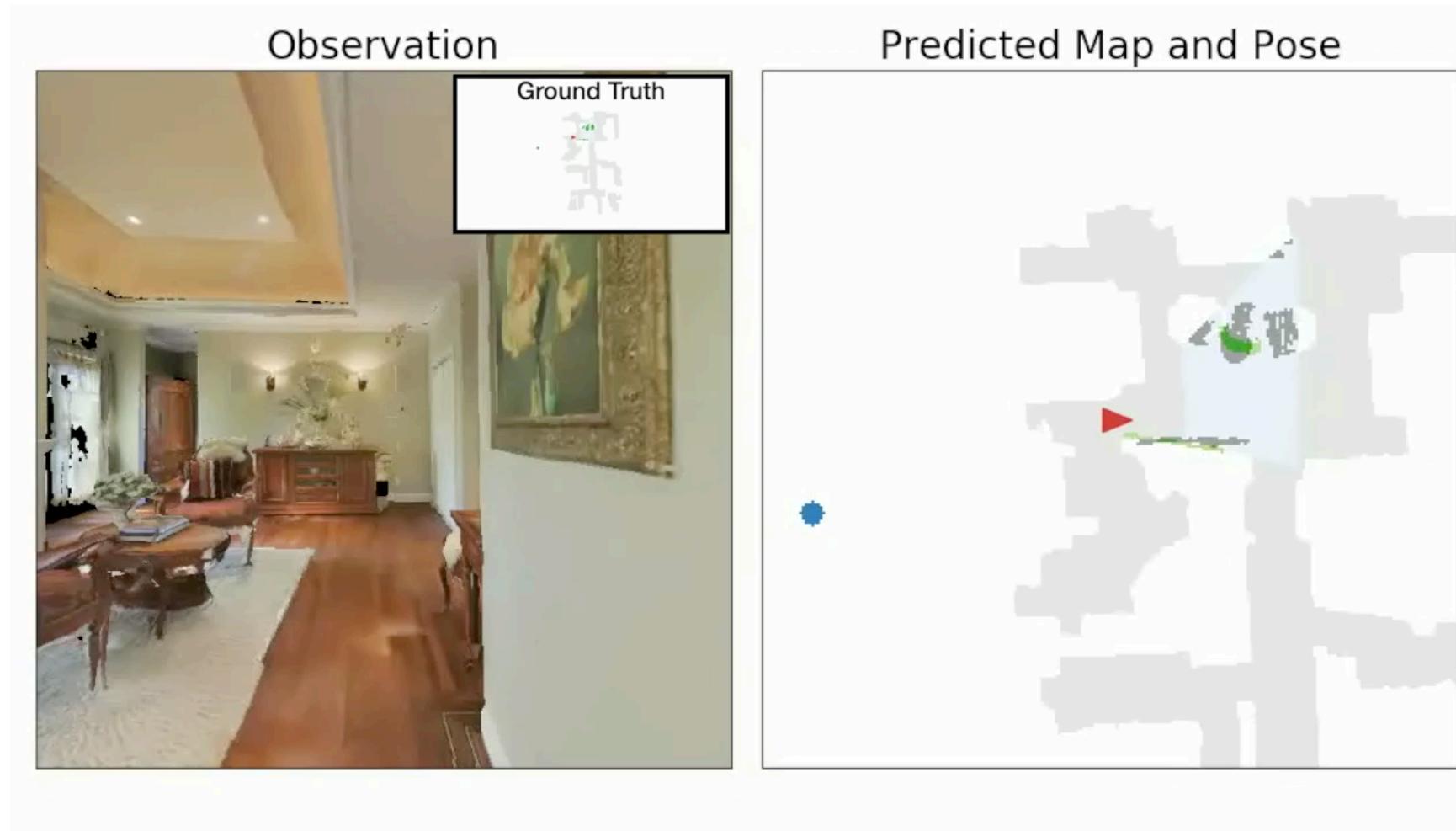
Active Neural SLAM: Overview



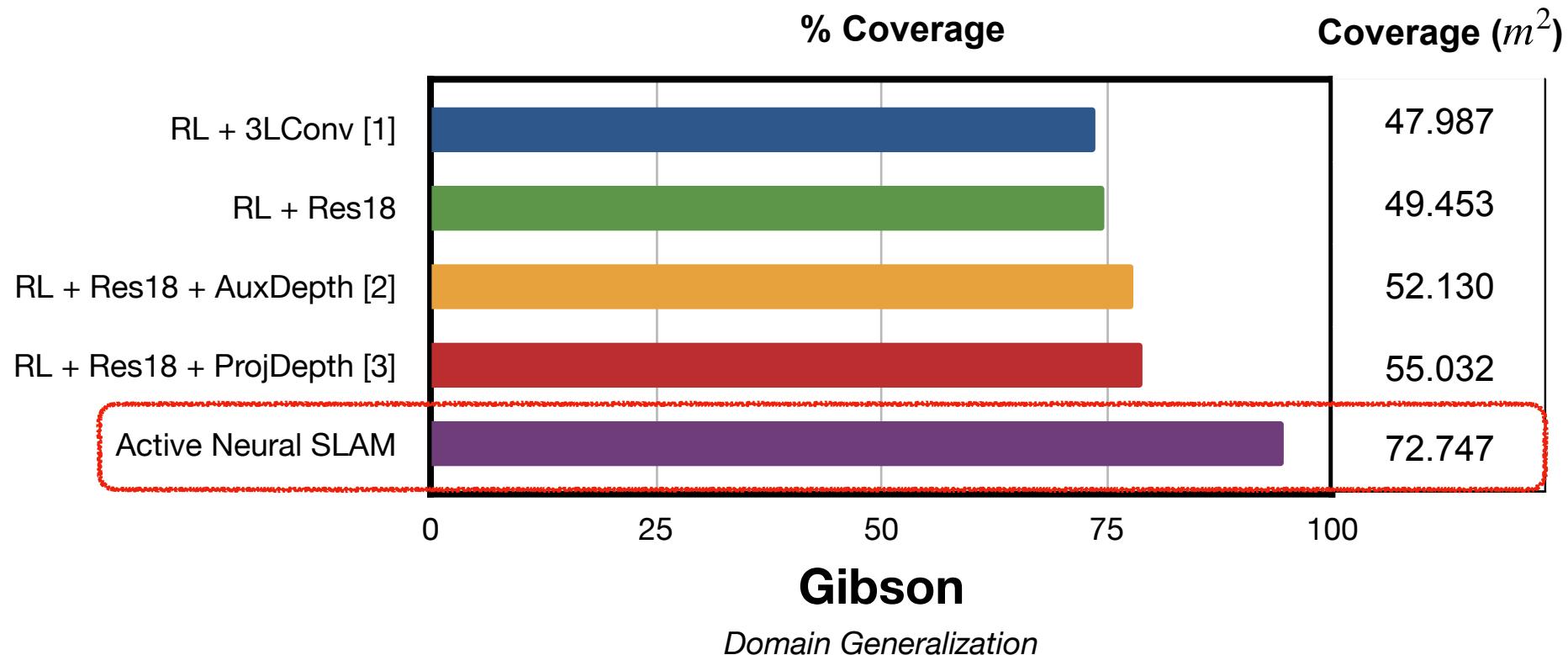
Neural SLAM Module



Domain Generalization: Matterport3D



Exploration Results



*Adapted from [1] Lample & Chaplot. AAAI-17, [2] Mirowski et al. ICLR-17, [3] Chen et al. ICLR-19

Goal-conditioned Navigation

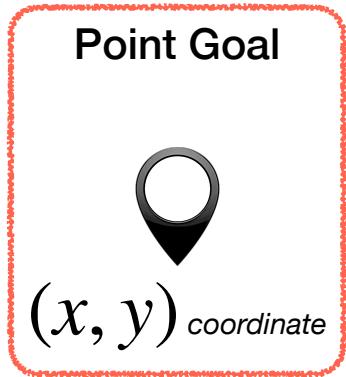


Image Goal



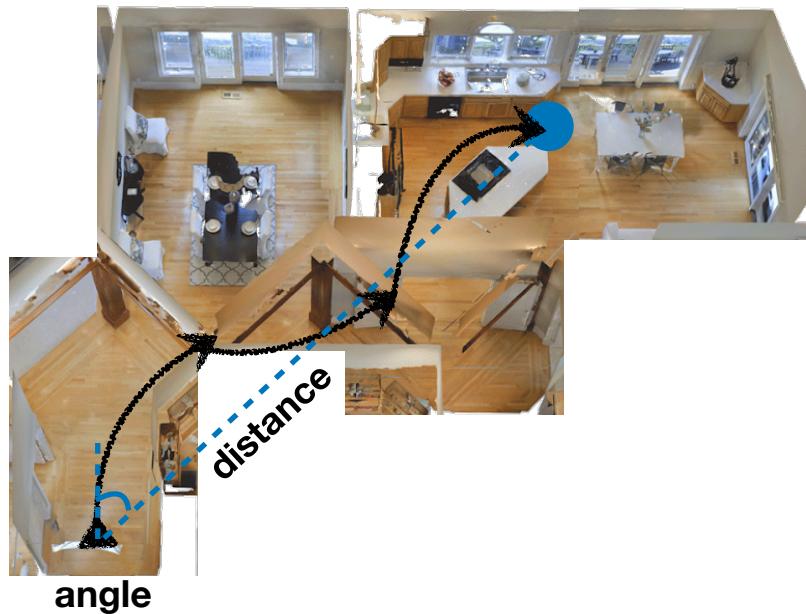
Object Goal

Chair
TV
Sofa

Language Goal

Blue Chair
Largest TV
White Sofa

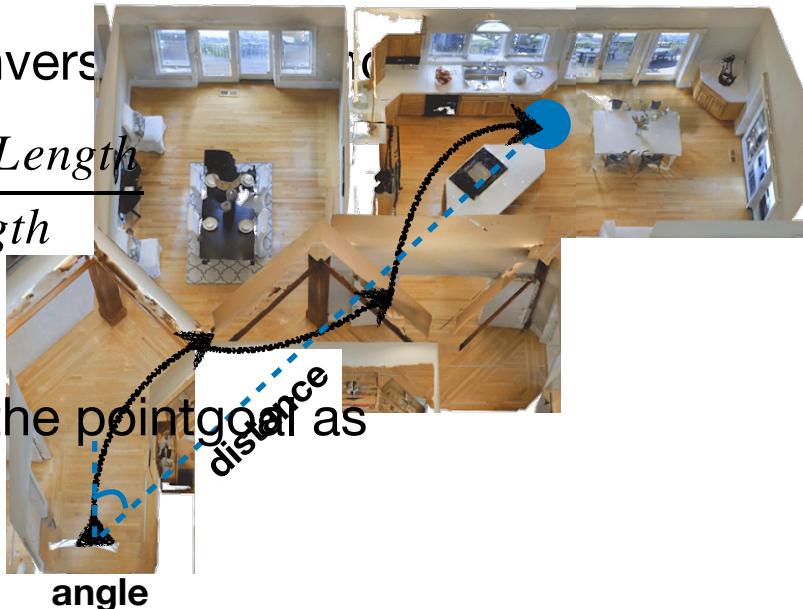
Point-Goal Navigation



Point-Goal Navigation

- Objective: Navigate to goal coordinates
- Metric: Success weighted by inverse path length

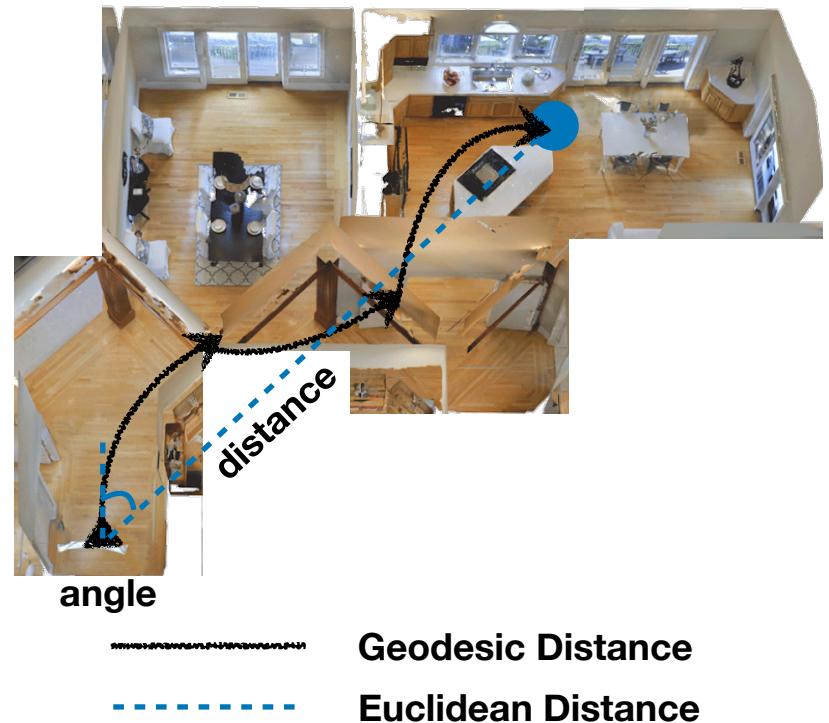
$$\frac{1}{N} \sum_{i=1}^N Success * \frac{\text{ShortestPathLength}}{\text{PathLength}}$$



