Embodied AI: Language and Perception

Russ Salakhutdinov

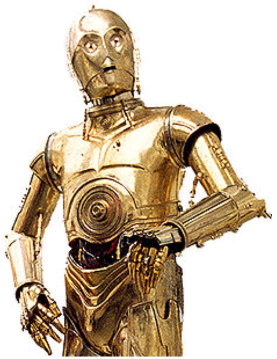
Machine Learning Department
Carnegie Mellon University

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Learning Behaviors

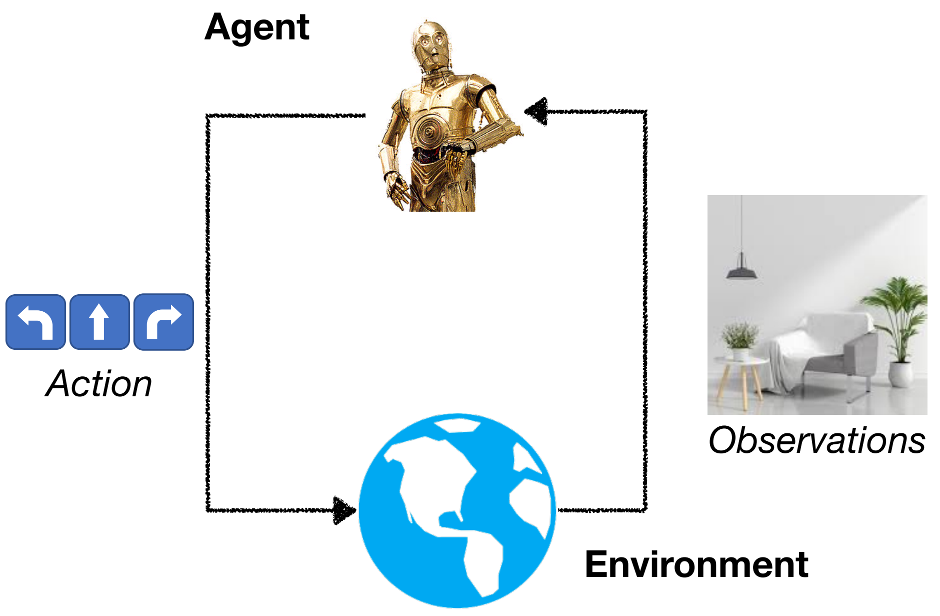
Observation → o_t



Action → a_t

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

Physical Intelligence

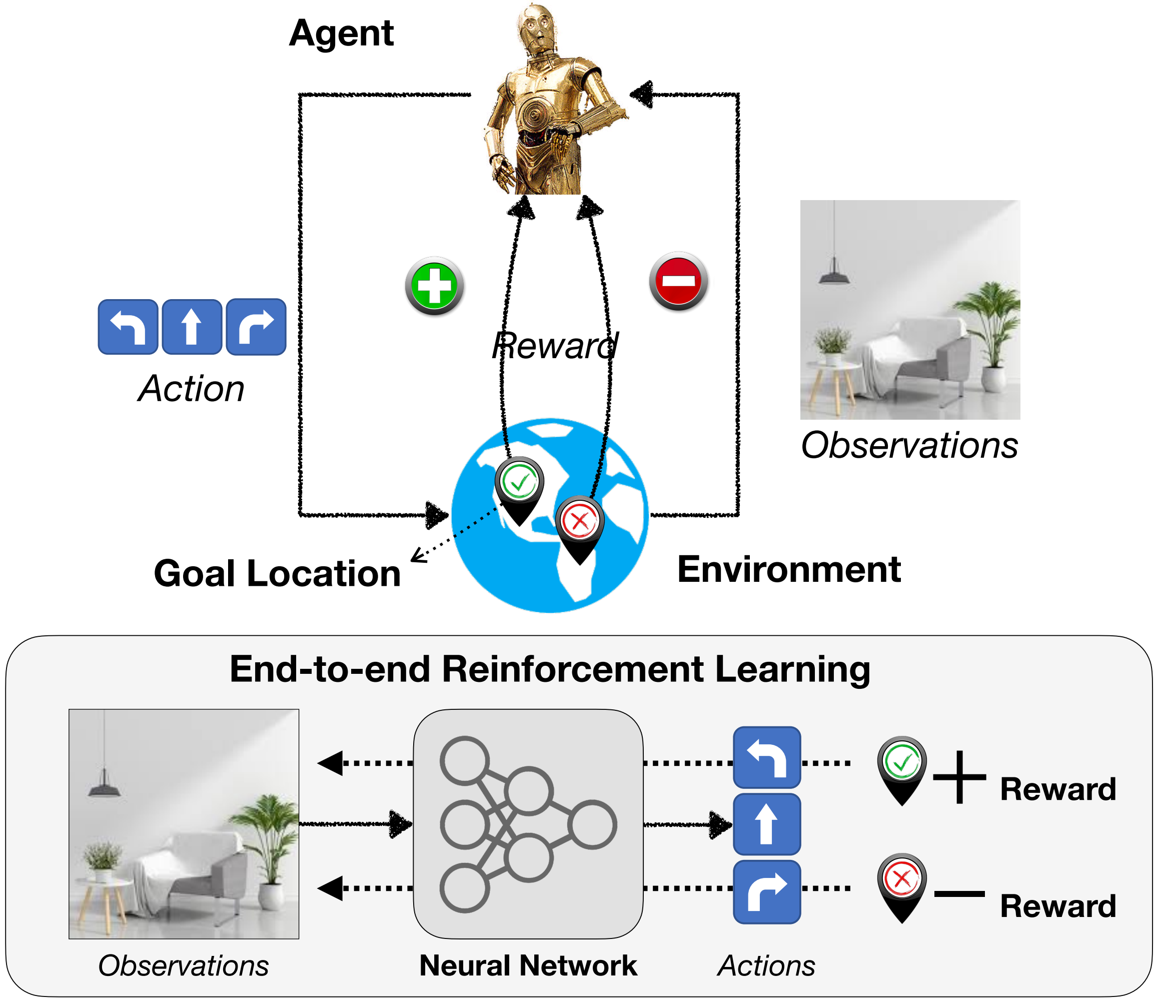


Agent needs to move in the world physically.

Current actions affect future observations.

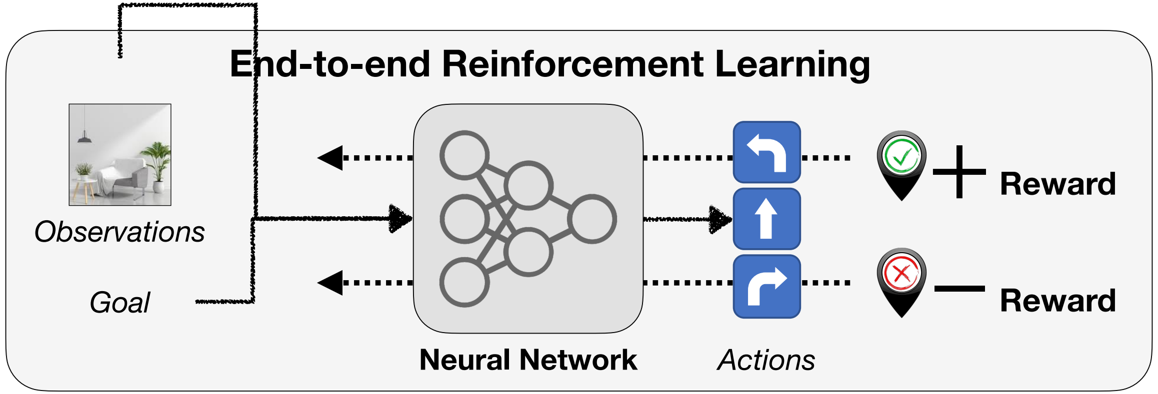
Require Spatial and Semantic Understanding.

Navigation



Agent
Action
Reward
Observations
Goal Location
Environment
End-to-end Reinforcement Learning
Observations
Neural Network
Actions
+ Reward
— Reward

Goal-conditioned Navigation



Point Goal

(x, y) *coordinate*

Image Goal

Object Goal

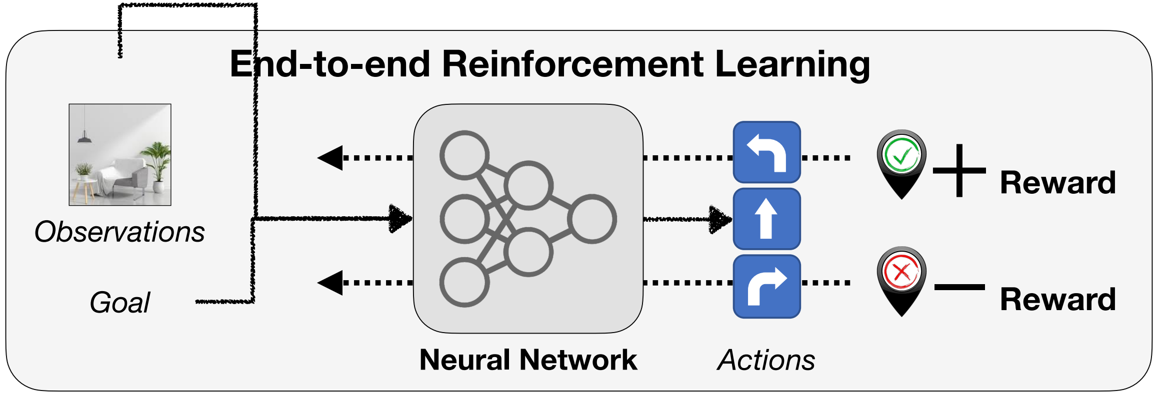
Chair
TV
Sofa

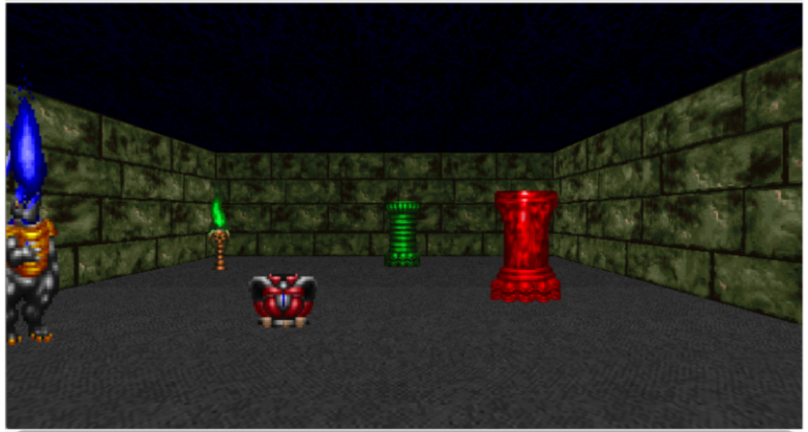
Language Goal

Blue Chair
Largest TV
White Sofa

- Convenient for humans
- Compositionality

Goal-conditioned Navigation





Go to the green torch

Train

Go to the short red torch
Go to the blue keycard
Go to the largest yellow object
Go to the green object