- Global Policy -> always gives the pointgoal as the long-term goal

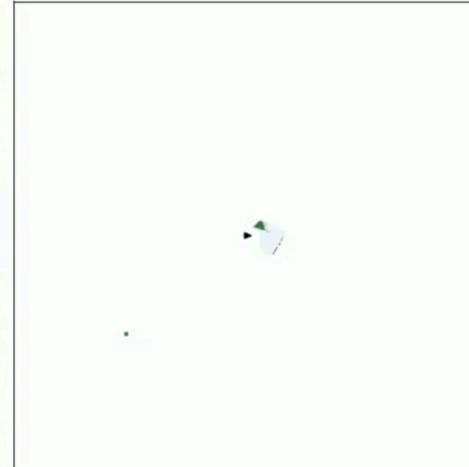
Harder Datasets

- **Hard-GEDR**
 - Higher Geodesic to Euclidean distance ratio (GEDR)
 - Avg GEDR 2.5 vs 1.37, minimum GEDR is 2
- **Hard-Dist**
 - Higher Geodesic distance
 - Avg Dist 13.5m vs 7.0m, minimum Dist is 10m

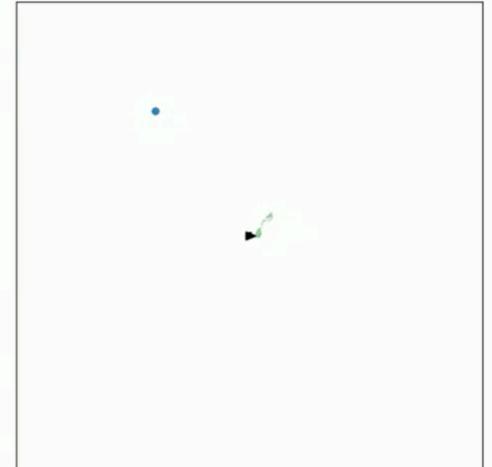


Point-Goal Navigation

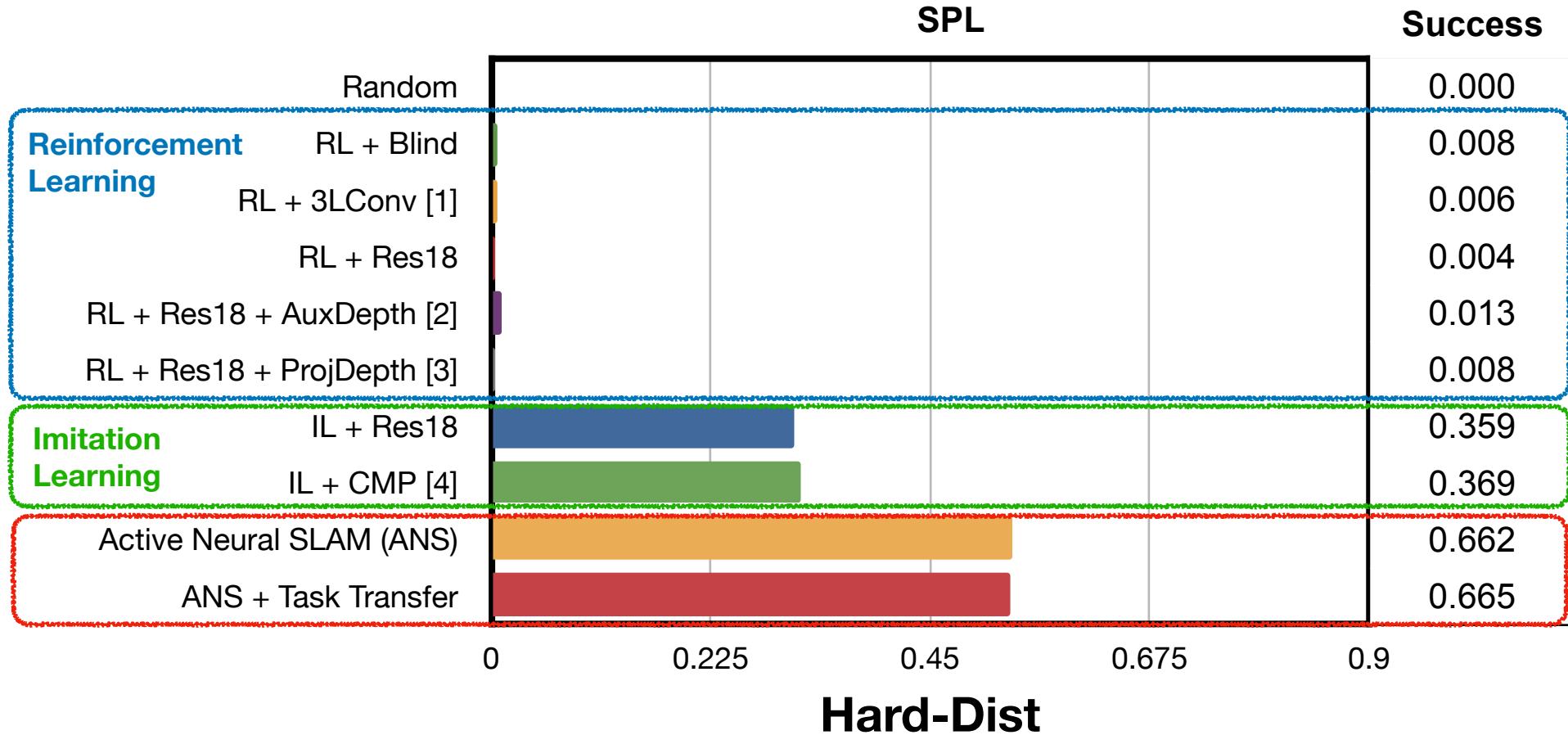
Gibson



MP3D

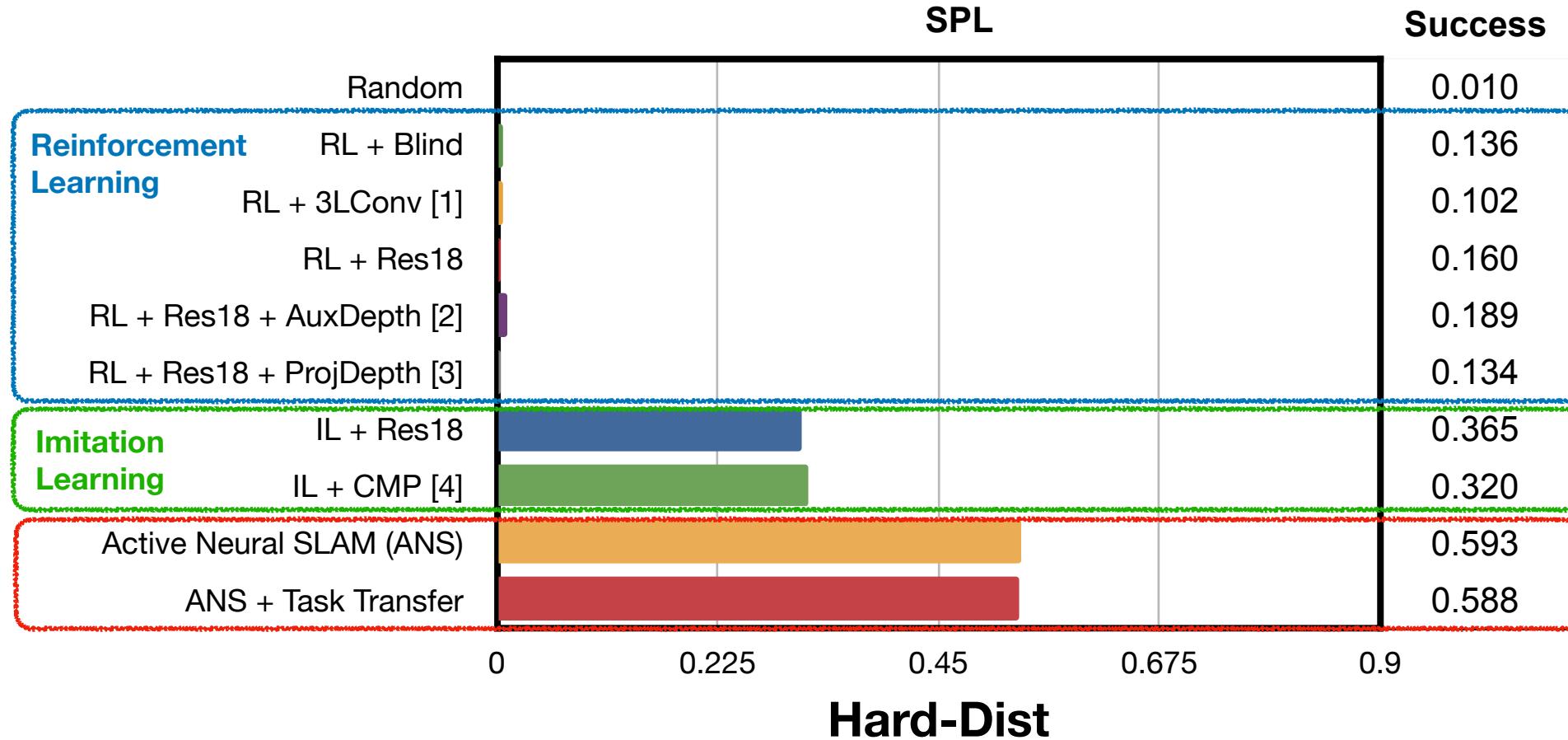


Results



*Adapted from [1] Lample & Chaplot. AAAI-17, [2] Mirowski et al. ICLR-17, [3] Chen et al. ICLR-19, [4] Gupta et al. CVPR-17

Results



*Adapted from [1] Lample & Chaplot. AAAI-17, [2] Mirowski et al. ICLR-17, [3] Chen et al. ICLR-19, [4] Gupta et al. CVPR-17

Navigation Tasks

Point Goal

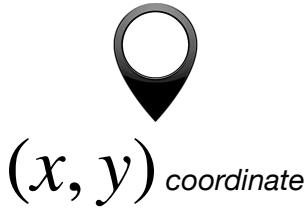


Image Goal



Object Goal

Chair
TV
Sofa

Language Goal

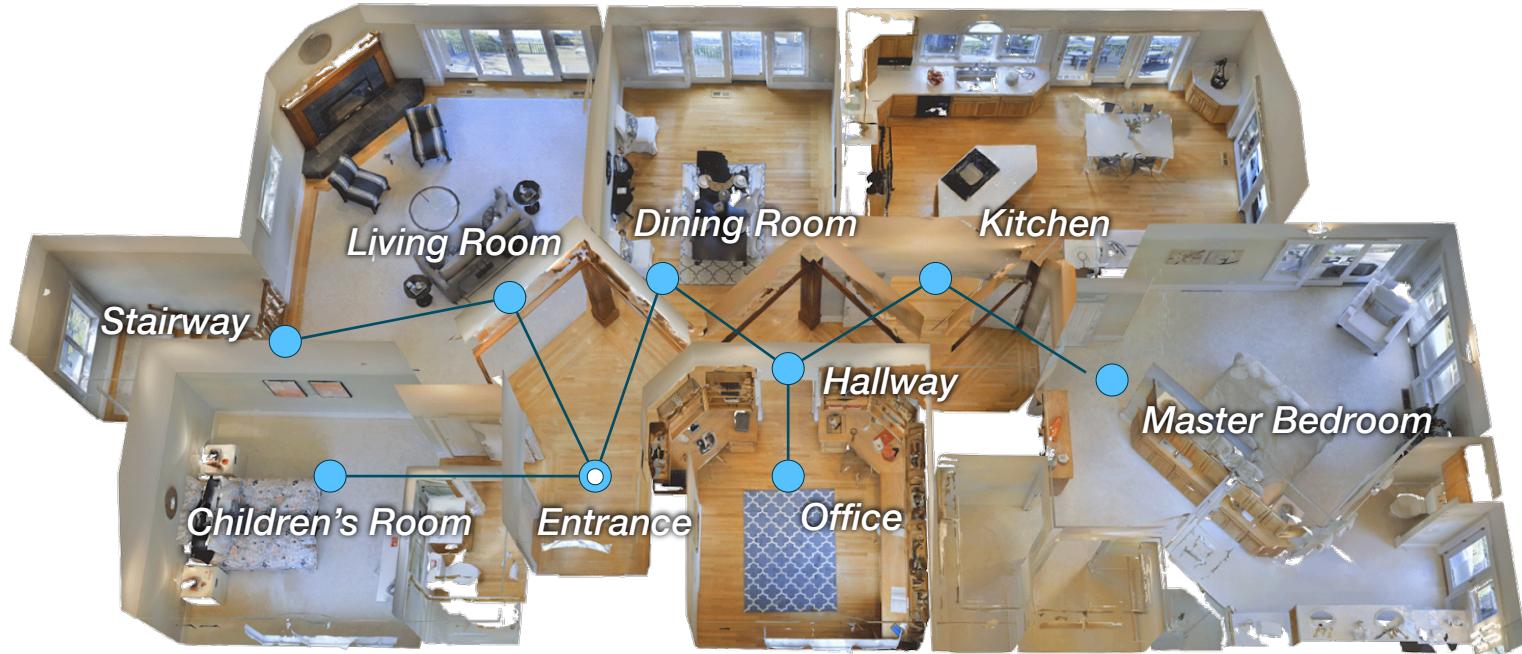
Blue Chair
Largest TV
White Sofa

Semantic Priors and Common-Sense



- Humans use semantic priors and common-sense to explore and navigate everyday
- Most navigation algorithms struggle to do so

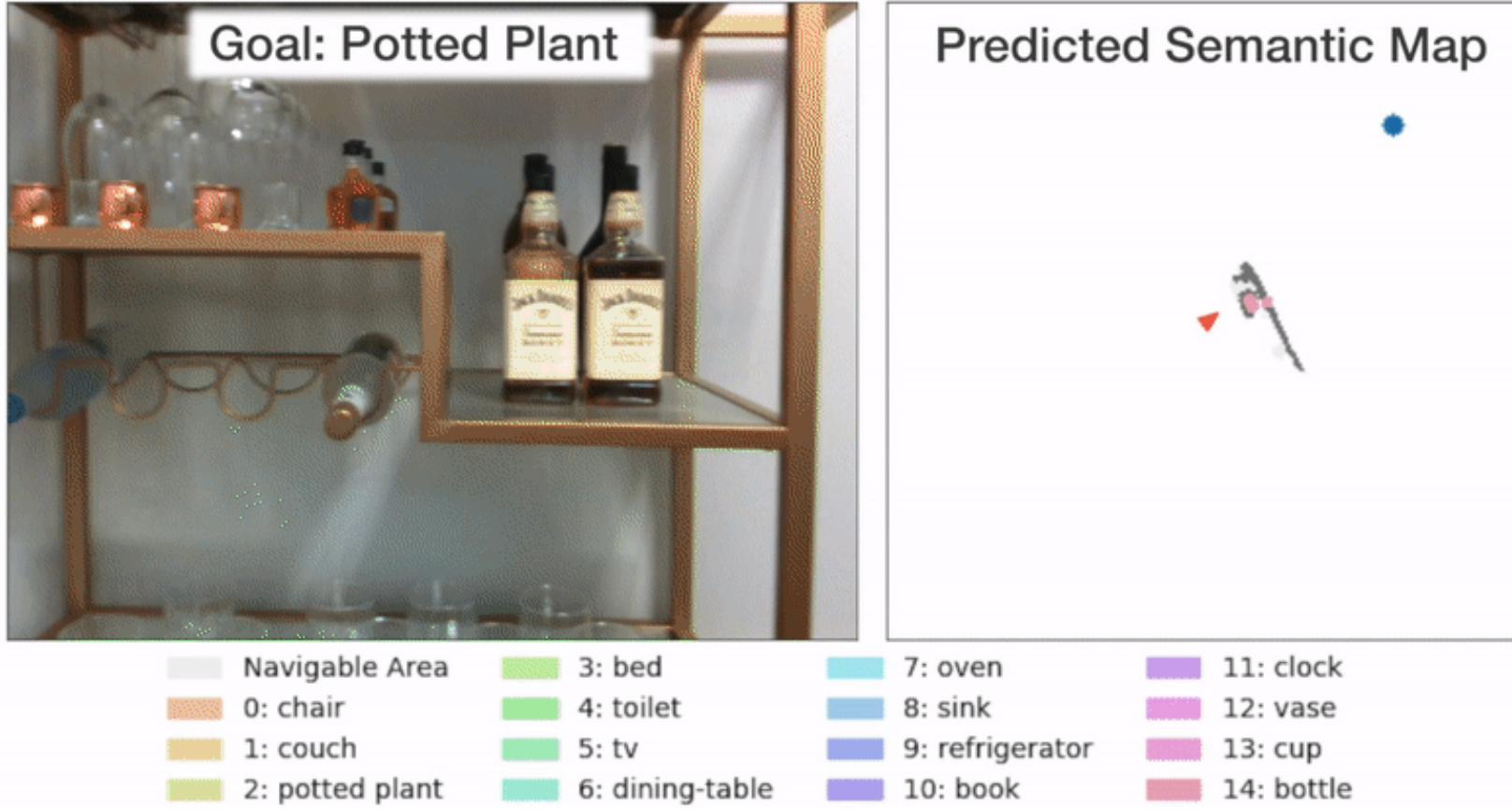
Topological Maps



Explicit Semantic Mapping

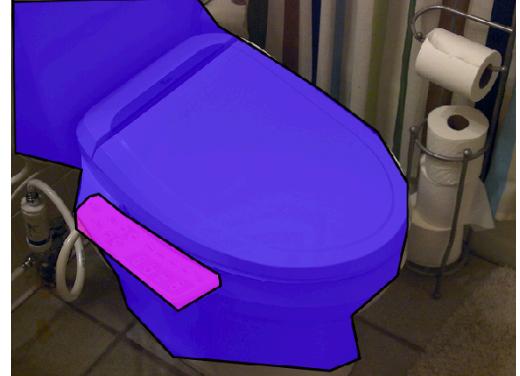


Explicit Semantic Mapping

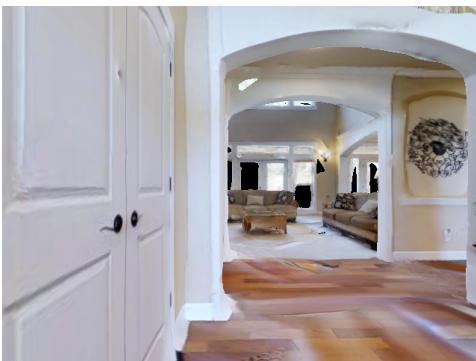


Internet vs Embodied Data

Static Internet Data



Active Embodied Data



Using Internet models for Embodied Agents



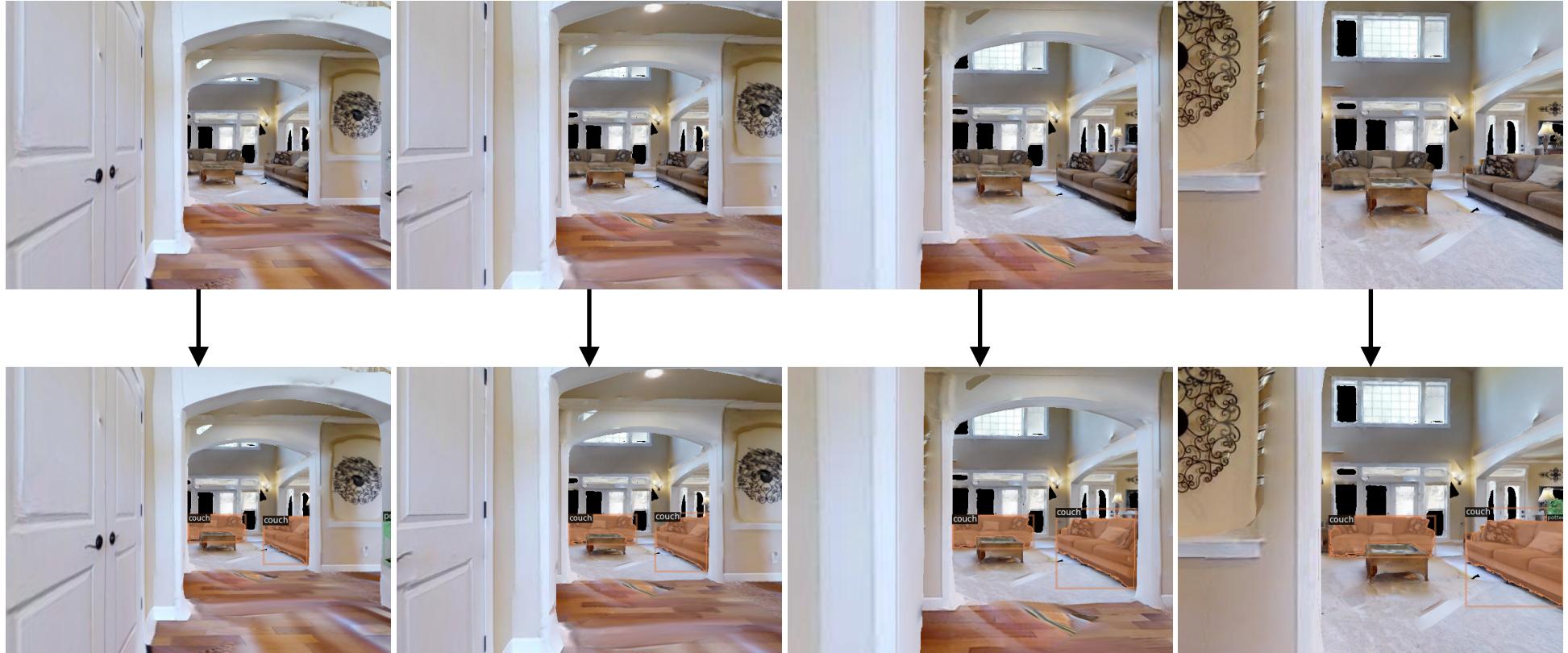
False positives



False negatives

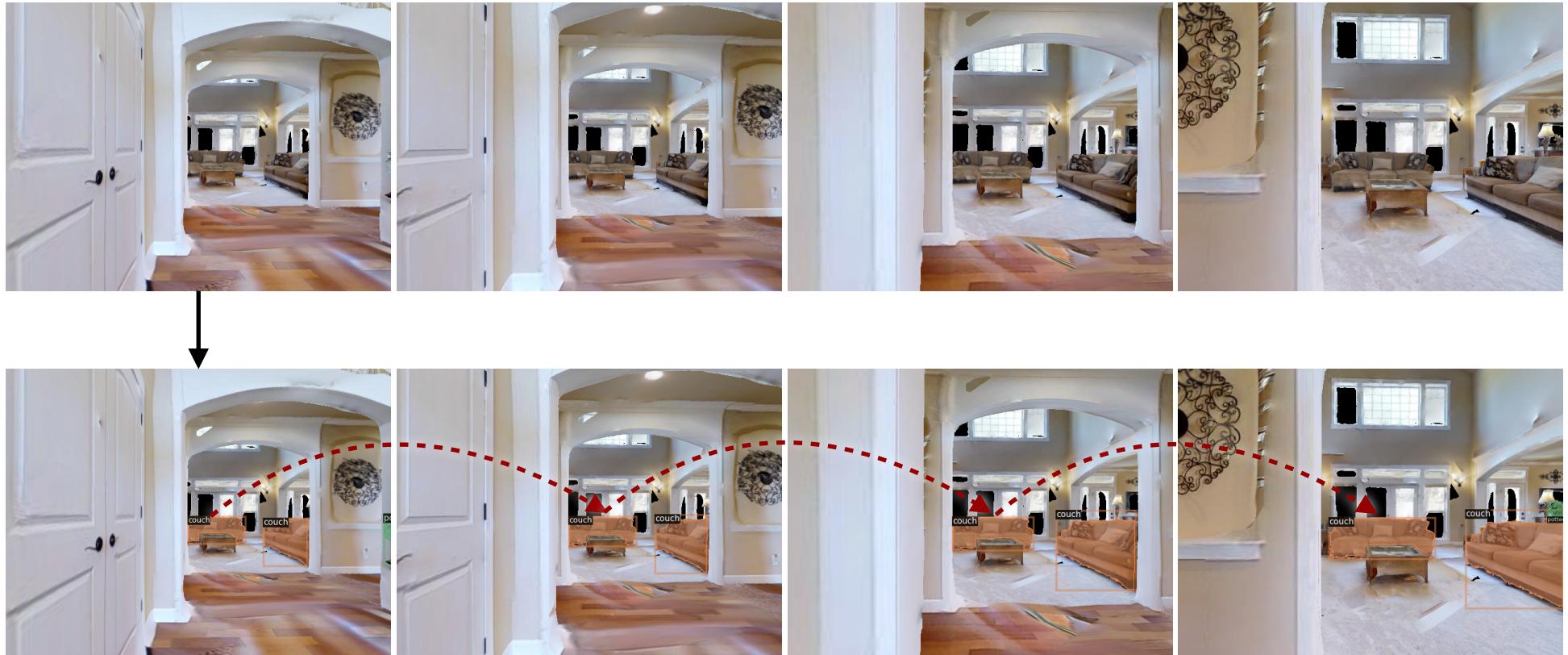
Embodied Perception

Active Embodied data