Test

Go to the tall green torch
Go to the red keycard
Go to the smallest blue object

Language Goal

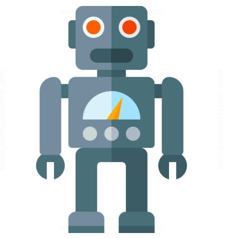
Blue Chair
Largest TV
White Sofa

- Convenient for humans
- Compositionality

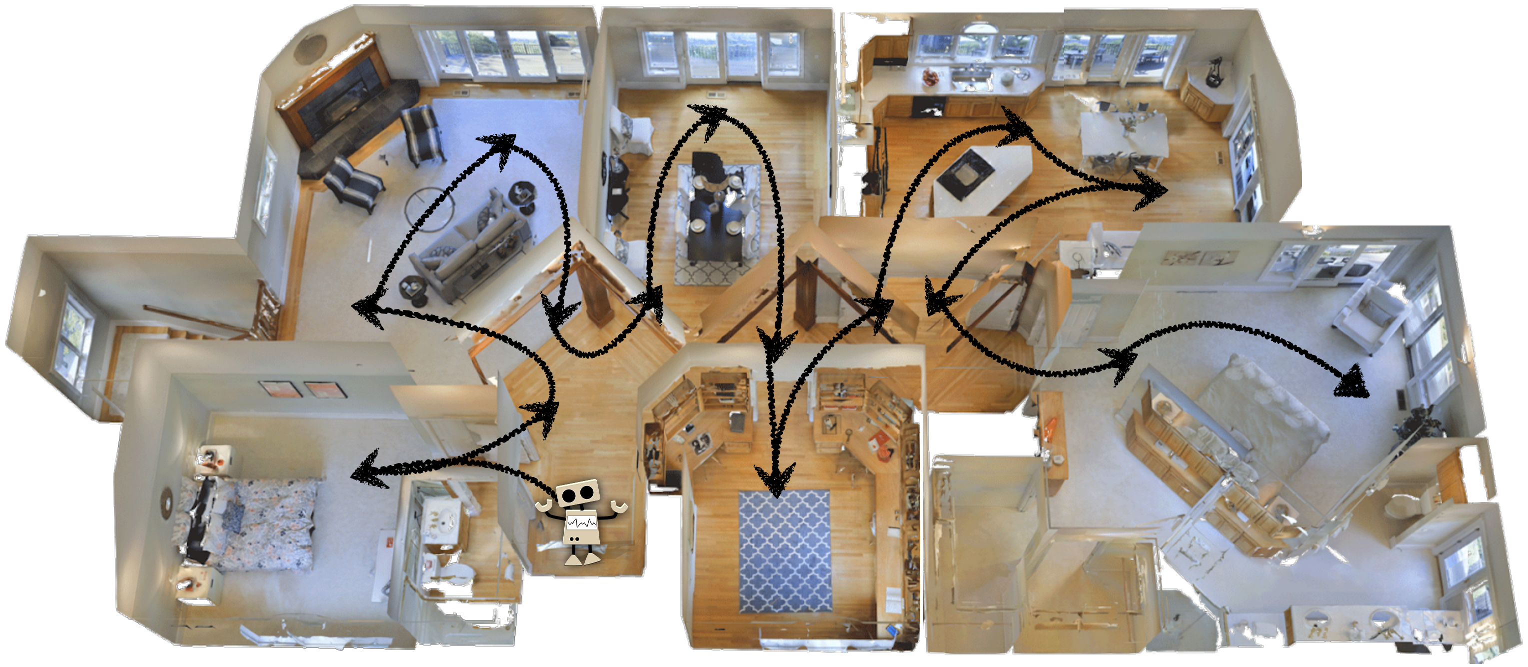
Navigation Tasks

Point Goal	Image Goal	Object Goal	Language Goal
 (x, y) *coordinate*		*Chair* *TV* *Sofa*	*Blue Chair* *Largest TV* *White Sofa*

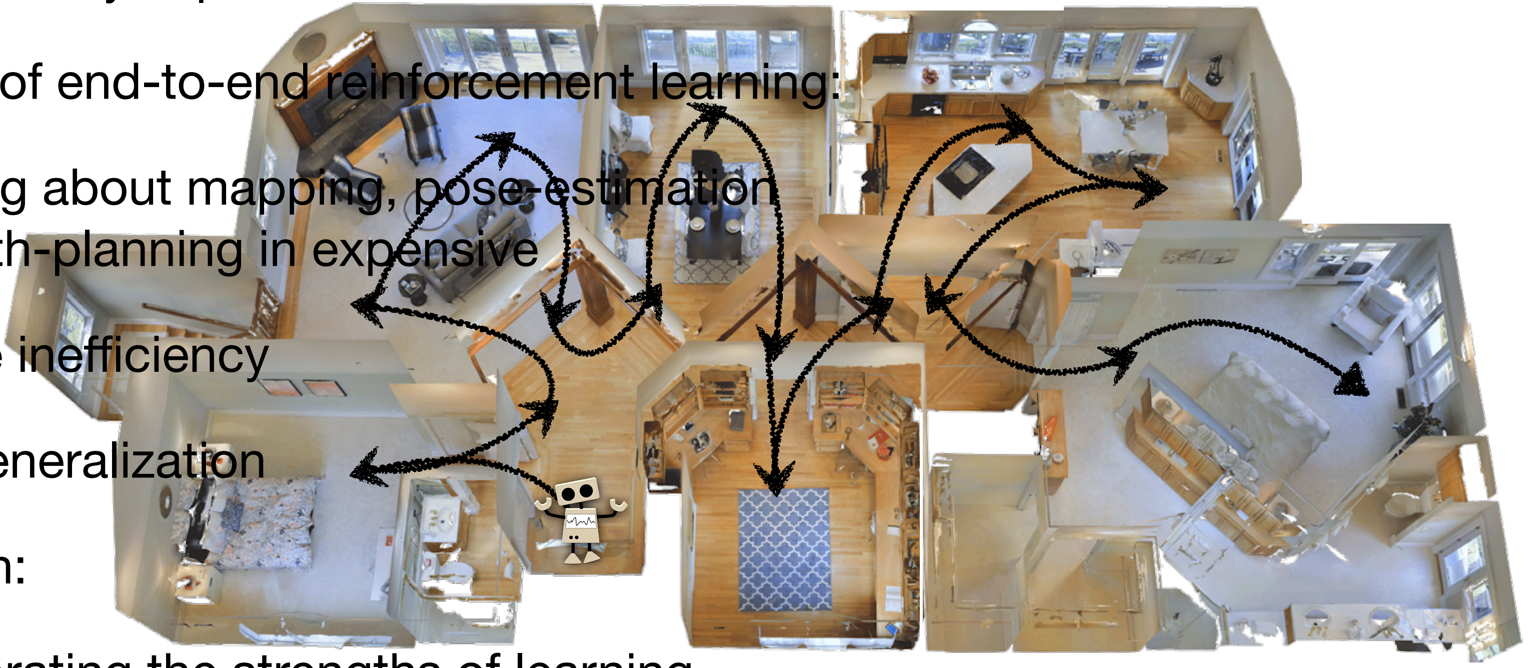
Require exploring the environment to find the goal



Exploration

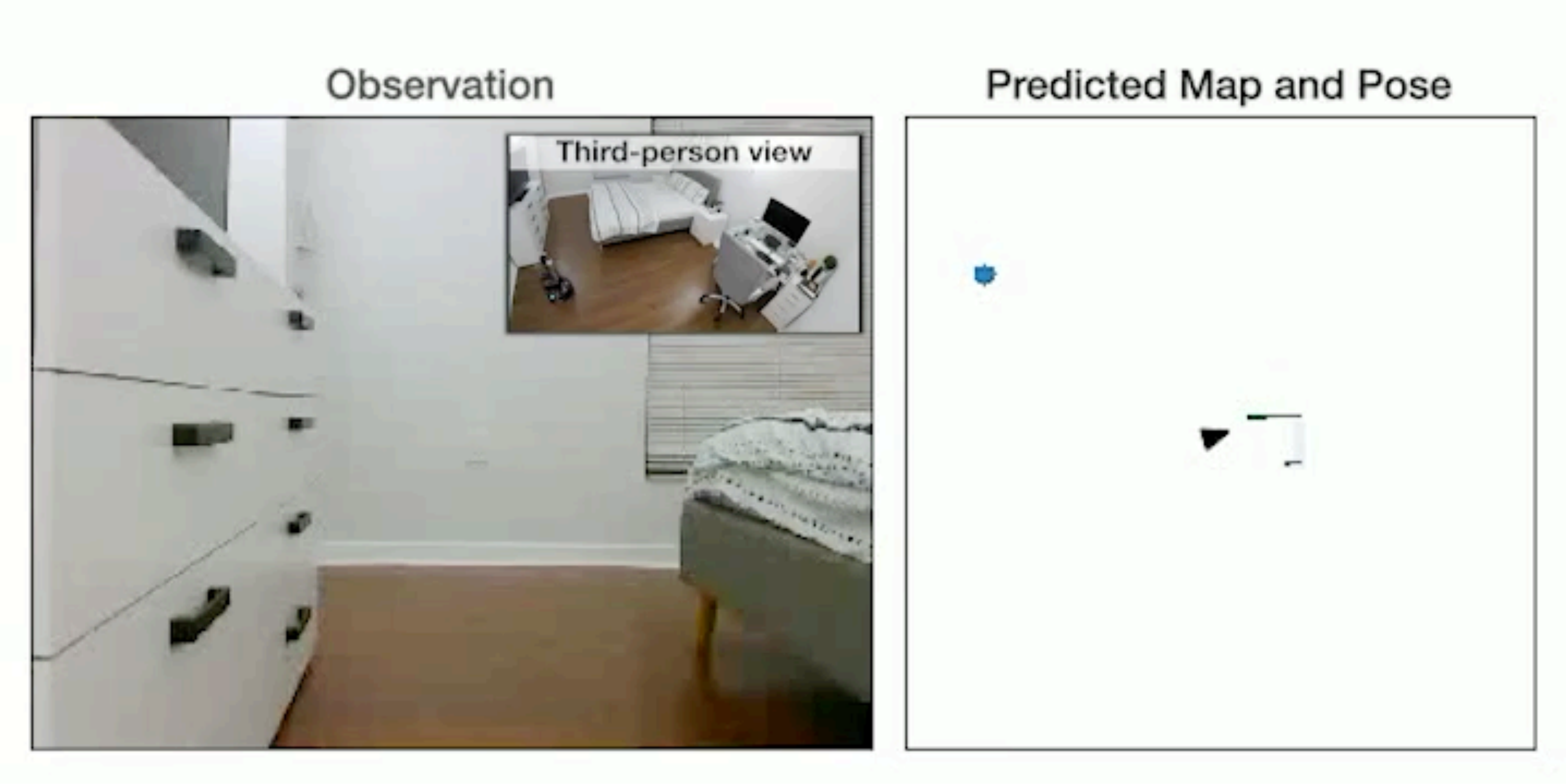


Exploration

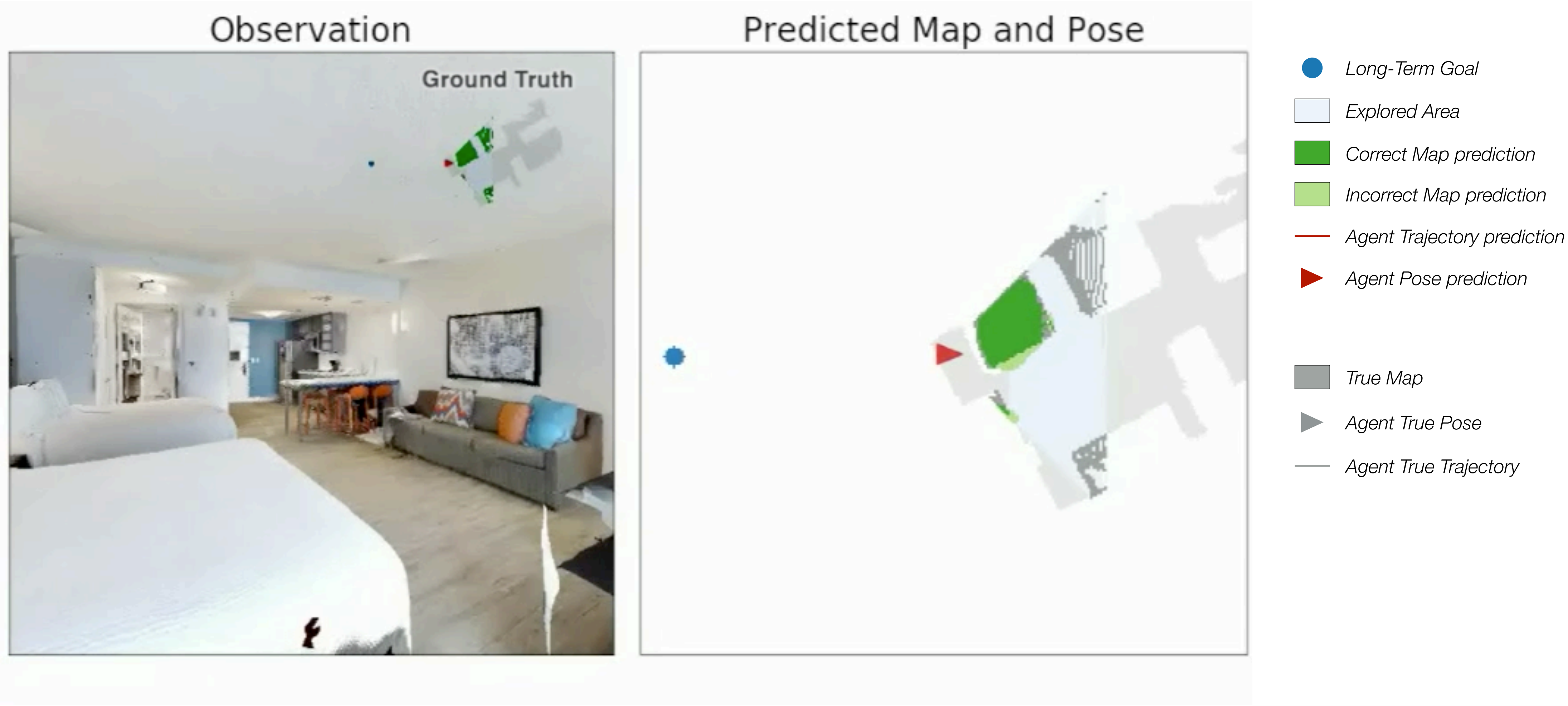


- How to efficiently explore an unseen environment?
- Limitations of end-to-end reinforcement learning:
  - Learning about mapping, pose-estimation and path-planning in expensive
  - Sample inefficiency
  - Poor generalization
- Our solution:
  - Incorporating the strengths of learning
  - Modular and hierarchical system

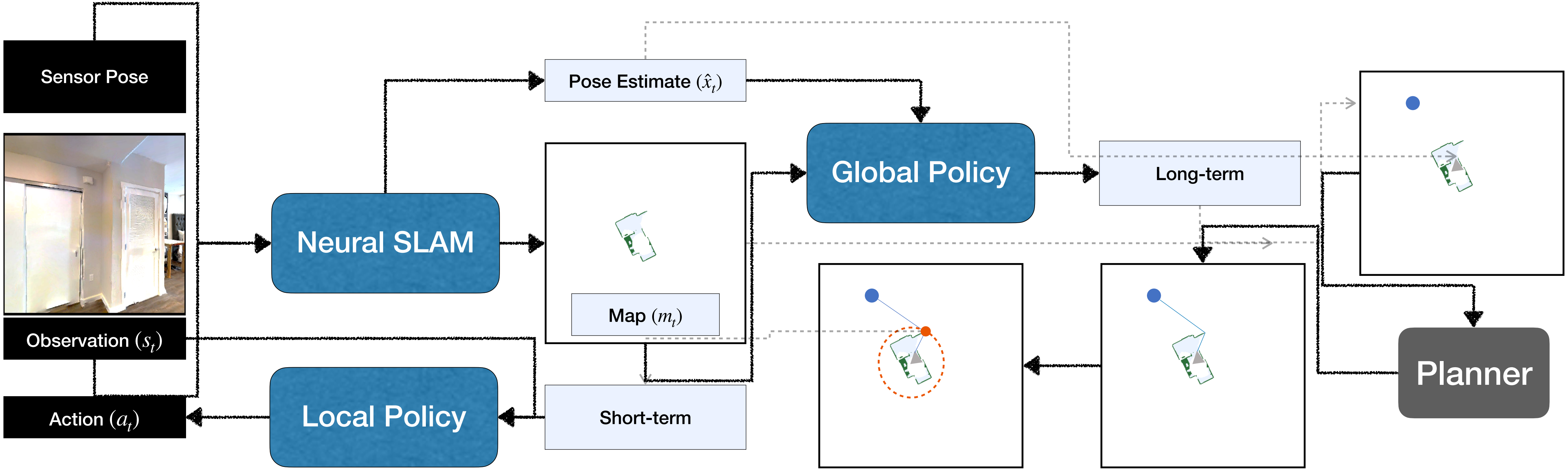
Preview: Visual Navigation in the Real World