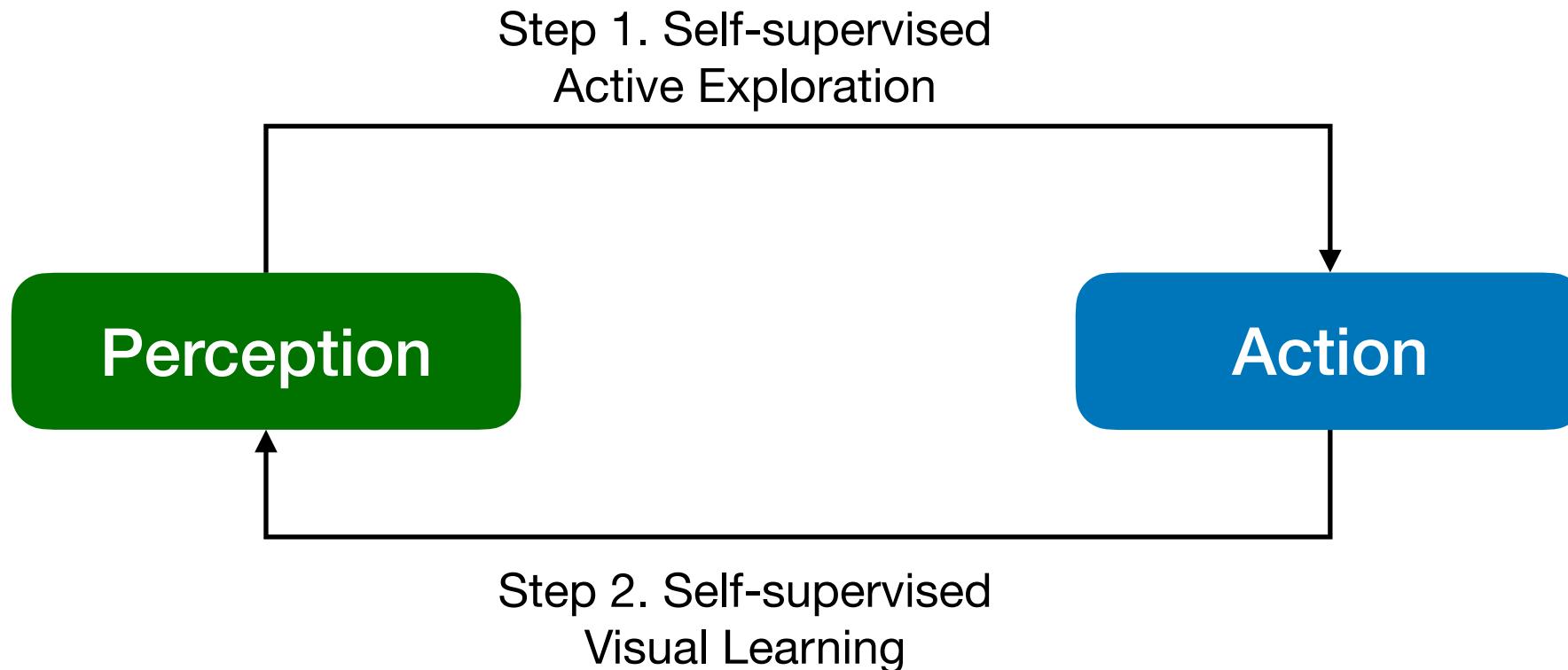


Embodied Perception

Active Embodied data



Perception-Action Loop



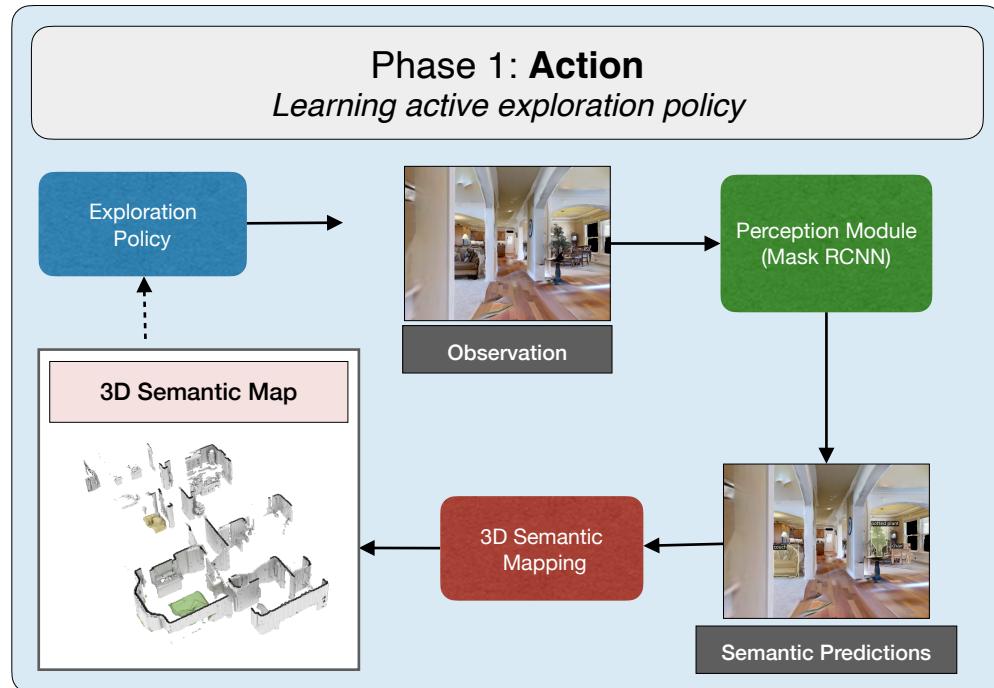
Pathak et al, Learning instance segmentation by interaction, 2018

Jang et al, Grasp2vec: Learning object representations from self-supervised grasping, 2018

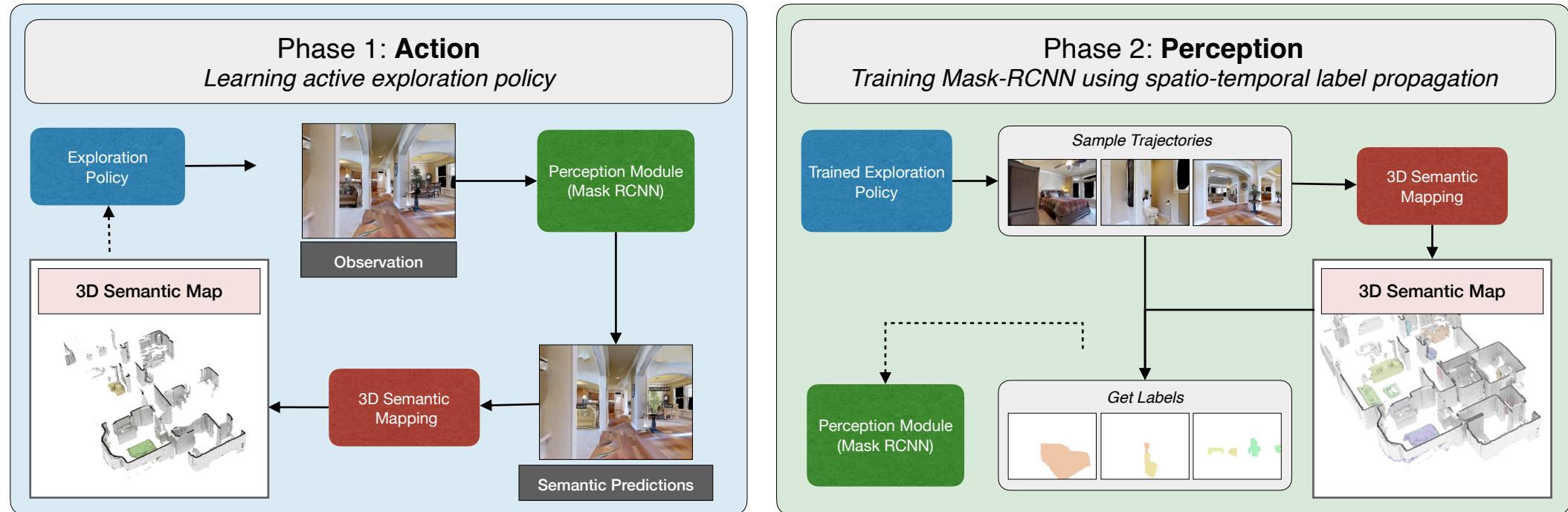
Eitel et al, Self-supervised transfer learning for instance segmentation through physical interaction, 2019

Fang et al., Move to See Better: Self-Improving Embodied Object Detection, 2021

SEAL: Self-supervised Embodied Active Learning

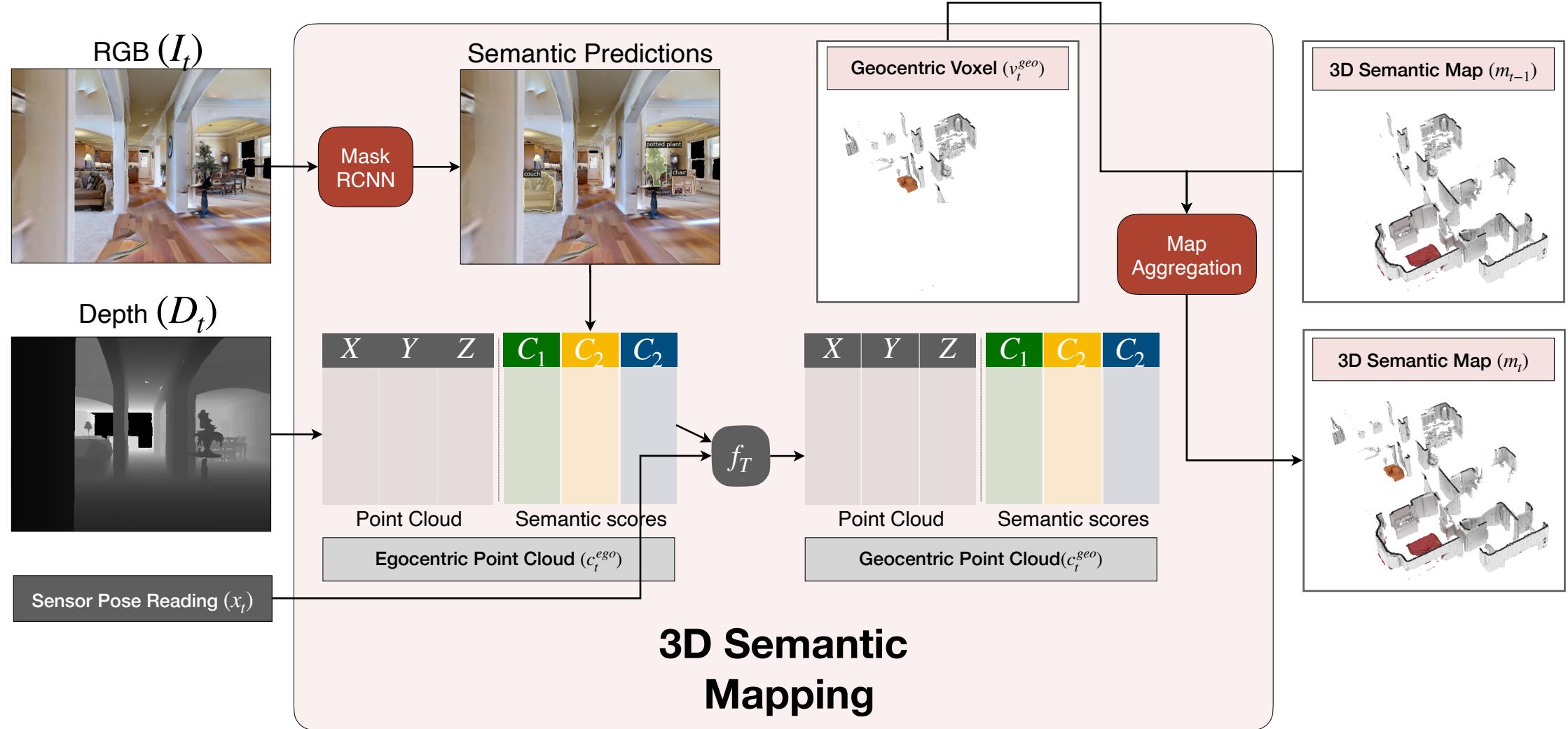


SEAL: Self-supervised Embodied Active Learning



Both phases do not require any additional labelled data

3D Semantic Mapping



3D Semantic Mapping



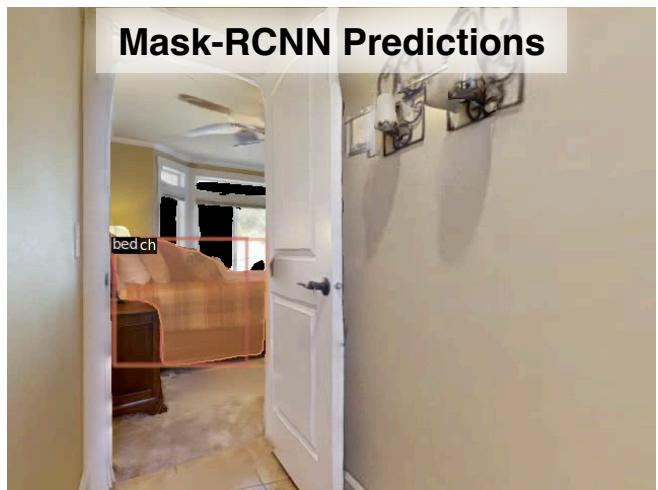
3D Semantic Map

$$M = K \times L \times W \times H$$

Chair
Couch
Potted Plant
Bed
Toilet
TV



3D Semantic Mapping

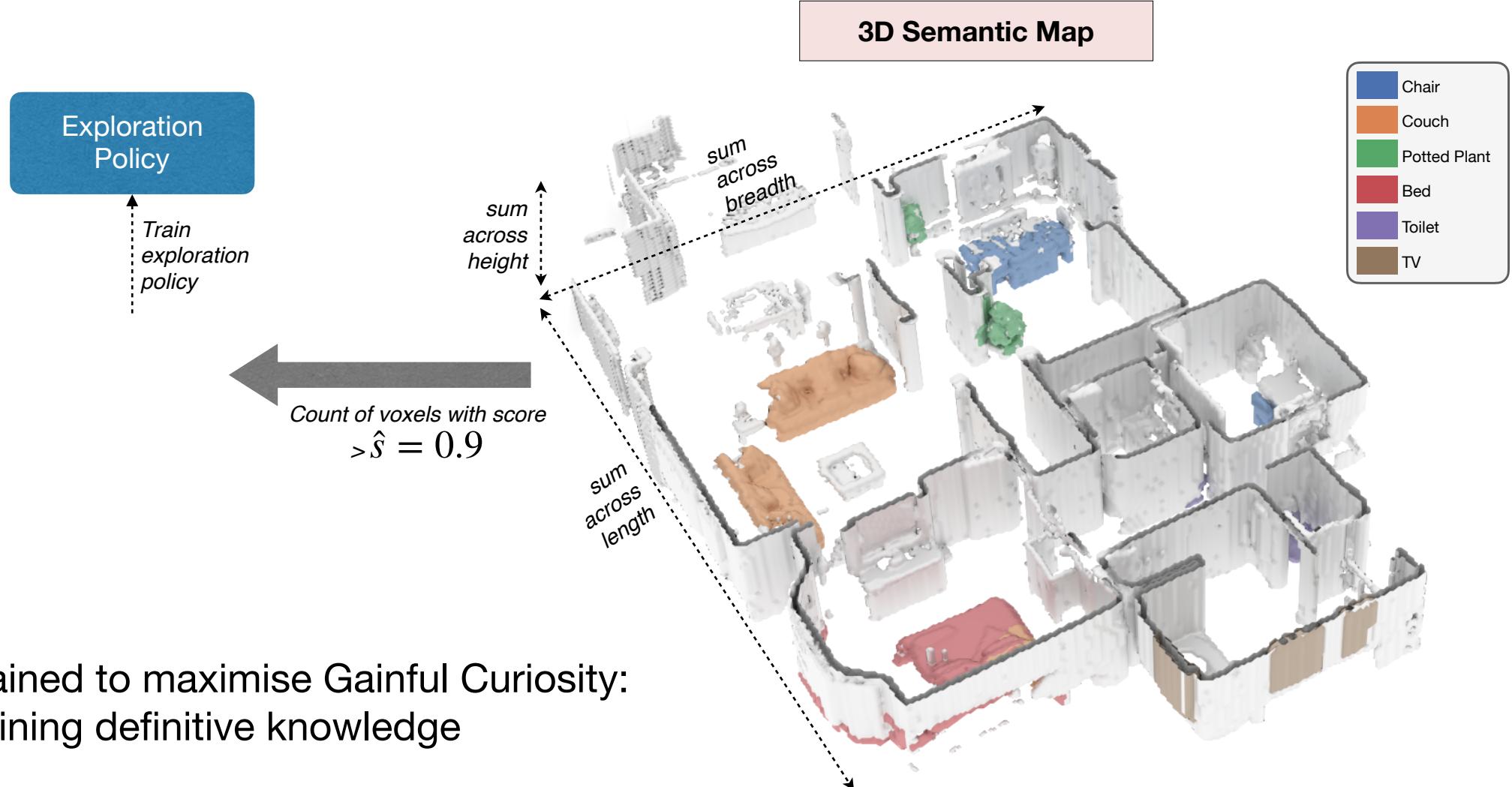


3D Semantic Map

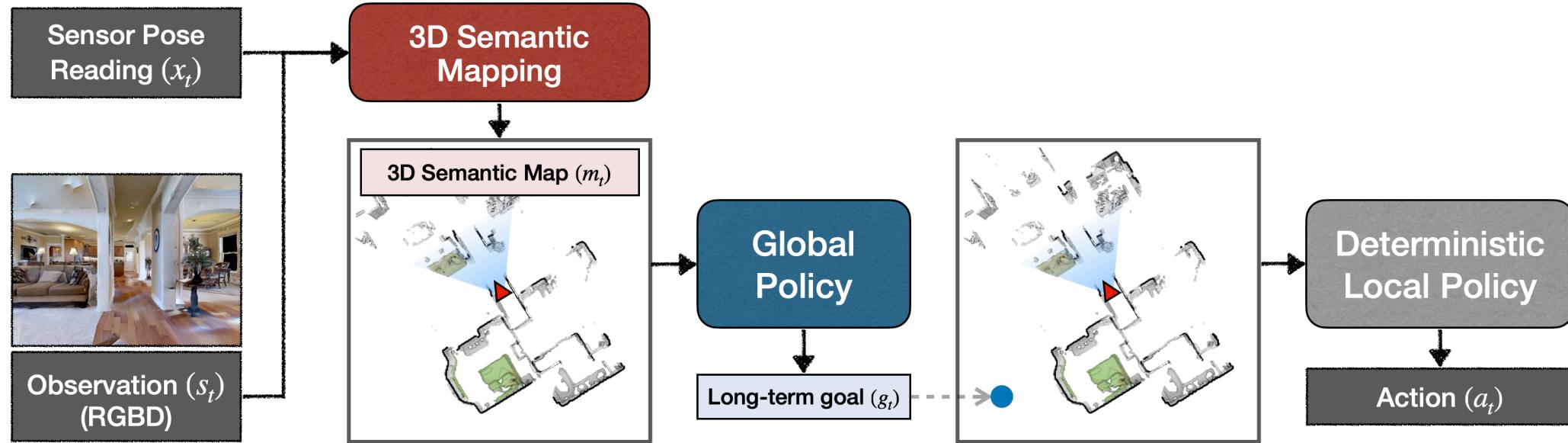
$$M = K \times L \times W \times H$$



Gainful Curiosity

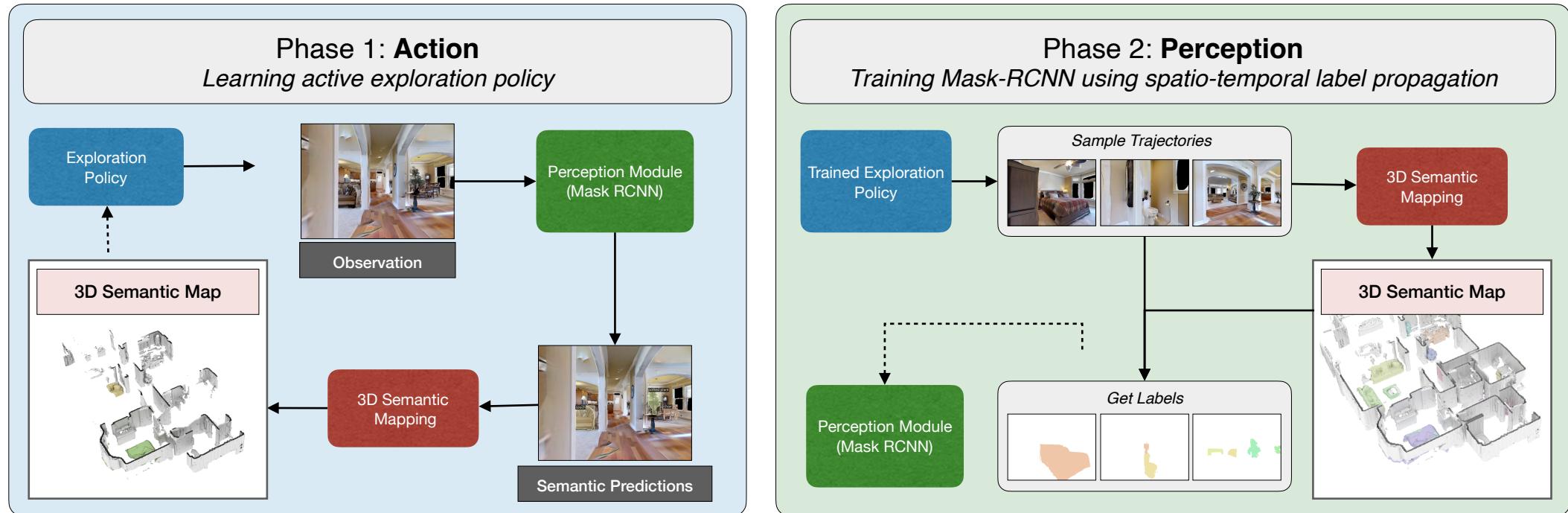


Policy Learning



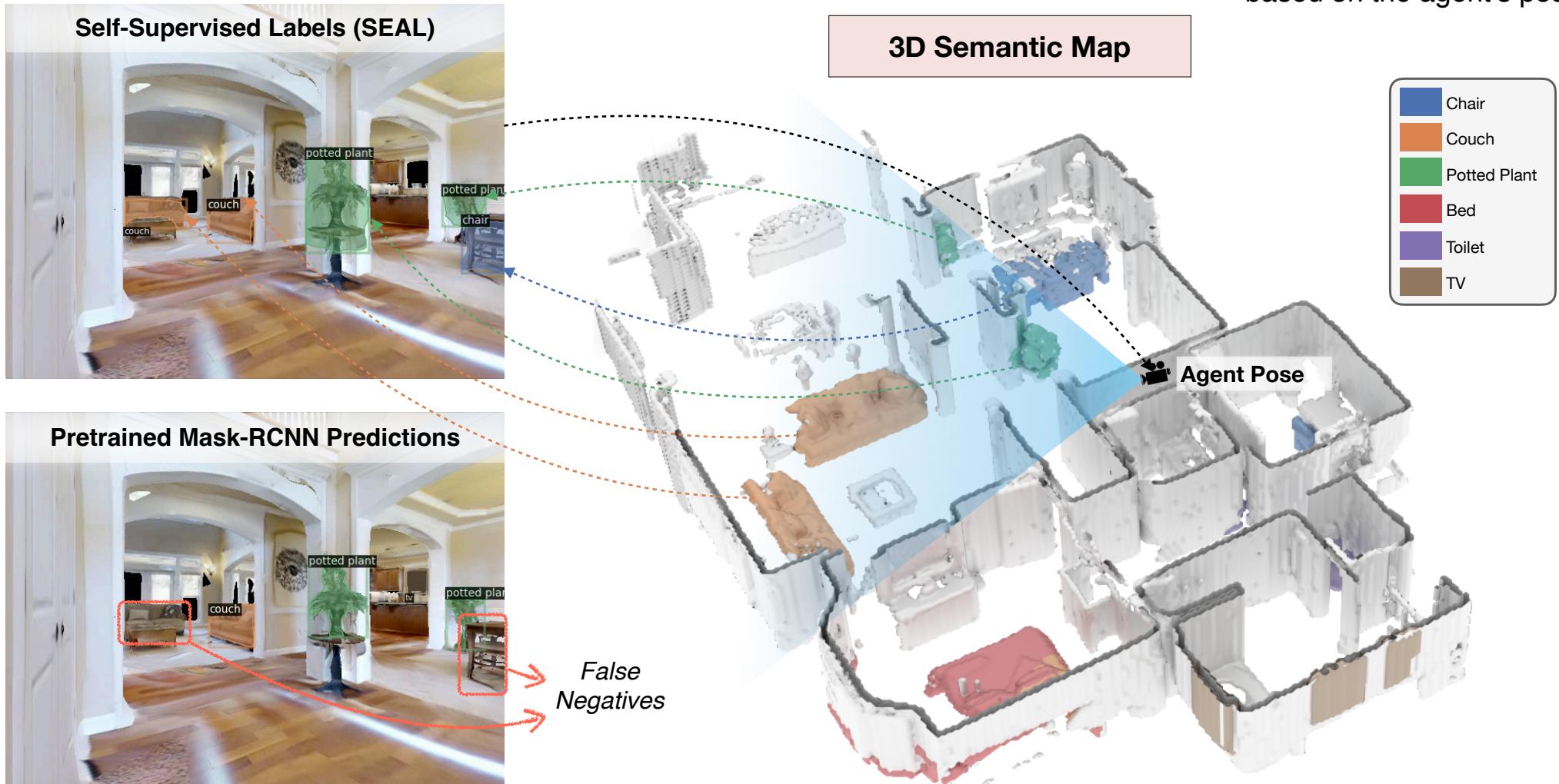
- Global Policy: samples a goal every 25 local steps
- Action Space: move forward (25cm), turn left or right (30 degrees)