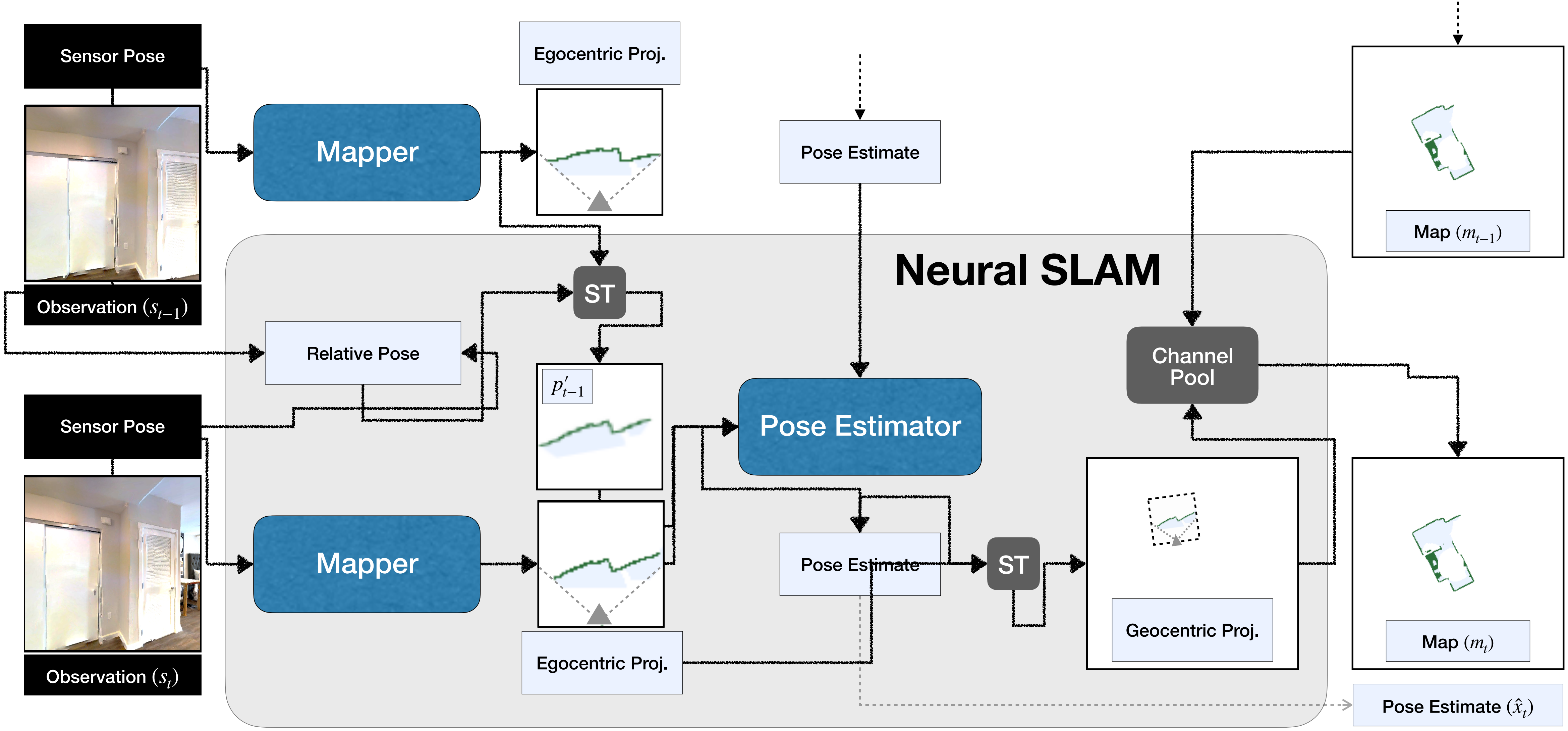
Exploration in Gibson Environment



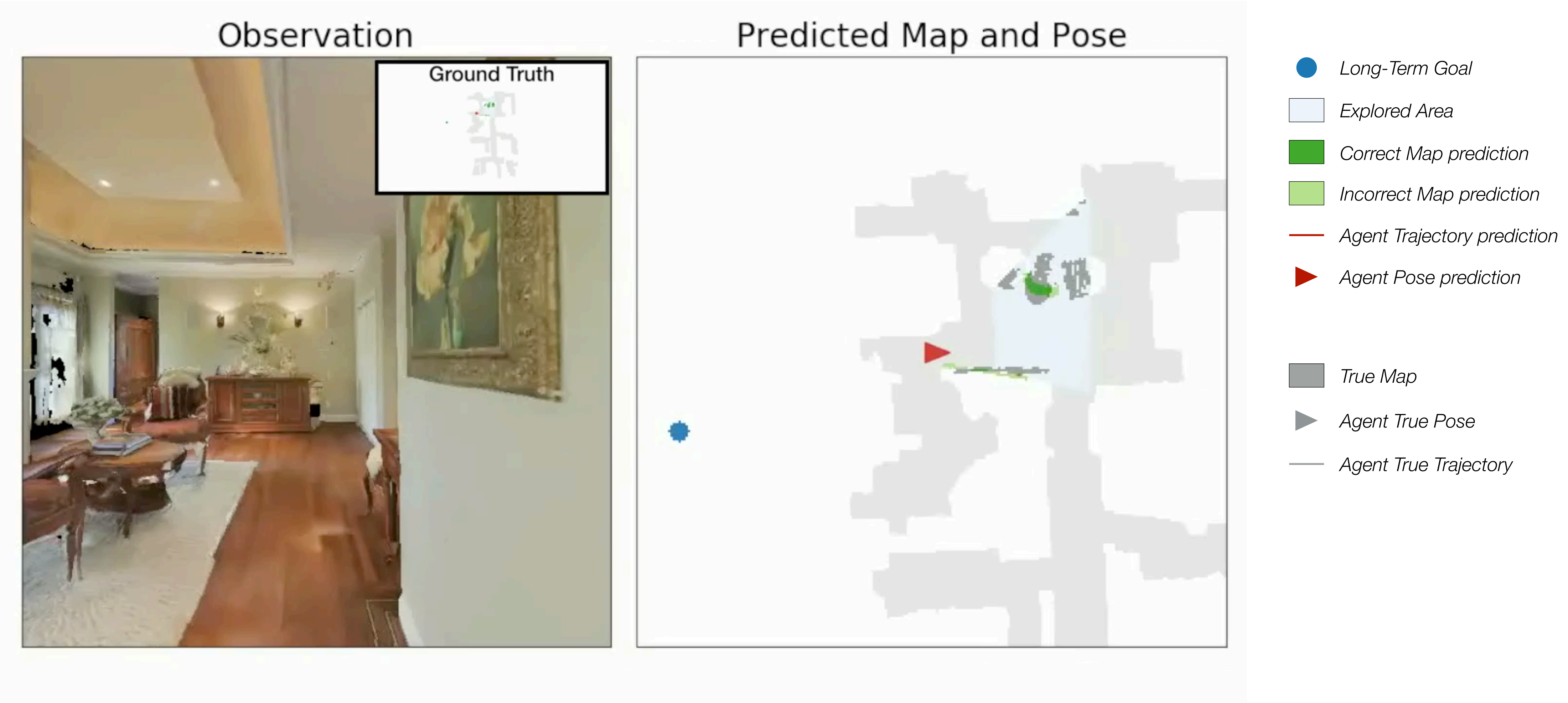
Active Neural SLAM: Overview



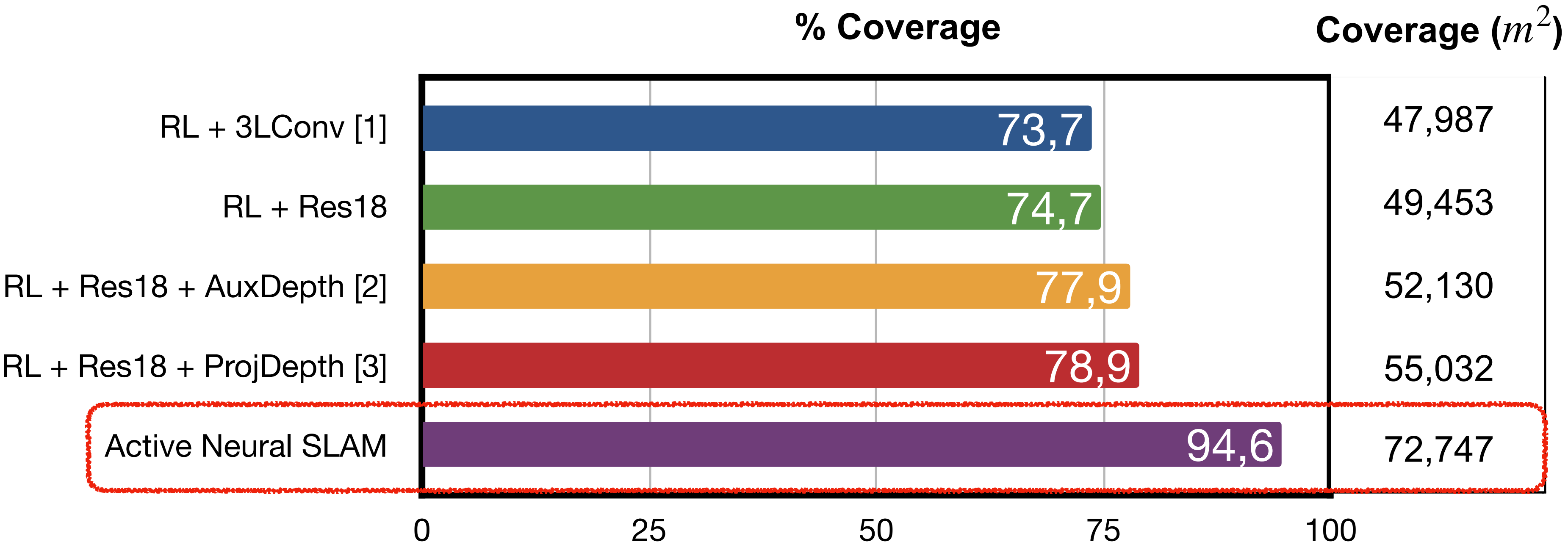
Neural SLAM Module



Domain Generalization: Matterport3D



Exploration Results



Gibson

Domain Generalization

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

Goal-conditioned Navigation

Point Goal

(x, y) *coordinate*

Image Goal

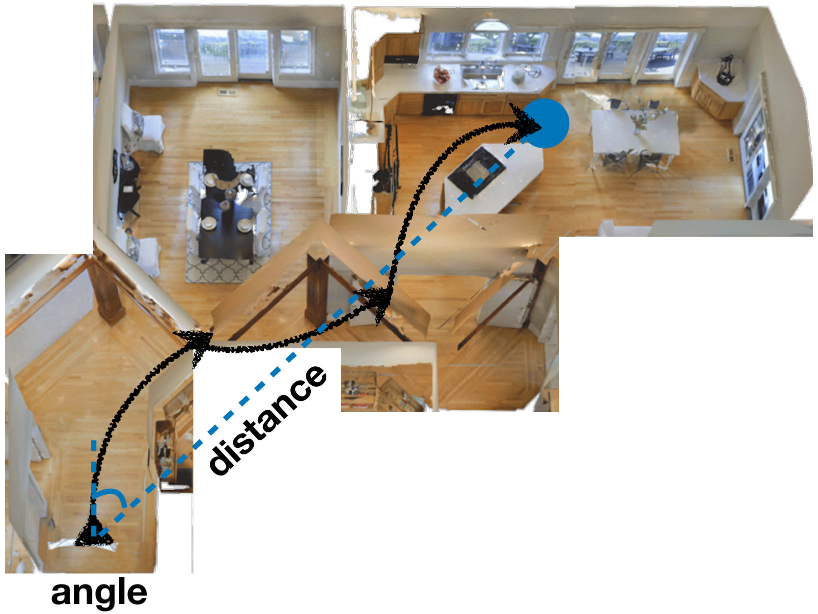
Object Goal

Chair
TV
Sofa

Language Goal

Blue Chair
Largest TV
White Sofa

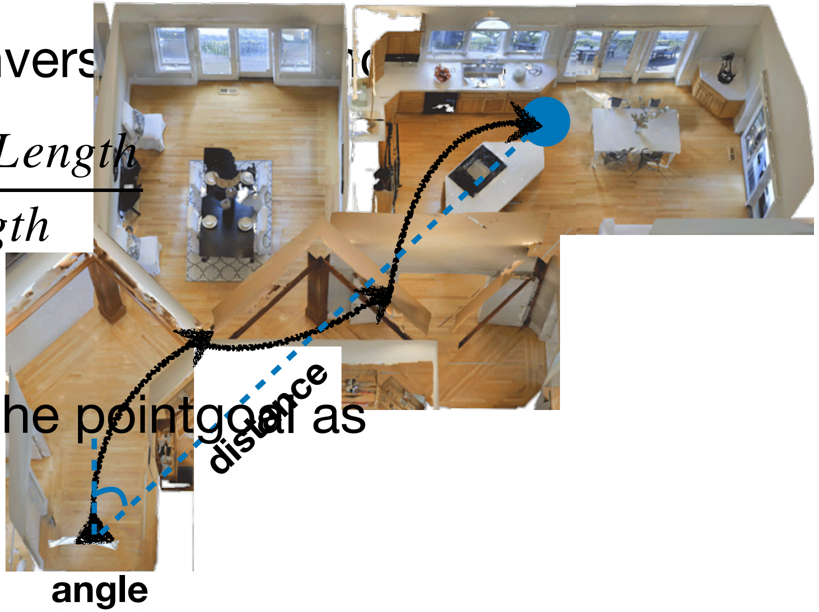
Point-Goal Navigation



Point-Goal Navigation

- Objective: Navigate to goal coordinates
- Metric: Success weighted by invers

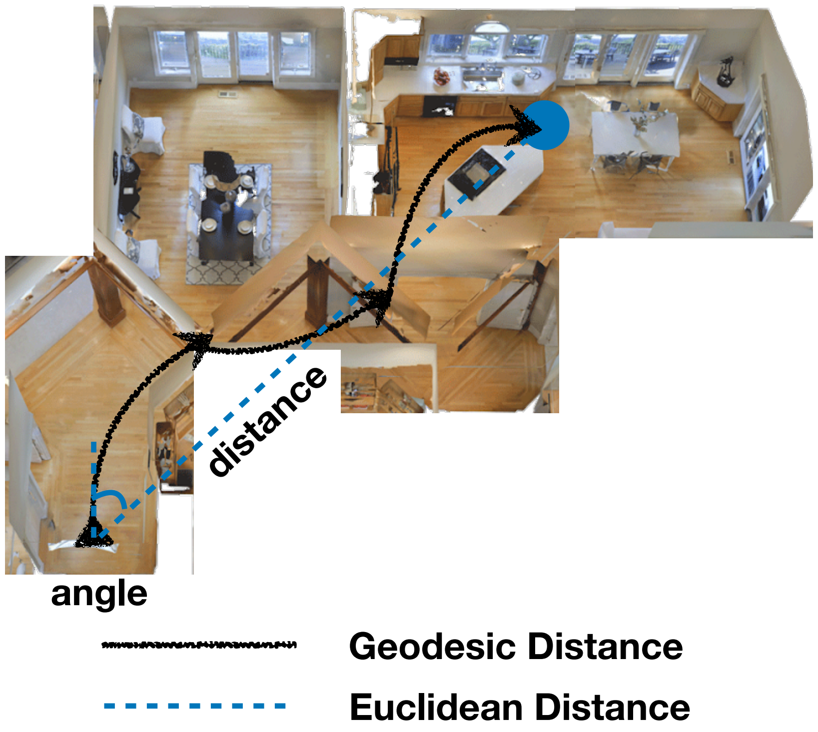
$$\frac{1}{N}\sum_{i=1}^{N} Success * \frac{ShortestPathLength}{PathLength}$$



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

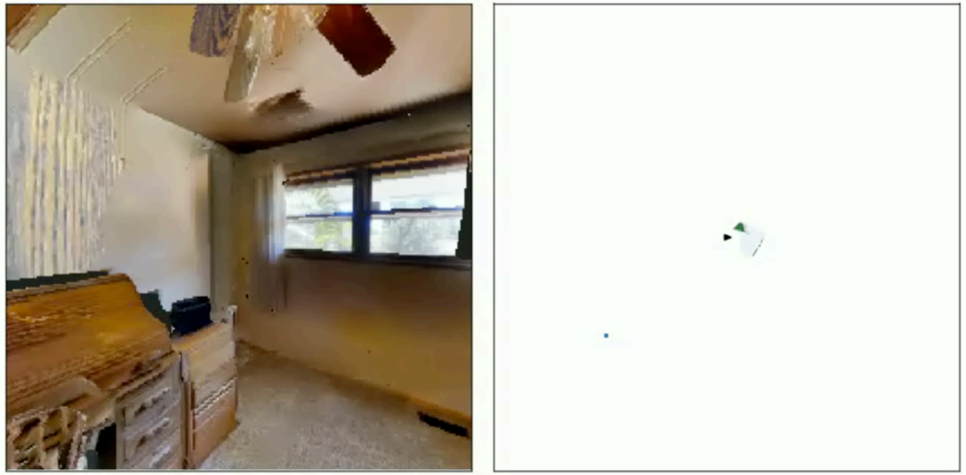
Harder Datasets

- **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



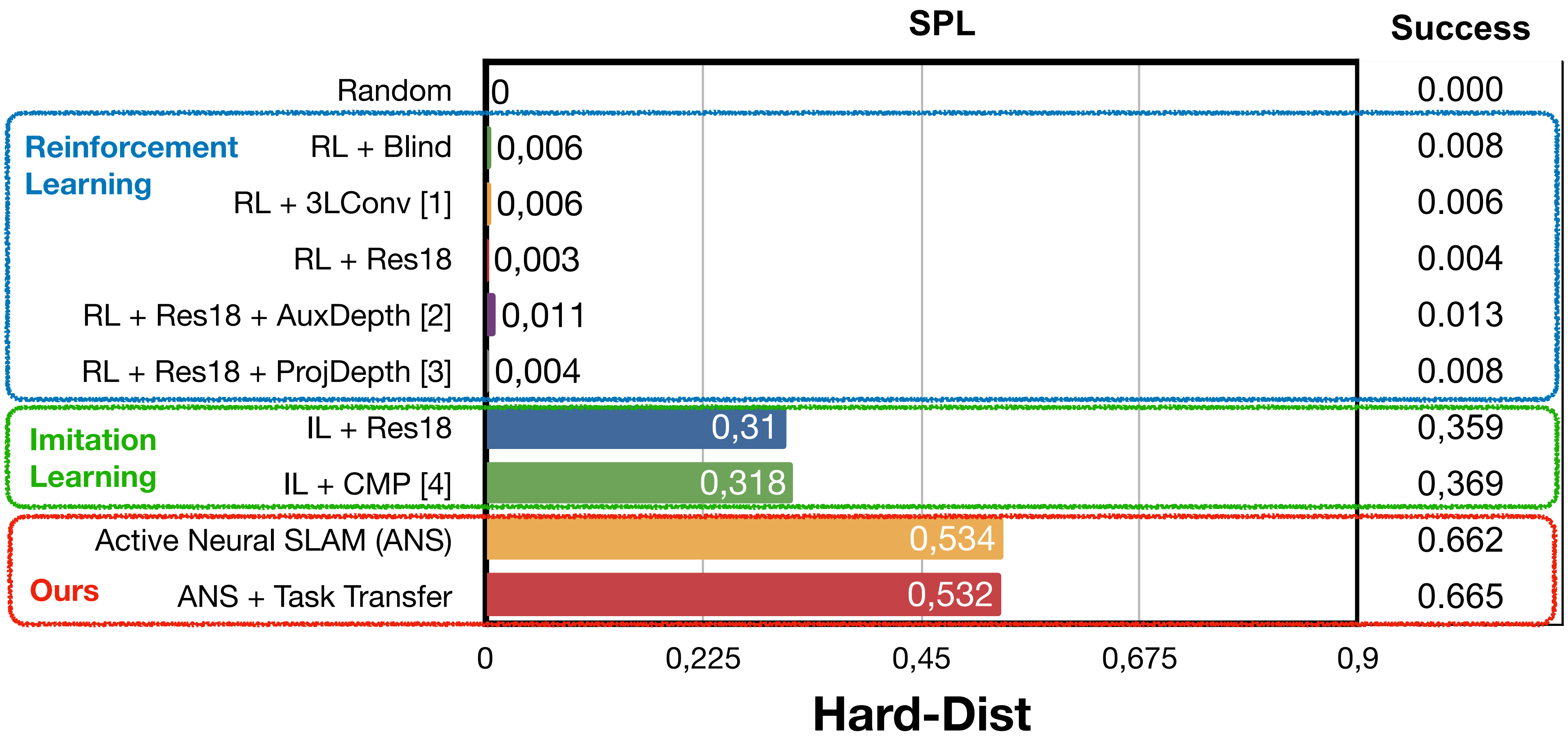
Point-Goal Navigation

Gibson



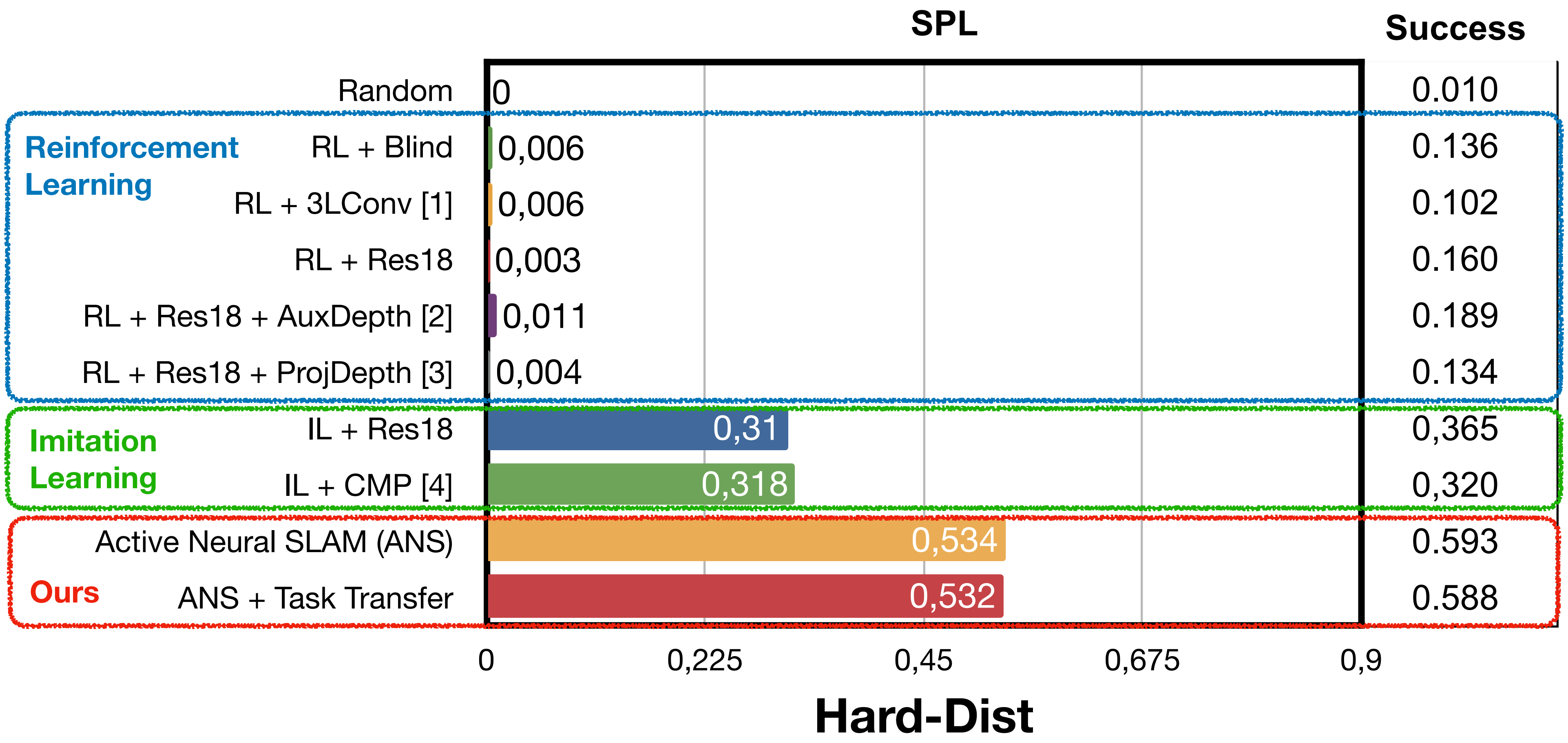
MP3D

Results



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

Navigation Tasks

Point Goal

(x, y) *coordinate*

Image Goal

Object Goal

Chair
TV
Sofa

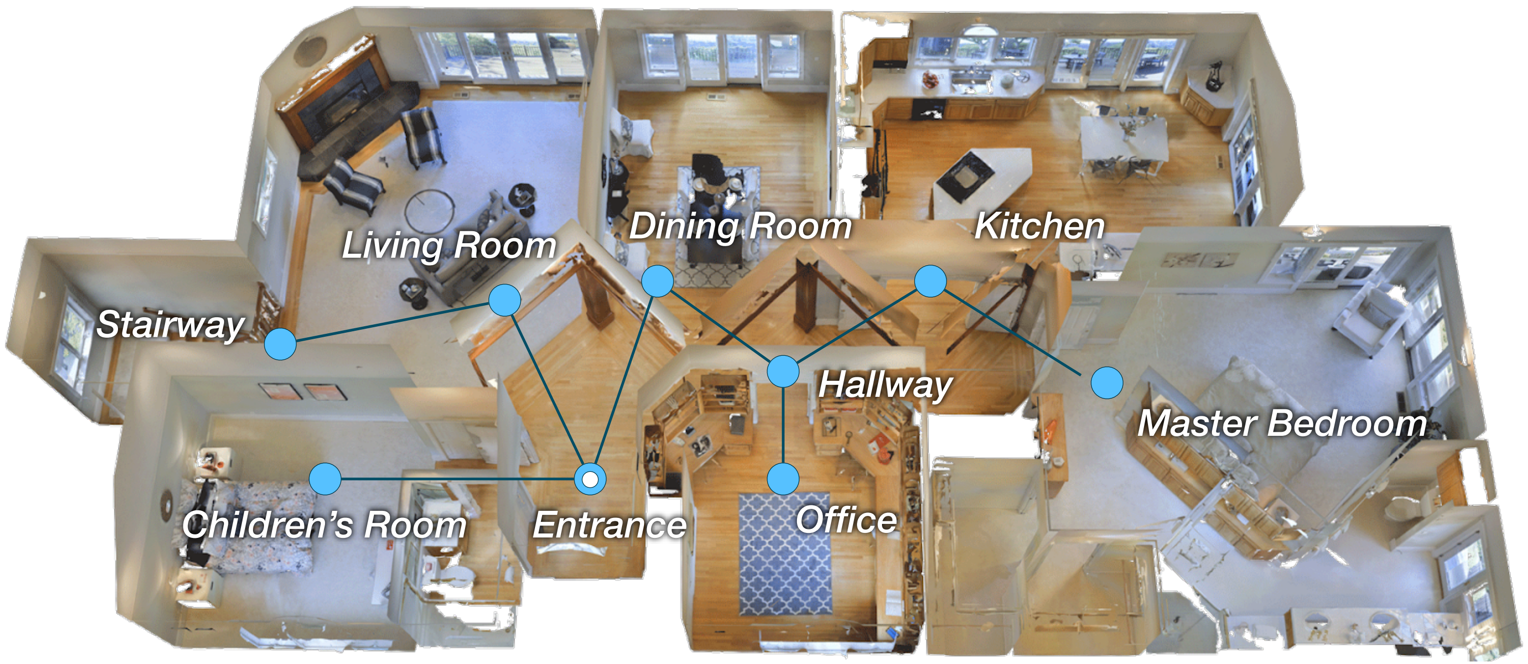
Language Goal

Blue Chair
Largest TV
White Sofa

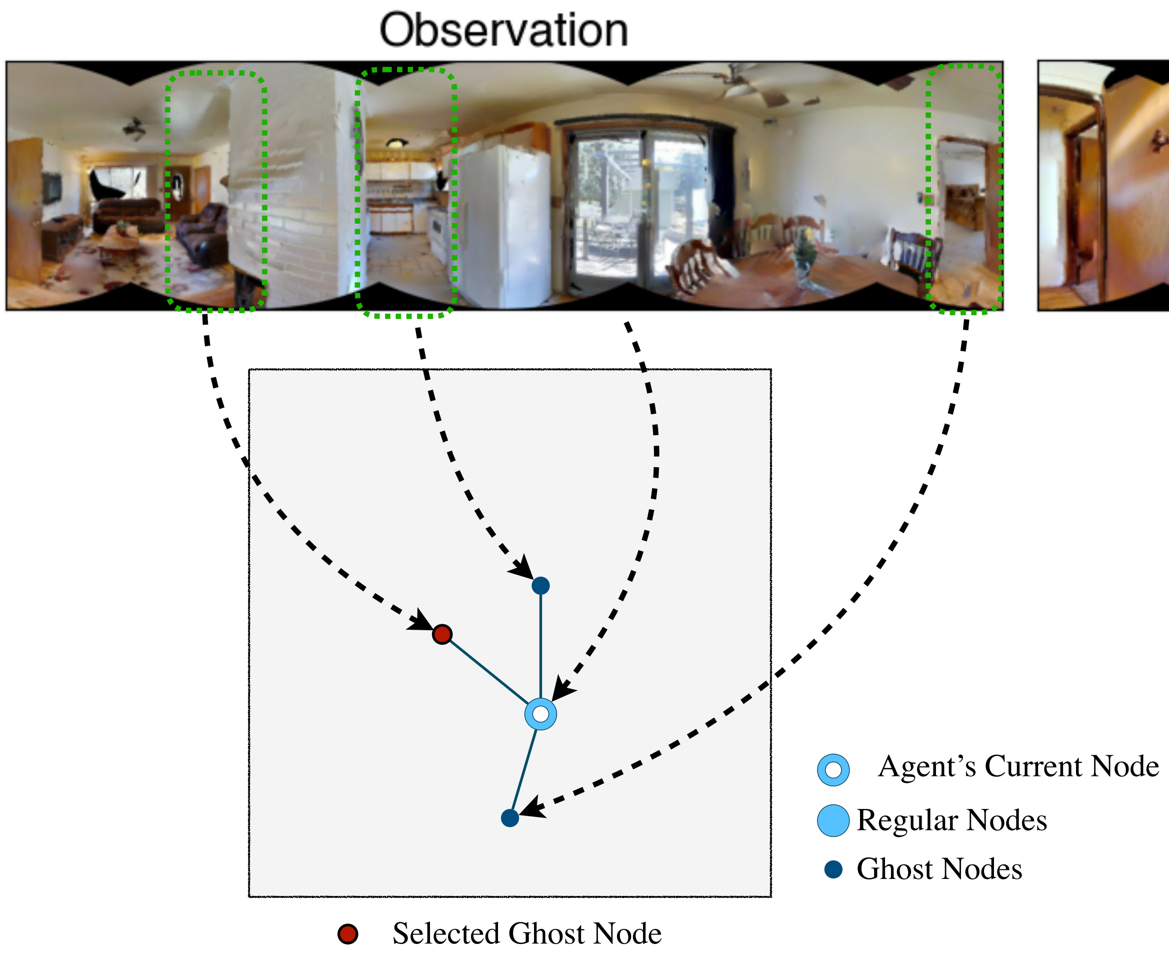
Semantic Priors and Common-Sense

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

Topological Maps

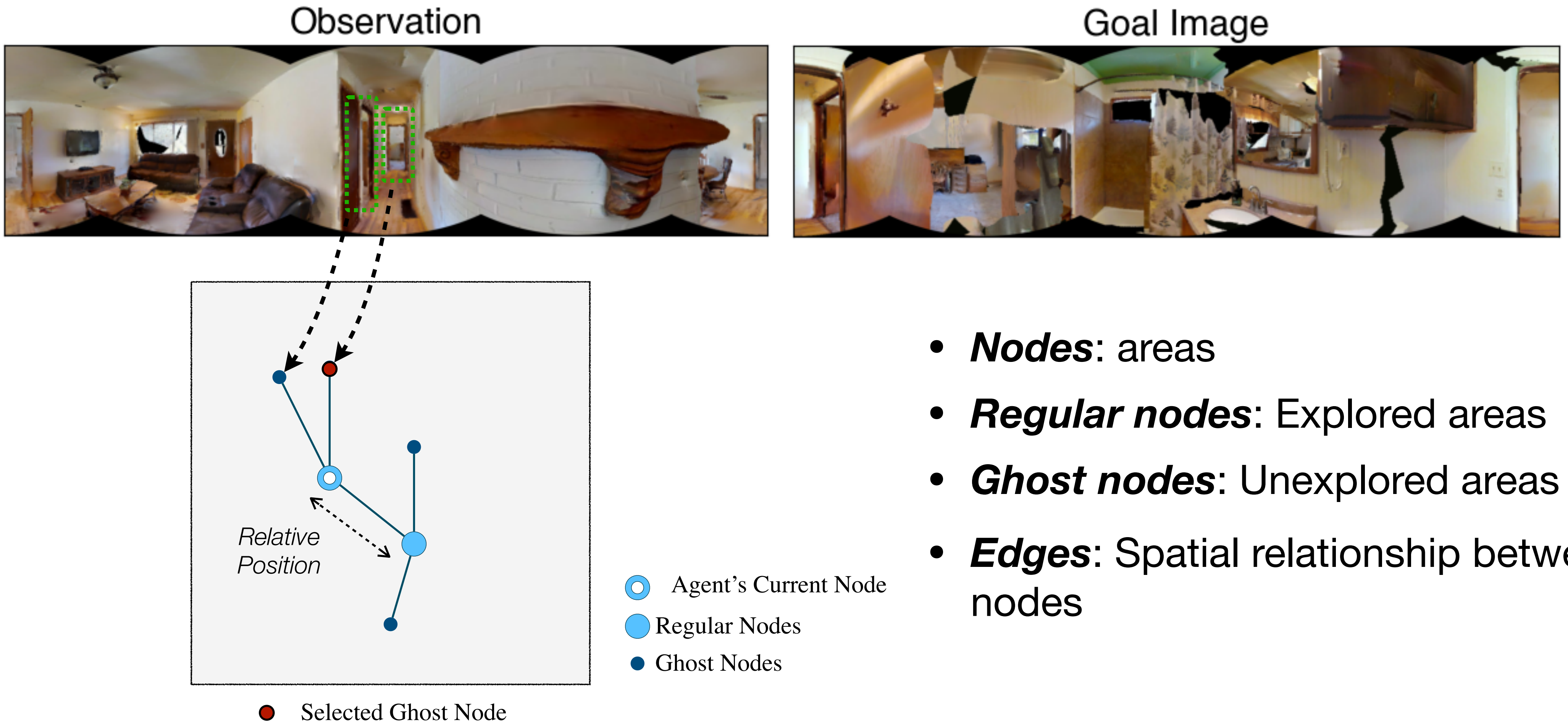


Topological Graph Representation