SEAL: Self-supervised Embodied Active Learning

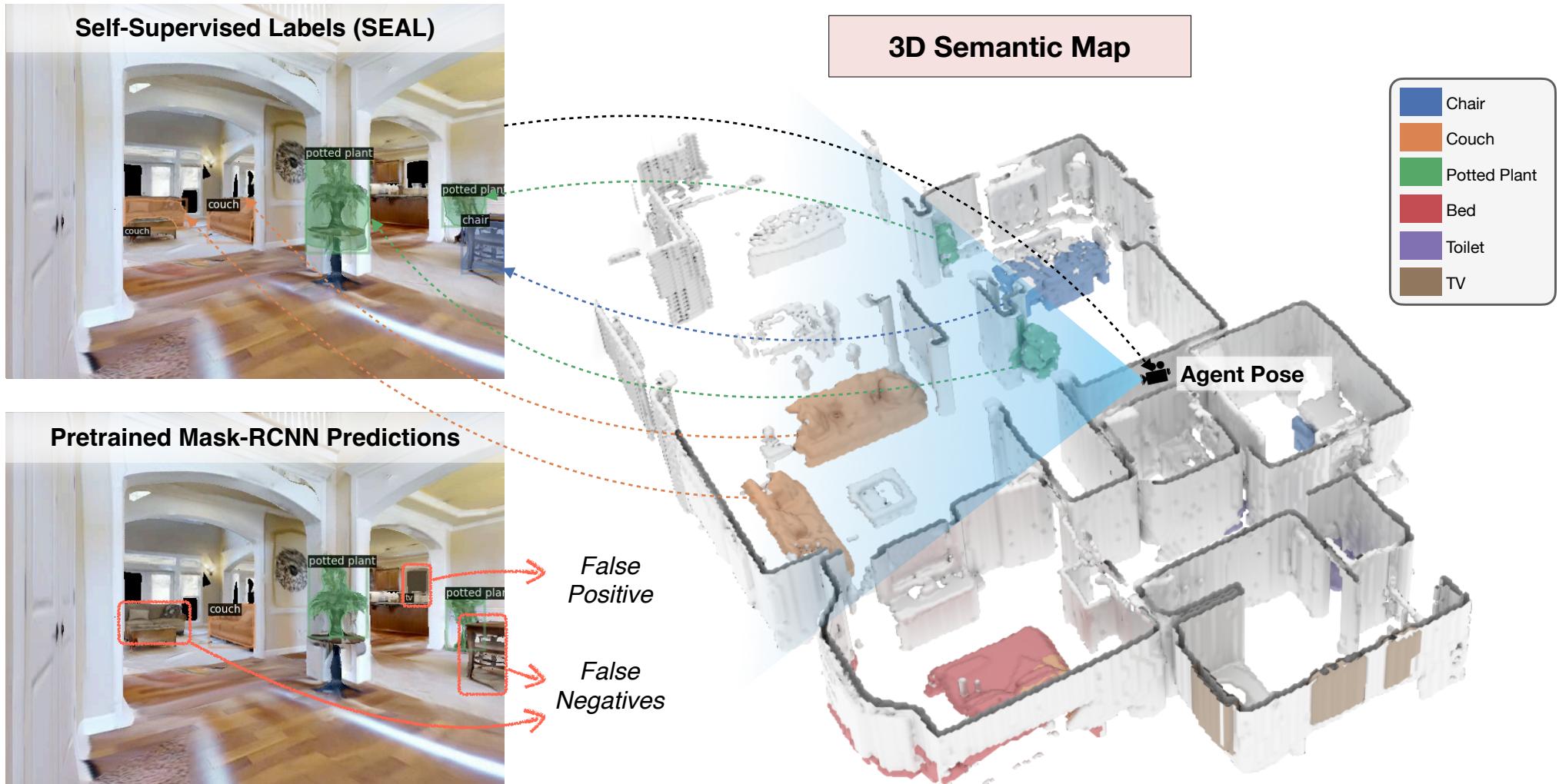


3D Label Propagation

Instance label for each pixel is obtained using ray tracing based on the agent's pose