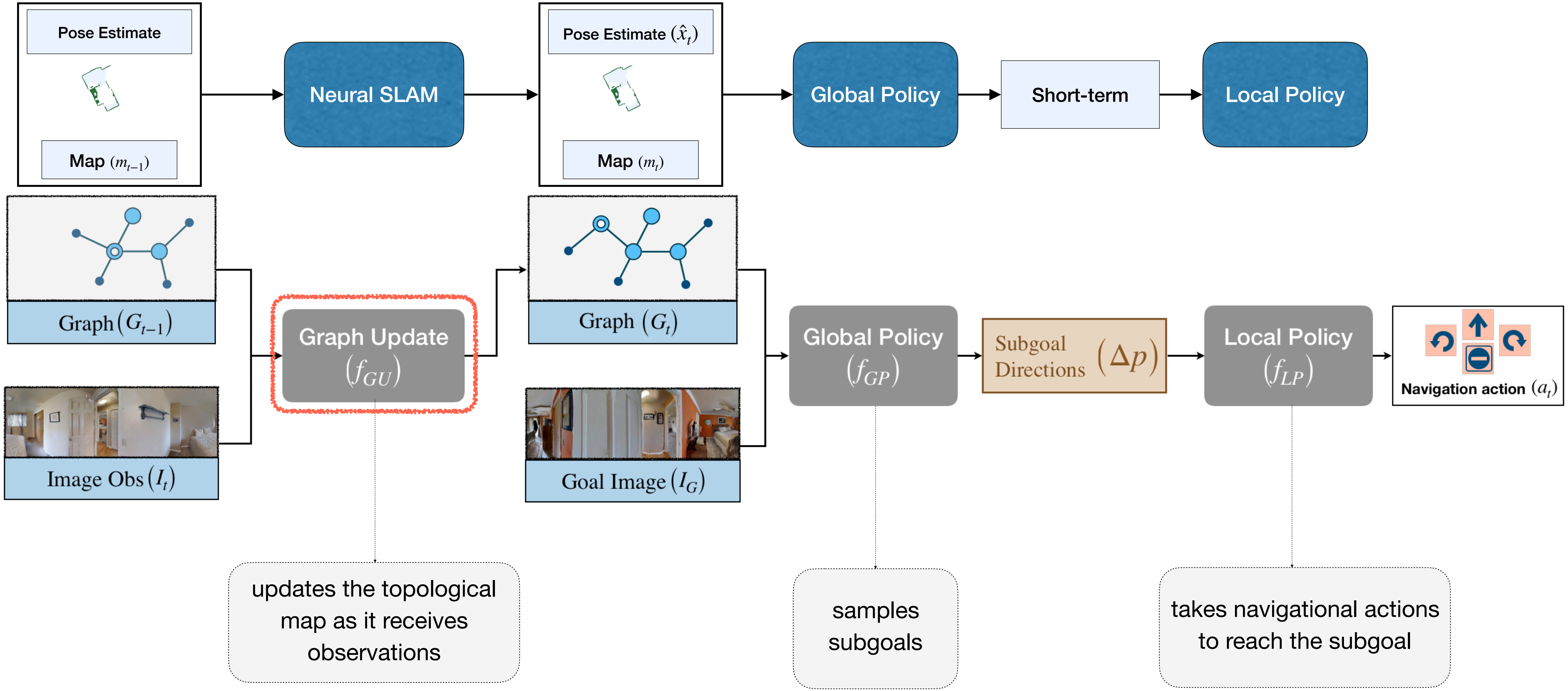
- ***Nodes***: areas
- ***Regular nodes***: Explored areas
- ***Ghost nodes***: Unexplored areas

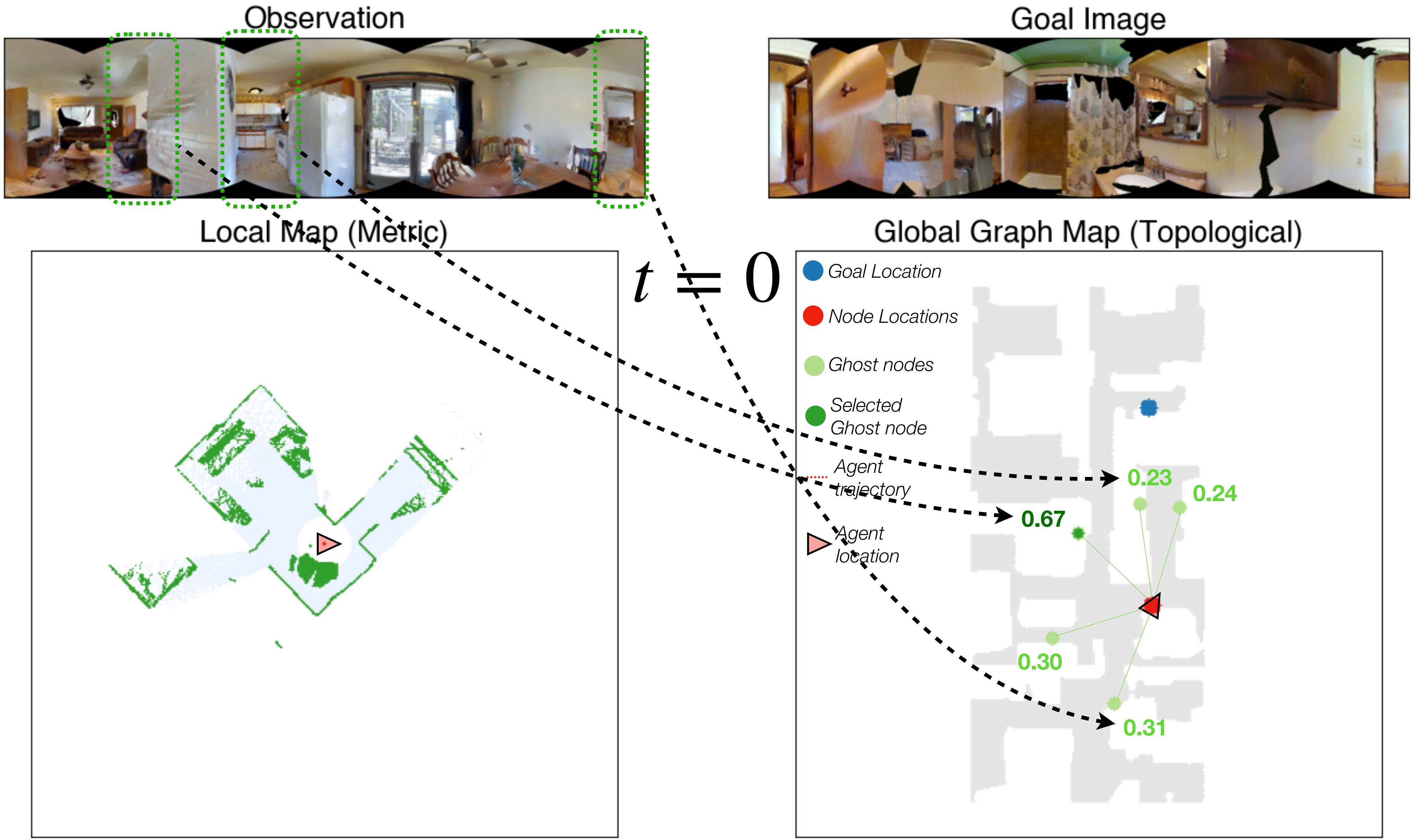
Topological Graph Representation



- ***Nodes***: areas
- ***Regular nodes***: Explored areas
- ***Ghost nodes***: Unexplored areas
- ***Edges***: Spatial relationship between nodes

Neural Topological SLAM




Observation
Goal Image
Local Map (Metric)
Global Graph Map (Topological)
$t = 0$
Goal Location
Node Locations
Ghost nodes
Selected Ghost node
Agent trajectory
Agent location
0.23
0.24
0.67
0.30
0.31

Observation

Goal Image

Local Map (Metric)

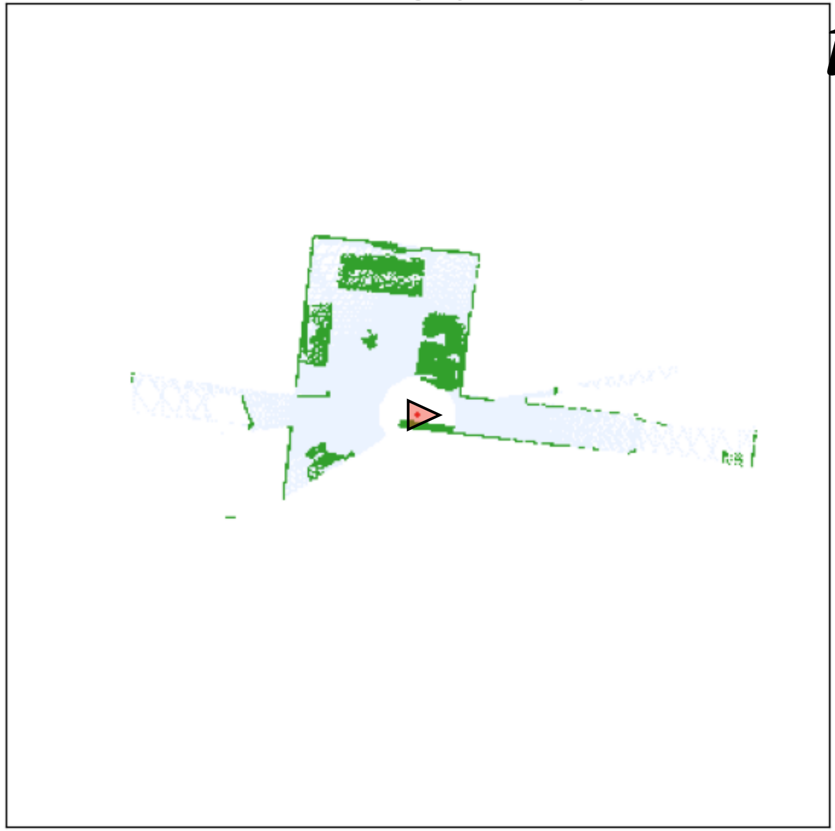
$t = 29$

Global Graph Map (Topological)

Observation

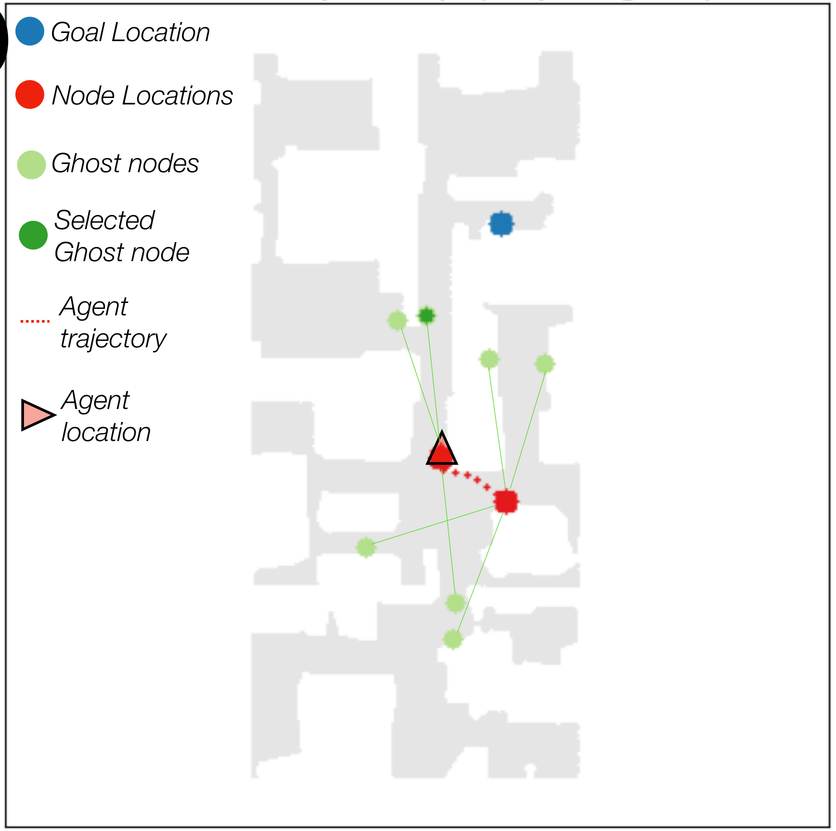
Goal Image

Local Map (Metric)



$t = 30$

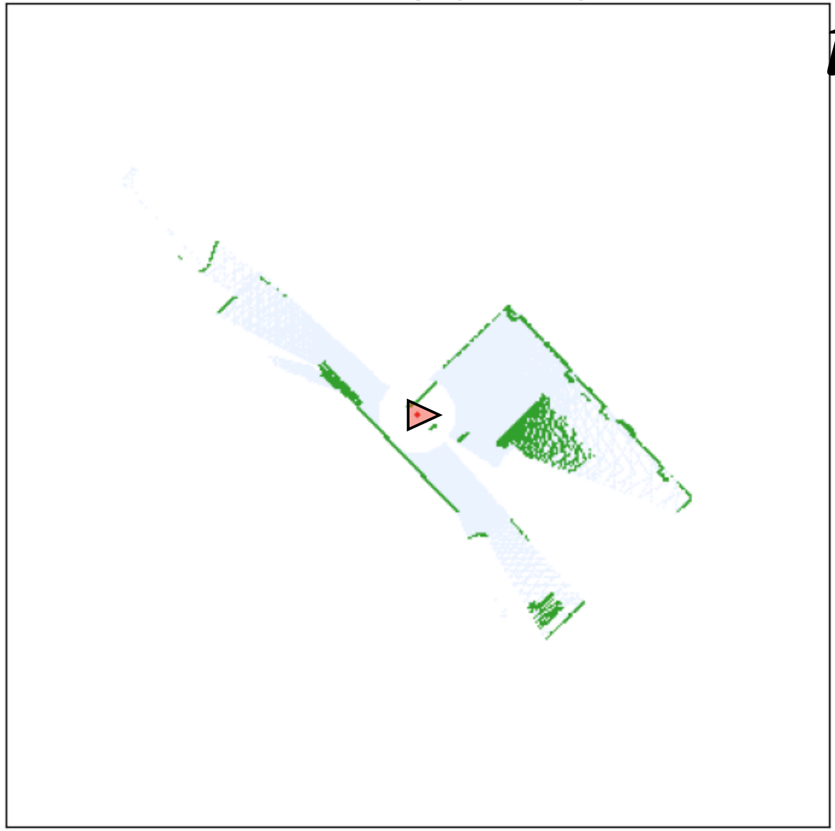
Global Graph Map (Topological)



Observation

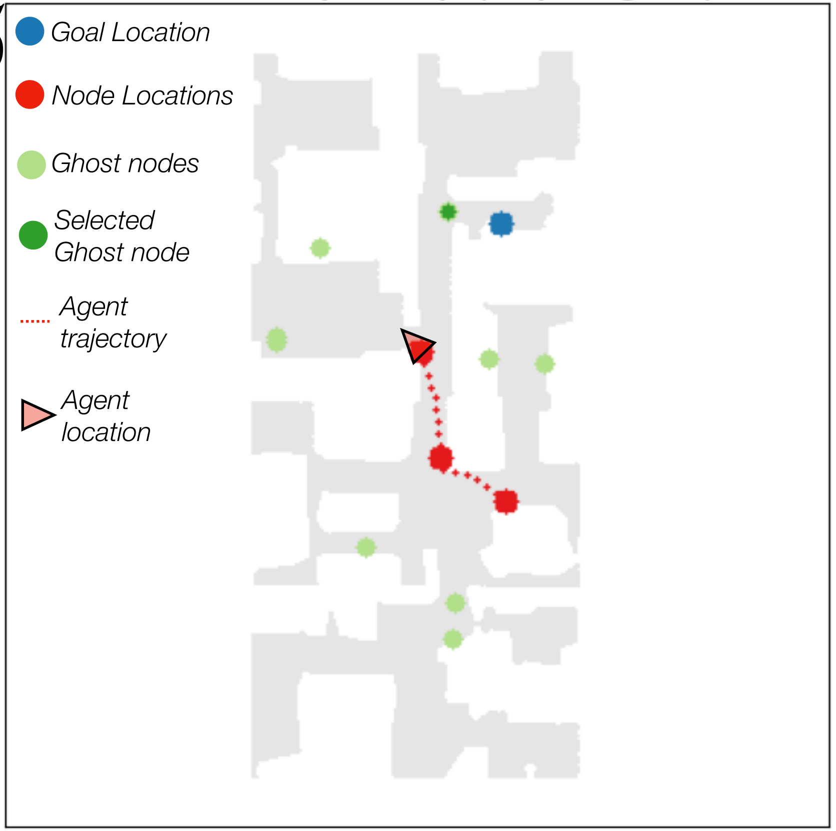
Goal Image

Local Map (Metric)



$t = 45$

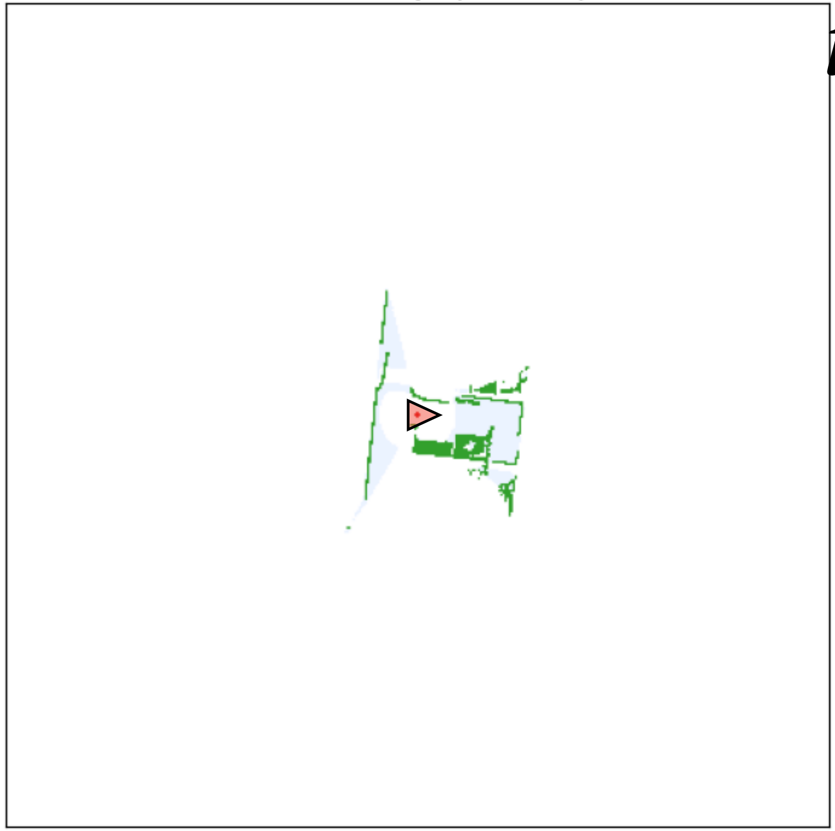
Global Graph Map (Topological)



Observation

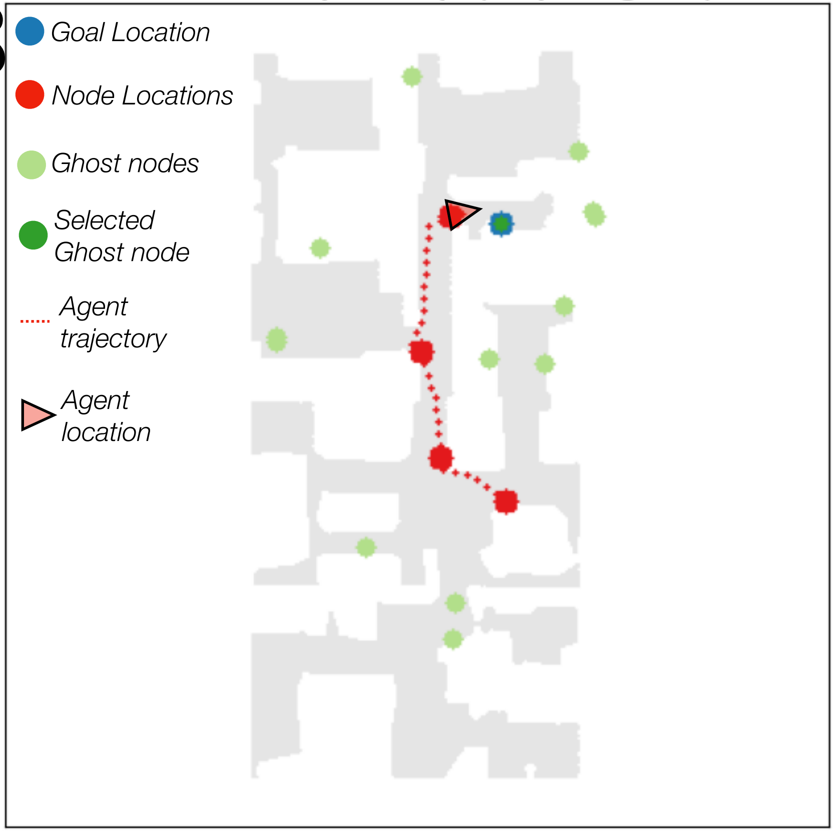
Goal Image

Local Map (Metric)



$t = 78$

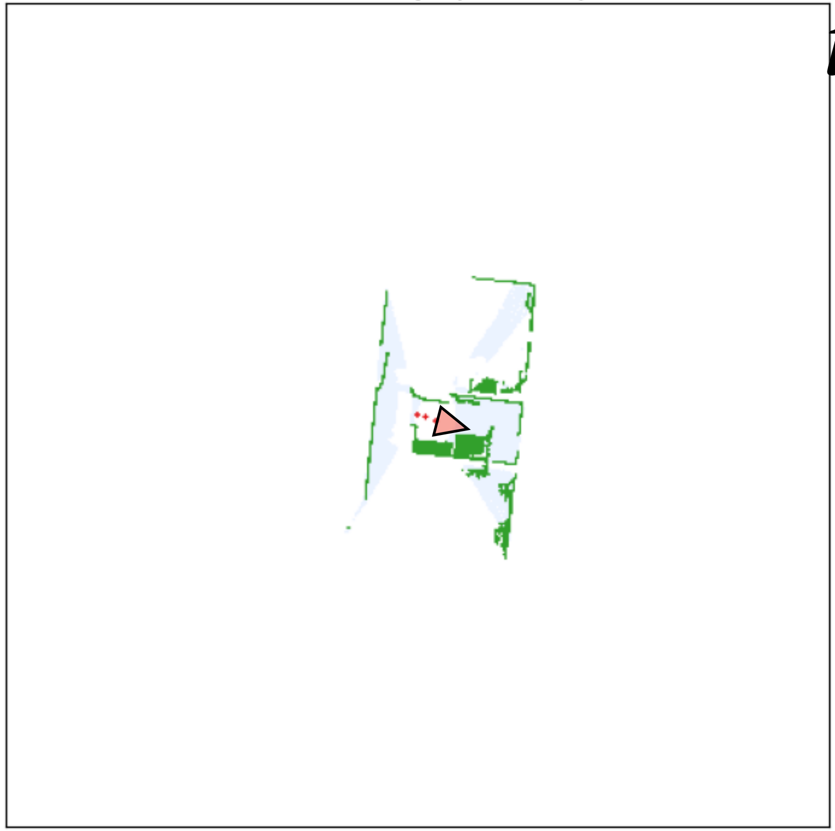
Global Graph Map (Topological)