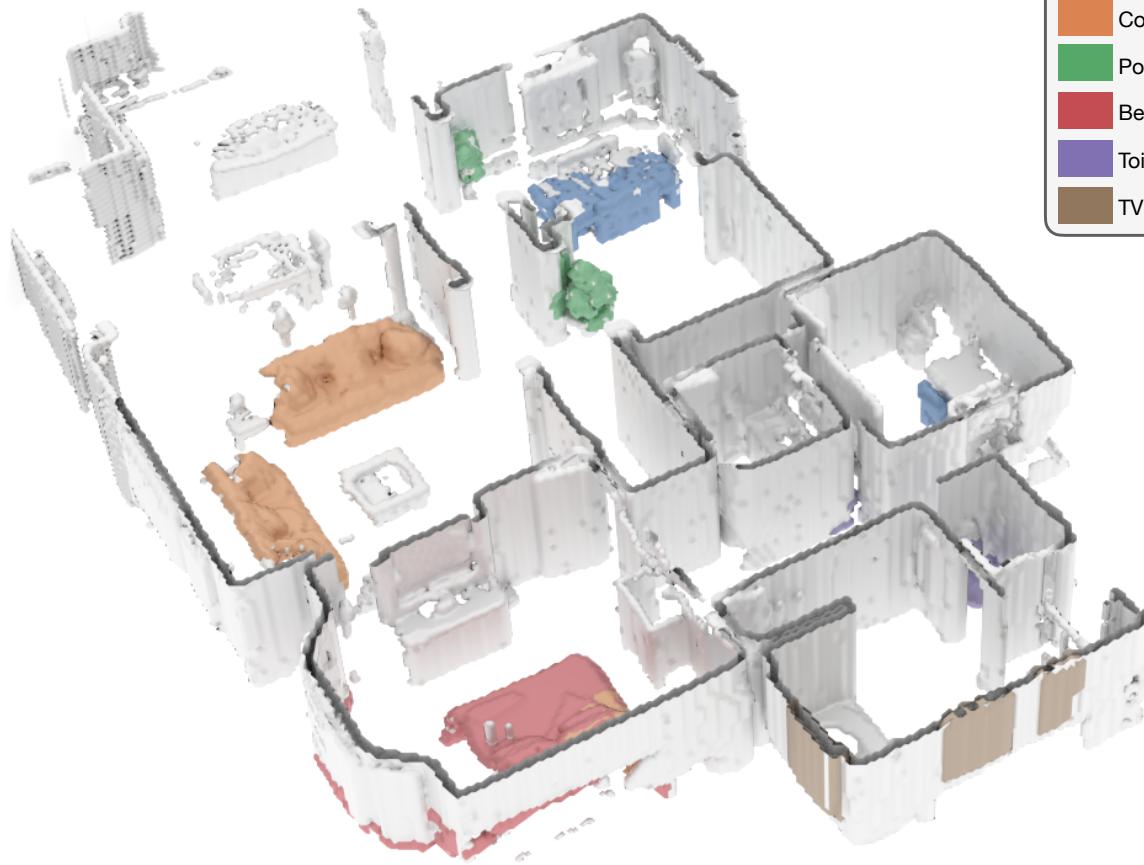
3D Label Propagation



3D Label Propagation



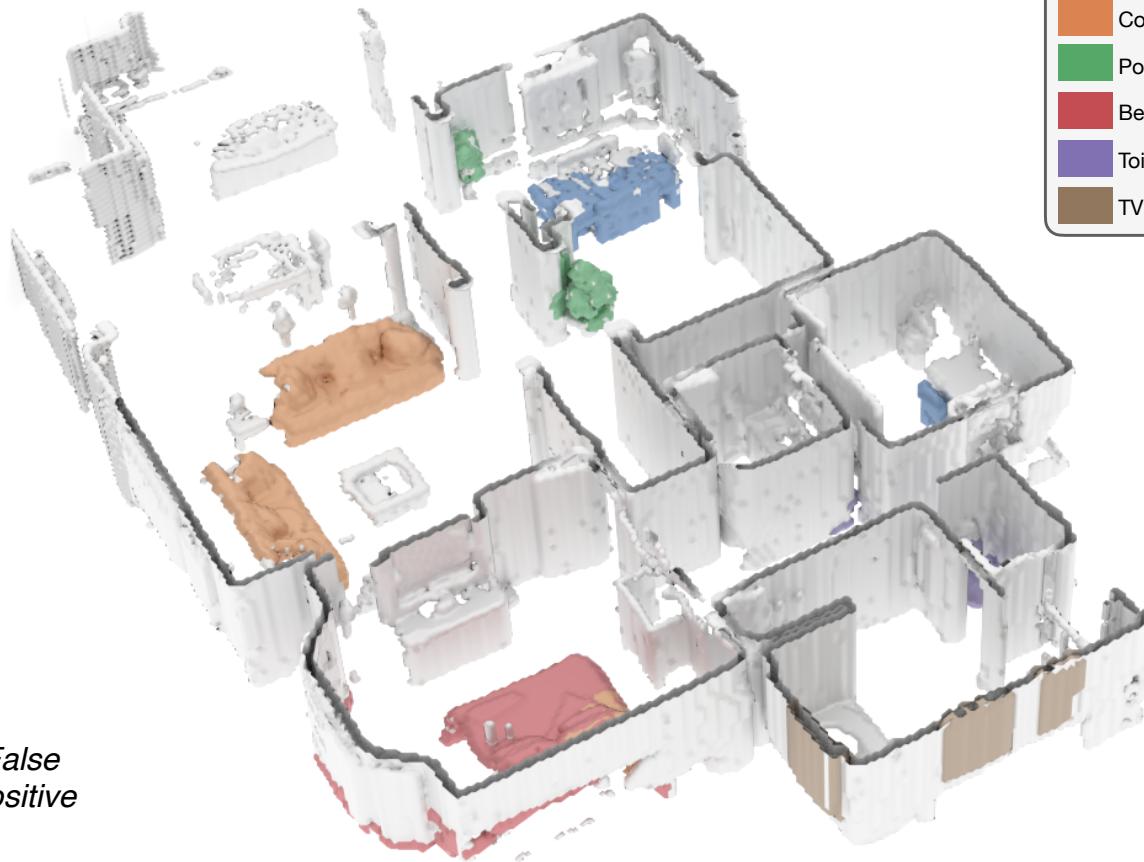
3D Semantic Map



3D Label Propagation



3D Semantic Map

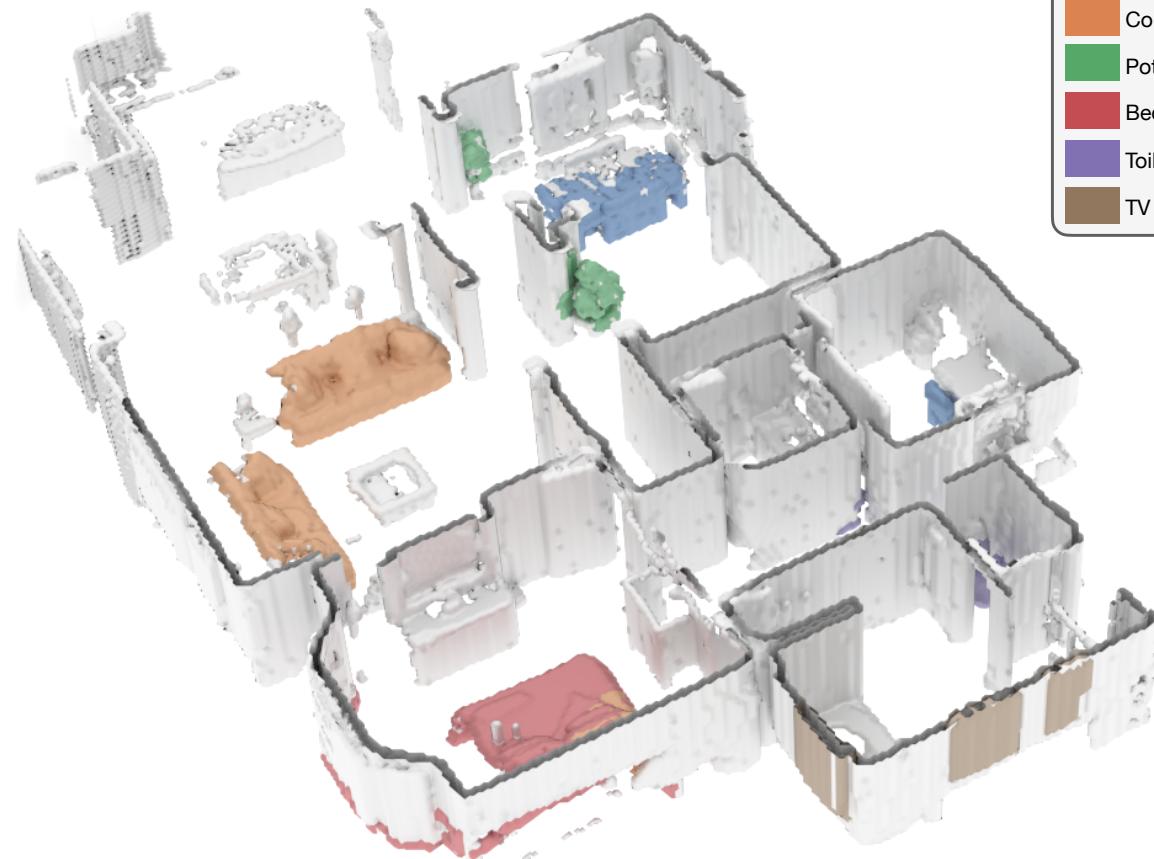


3D Label Propagation

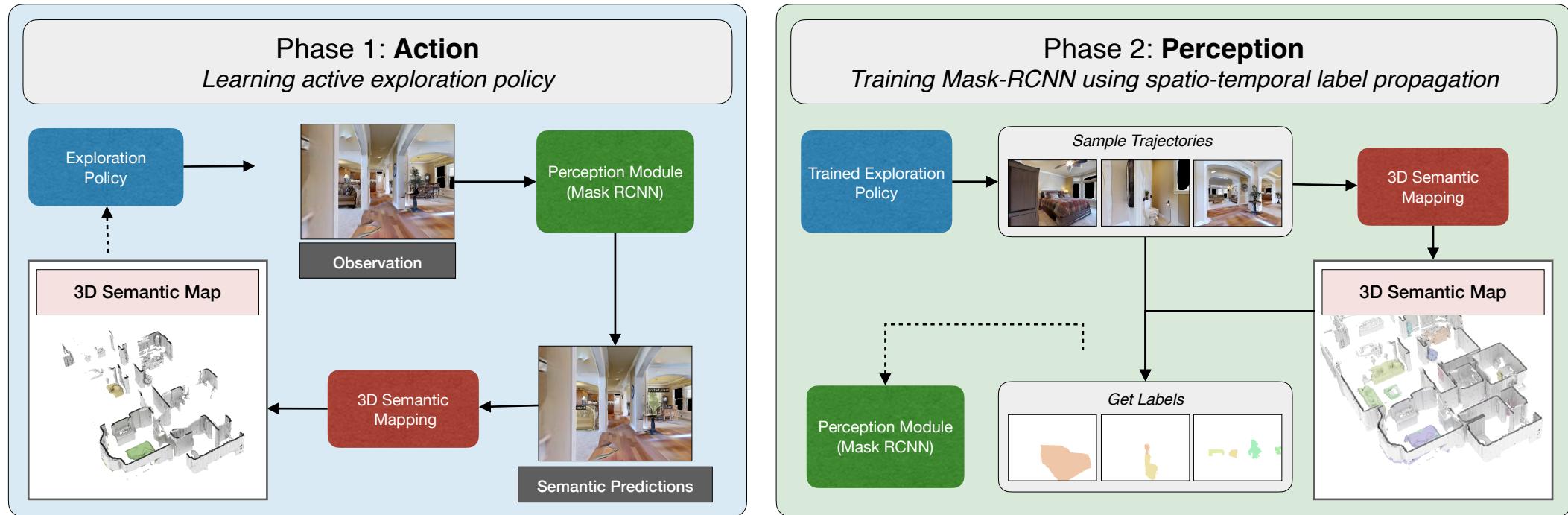


Perception Model
(Mask RCNN)

3D Semantic Map



SEAL: Self-supervised Embodied Active Learning



	Action	Perception
Generalization	Train	Train
Specialization	Train	Train + 1 episode test

Dataset

- Gibson dataset: 25 training and 5 test scenes
- 6 object categories: chair, couch, bed, toilet, TV, potted plant.
- Training Set: randomly sample 2500 images (500 per test scene)
- Evaluation Set: randomly sample 12,500 images (500 per training scene)
- Report bounding box and mask AP50 scores for detection and instance segmentation

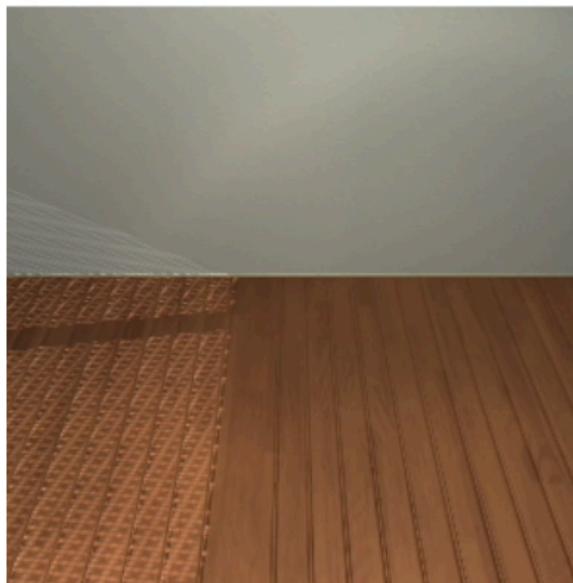
Results

Method	Generalization		Specialization	
	Object Detection	Instance Segmentation	Object Detection	Instance Segmentation
Pretrained Mask-RCNN	34.82	32.54	34.82	32.54
Random Policy + Self-training [51]	33.41	31.89	34.11	31.23
Random Policy + Optical Flow [22]	33.97	32.34	34.33	32.22
Frontier Exploration [52] + Self-training [51]	33.78	32.45	33.29	32.50
Frontier Exploration [52] + Optical Flow [22]	35.22	31.90	34.19	32.12
Active Neural SLAM [10] + Self-training [51]	34.35	31.20	34.84	32.44
Active Neural SLAM [10] + Optical Flow [22]	35.85	32.22	35.90	33.12
Semantic Curiosity [11] + Self-training [51]	35.04	32.19	35.23	32.88
Semantic Curiosity [11] + Optical Flow [22]	35.61	32.57	35.71	33.29
SEAL	40.02	36.23	41.23	37.28

EIF: Embodied Instruction Following: ALFRED

Instruction: place a cold lettuce slice in a waste basket.

RGB

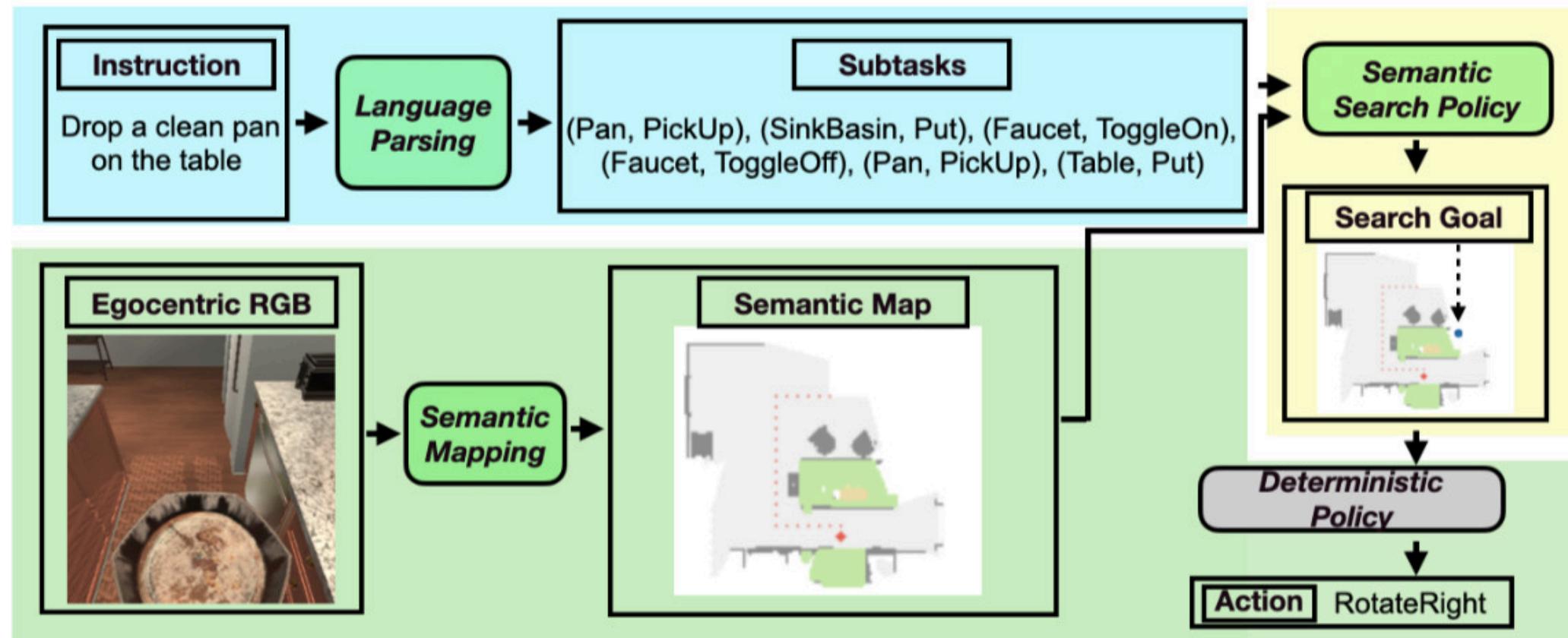


Completed Subgoals

- X PickUp, Knife
- X Slice, Lettuce
- X Put, Knife, Sink
- X PickUp SlicedLettuce
- X Open, Fridge
- X Put, SlicedLettuce, Fridge
- X Close, Fridge
- X Open, Fridge
- X PickUp, SlicedLettuce
- X Close, Fridge
- X Put, SlicedLettuce, GarbageCan

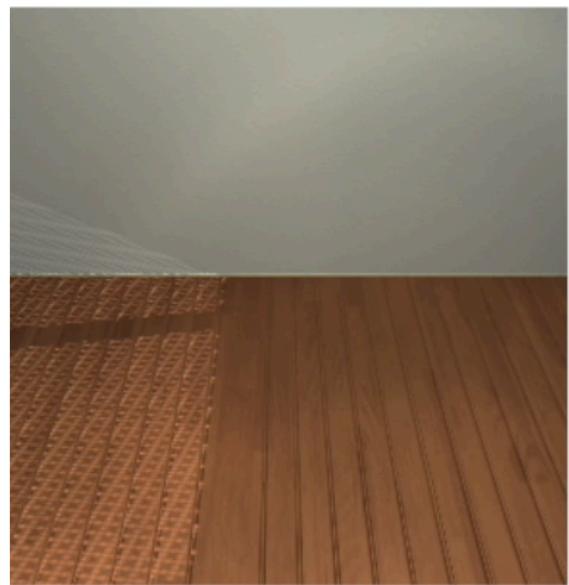
Predicted Action RotateLeft_90

FILM: Following Instructions in Language with Modular Methods



FILM: Following Instructions in Language with Modular Methods

Instruction: place a cold lettuce slice in a waste basket.



Predicted Action

Semantic Map



RotateLeft_90

Completed Subgoals

- X PickUp, Knife
- X Slice, Lettuce
- X Put, Knife, Sink
- X PickUp SlicedLettuce
- X Open, Fridge
- X Put, SlicedLettuce, Fridge
- X Close, Fridge
- X Open, Fridge
- X PickUp, SlicedLettuce
- X Close, Fridge
- X Put, SlicedLettuce, GarbageCan

Results

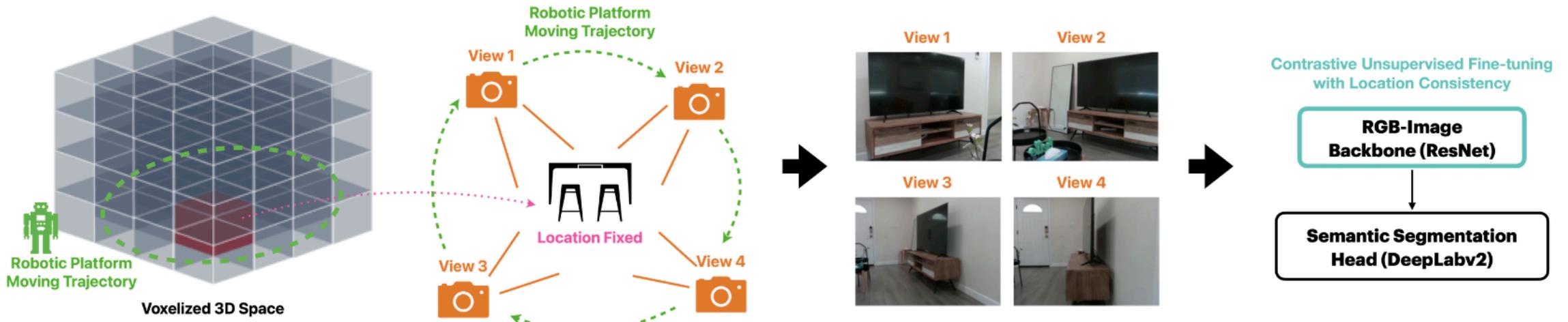
Table 1: Test results. Top section uses step-by-step instructions; the bottom section does not.

Method		Tests Seen				Tests Unseen			
		PLWGC	GC	PLWSR	SR	PLWGC	GC	PLWSR	SR
Low-level Sequential Instructions + High-level Goal Instruction									
SEQ2SEQ	(Shridhar et al., 2020)	6.27	9.42	2.02	3.98	4.26	7.03	0.08	3.9
MOCA	(Singh et al., 2020)	22.05	28.29	15.10	22.05	9.99	14.28	2.72	5.30
E.T.	(Pashevich et al., 2021)	-	36.47	-	28.77	-	15.01	-	5.04
E.T. + synth. data	(Pashevich et al., 2021)	34.93	45.44	27.78	38.42	11.46	18.56	4.10	8.57
LWIT	(Nguyen et al., 2021)	23.10	40.53	43.10	30.92	16.34	20.91	5.60	9.42
HiTUT	(Zhang & Chai, 2021)	17.41	29.97	11.10	21.27	11.51	20.31	5.86	13.87
ABP	(Kim et al., 2021)	4.92	51.13	3.88	44.55	2.22	24.76	1.08	15.43
FILM w.o. SEMANTIC SEARCH		<u>13.10</u>	<u>35.59</u>	<u>9.43</u>	<u>25.90</u>	<u>13.37</u>	<u>35.51</u>	<u>10.17</u>	<u>23.94</u>
FILM 		<u>15.06</u>	<u>38.51</u>	<u>11.23</u>	<u>27.67</u>	<u>14.30</u>	<u>36.37</u>	<u>10.55</u>	<u>26.49</u>
High-level Goal Instruction Only									
LAV	(Nottingham et al., 2021)	13.18	23.21	6.31	13.35	10.47	17.27	3.12	6.38
HiTUT G-only	(Zhang & Chai, 2021)	-	21.11	-	13.63	-	17.89	-	11.12
HLSM	(Blukis et al., 2021)	11.53	35.79	6.69	25.11	8.45	27.24	4.34	16.29
FILM w.o. SEMANTIC SEARCH		<u>12.22</u>	<u>34.41</u>	<u>8.65</u>	<u>24.72</u>	<u>12.69</u>	<u>34.00</u>	<u>9.44</u>	<u>22.56</u>
FILM 		<u>14.17</u>	<u>36.15</u>	<u>10.39</u>	<u>25.77</u>	<u>13.13</u>	<u>34.75</u>	<u>9.67</u>	<u>24.46</u>

FILM: Following Instructions in Language with Modular Methods

So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, Ruslan Salakhutdinov, ICLR 2022

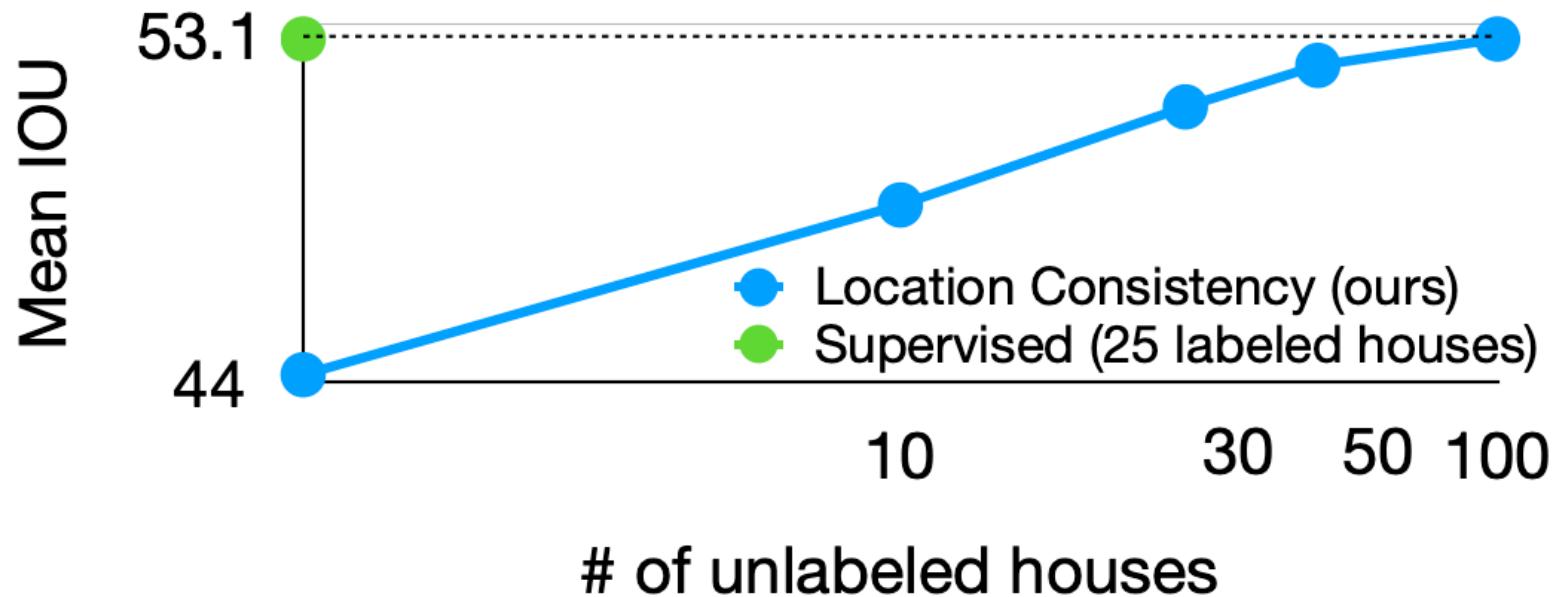
Self-supervision with Location Consistency



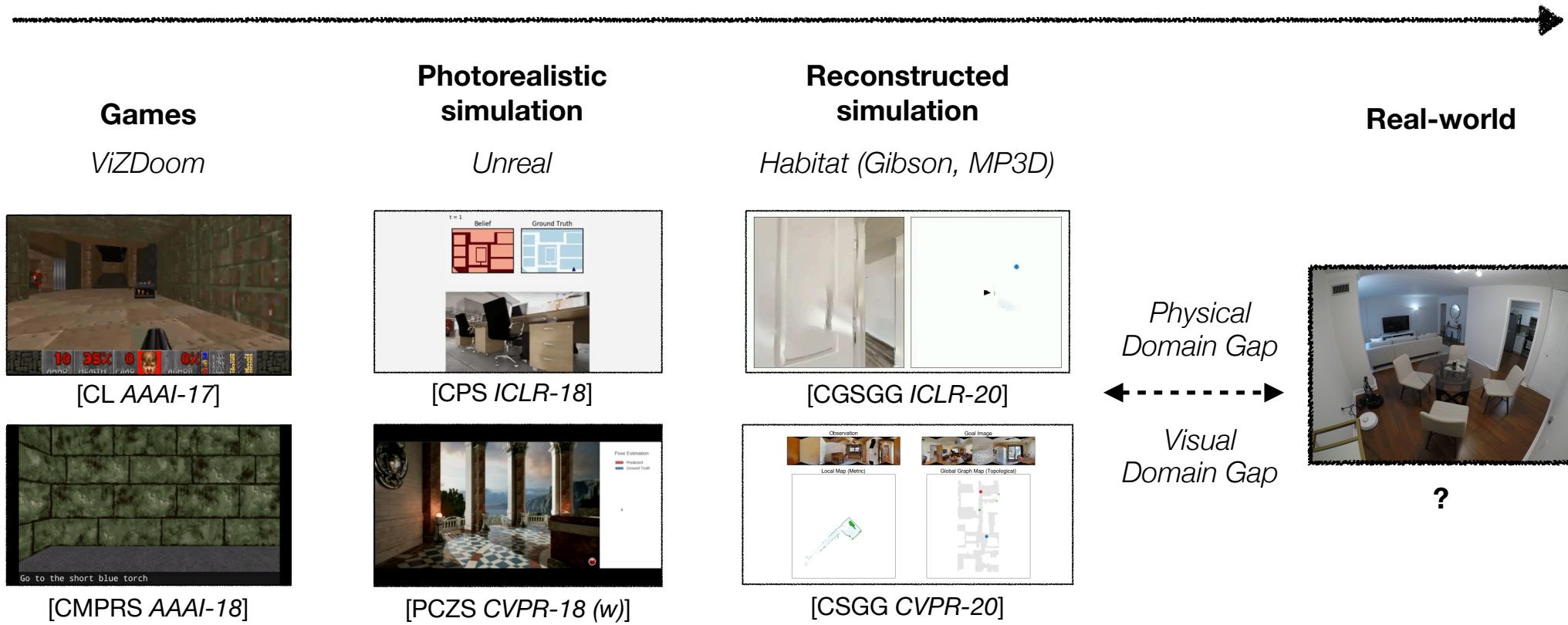
Finding Bed



Self-Supervision: Semantic Segmentation



Simulation to Real



Simulation to Real

- Physical Domain Gap
 - Actuation noise models
 - Sensor noise models
- Visual Domain Gap
 - Image Translation
 - Policy-based



PyRobot is a light weight, high-level interface which provides hardware independent APIs for robotic manipulation and navigation.
This repository also contains the low-level stack for *LoCoBot*, a low cost mobile manipulator hardware platform.

- [What can you do with PyRobot?](#)
- [Installation](#)
- [Getting Started](#)
- [The Team](#)
- [Citation](#)
- [License](#)
- [Future features](#)

What can you do with PyRobot?



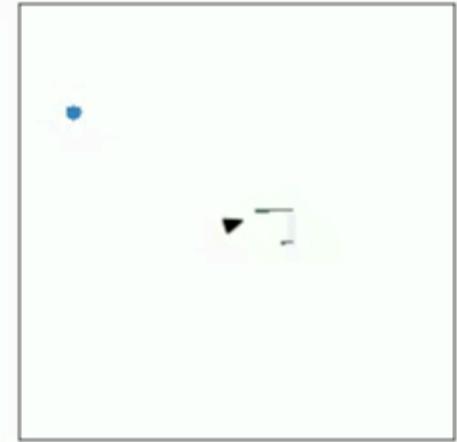
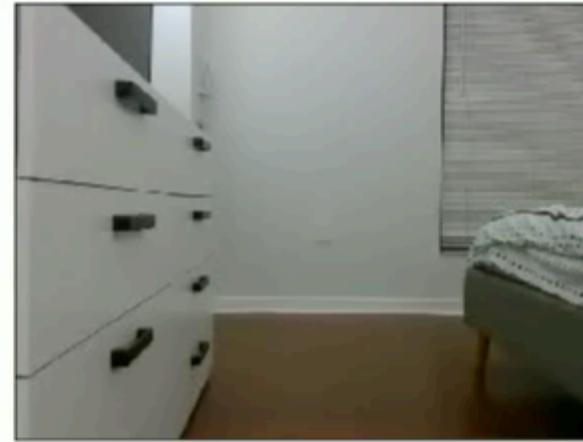
pyrobot.org

LoCoBot



locobot.org

Simulation to Real



Building Intelligent Agents

Navigate Autonomously
Localize and Plan
Multi-modal Input
Perceptive Human Speech
Reason & Understand Language
Recognize objects