Observation

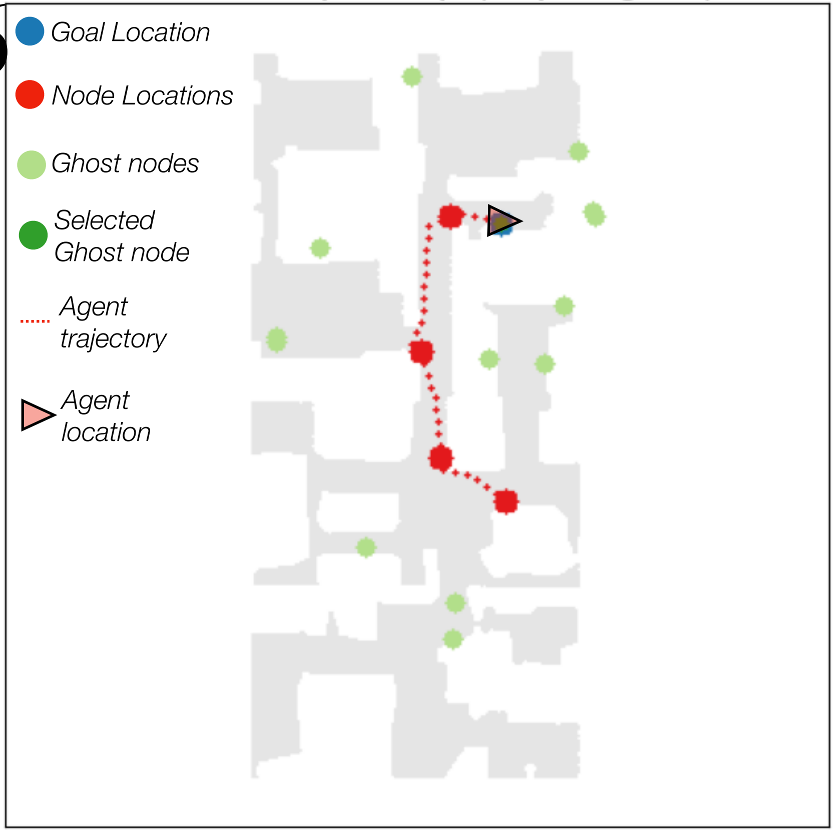
Goal Image

Local Map (Metric)



$t = 86$

Global Graph Map (Topological)

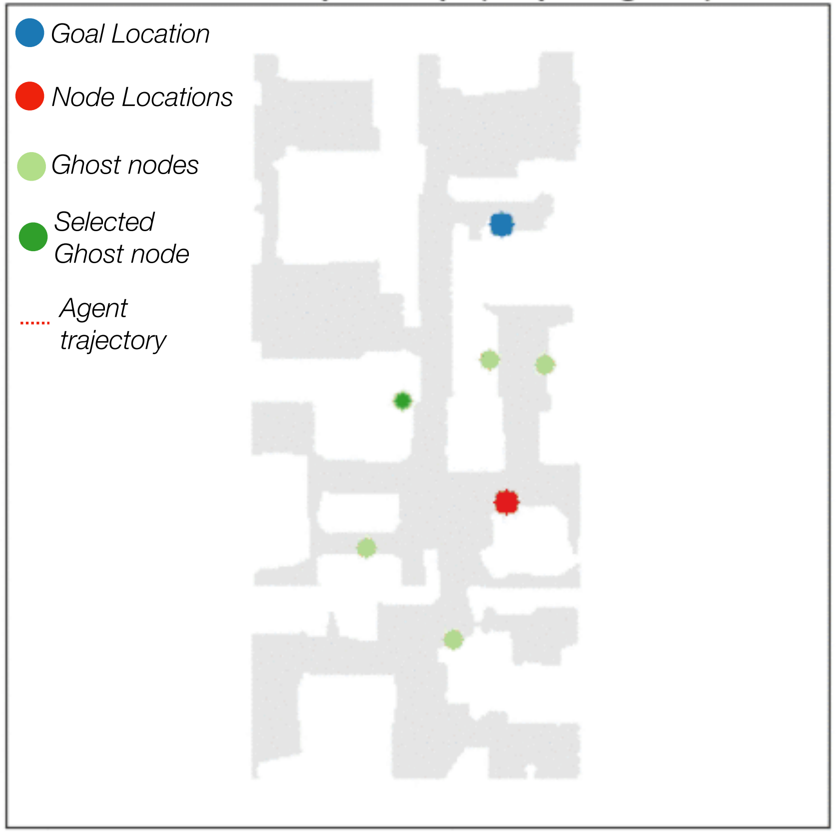


Observation

Goal Image

Local Map (Metric)

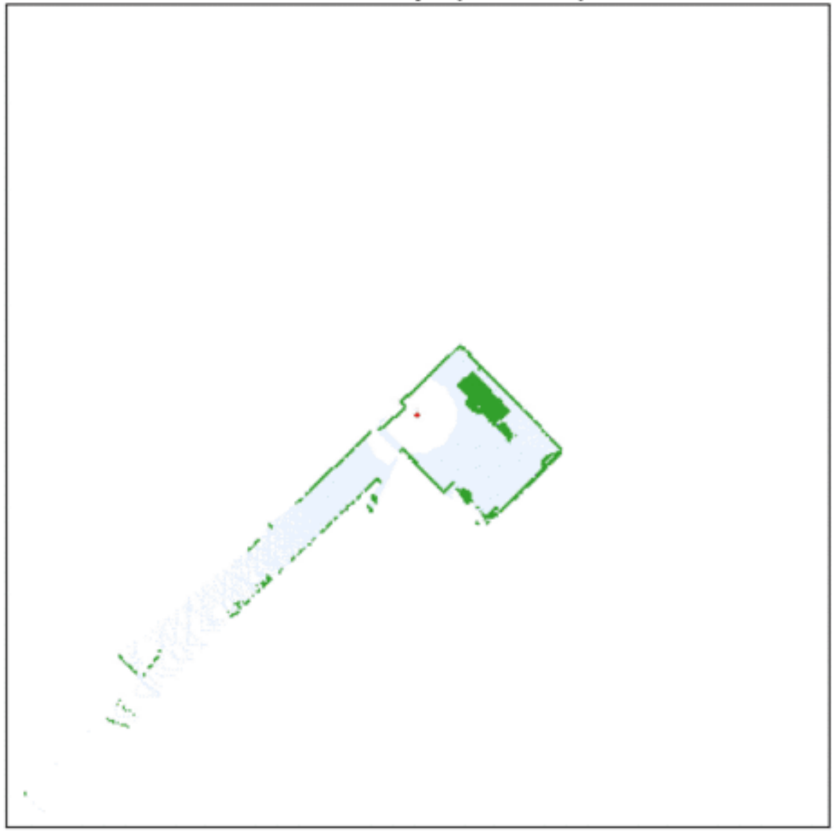
Global Graph Map (Topological)



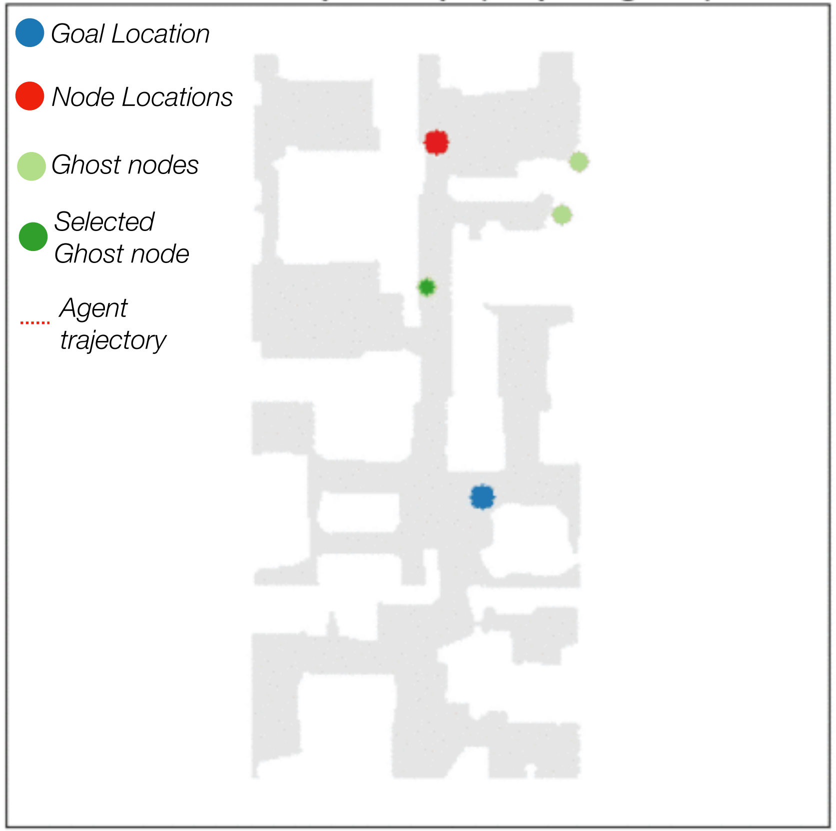
Observation

Goal Image

Local Map (Metric)



Global Graph Map (Topological)



Results

Robustness to Pose Noise

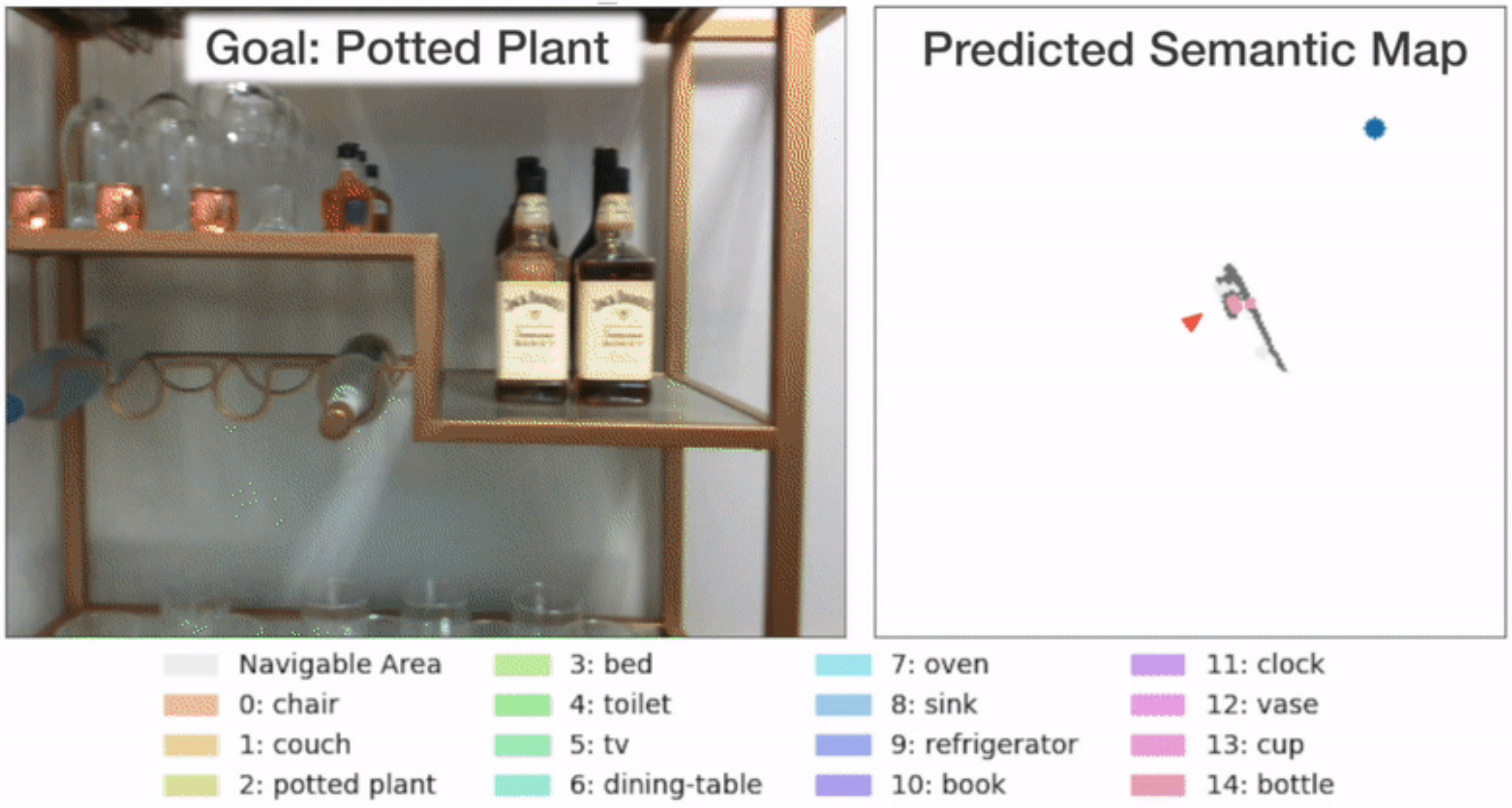
		RGB	RGBD	RGBD (No Noise)	RGBD (No Stop)
End-to-end Learning	LSTM + Imitation	0,10	0,14	0,15	0,18
	LSTM + RL	0,10	0,13	0,14	0,17
Modular Metric Maps	Occupancy Maps + FBE + RL	N/A	0,26	0,31	0,24
	Active Neural SLAM	0,23	0,29	0,35	0,39
Topological Maps	**Neural Topological SLAM**	**0,38**	**0,43**	**0,45**	**0,60**

Map based methods are better than vanilla learning methods even in presence of noise

NTS is better than occupancy map models, captures and uses semantic priors.

Explicit Semantic Mapping

Explicit Semantic Mapping



Internet vs Embodied Data

Static Internet Data

Active Embodied Data

Using Internet models for Embodied Agents

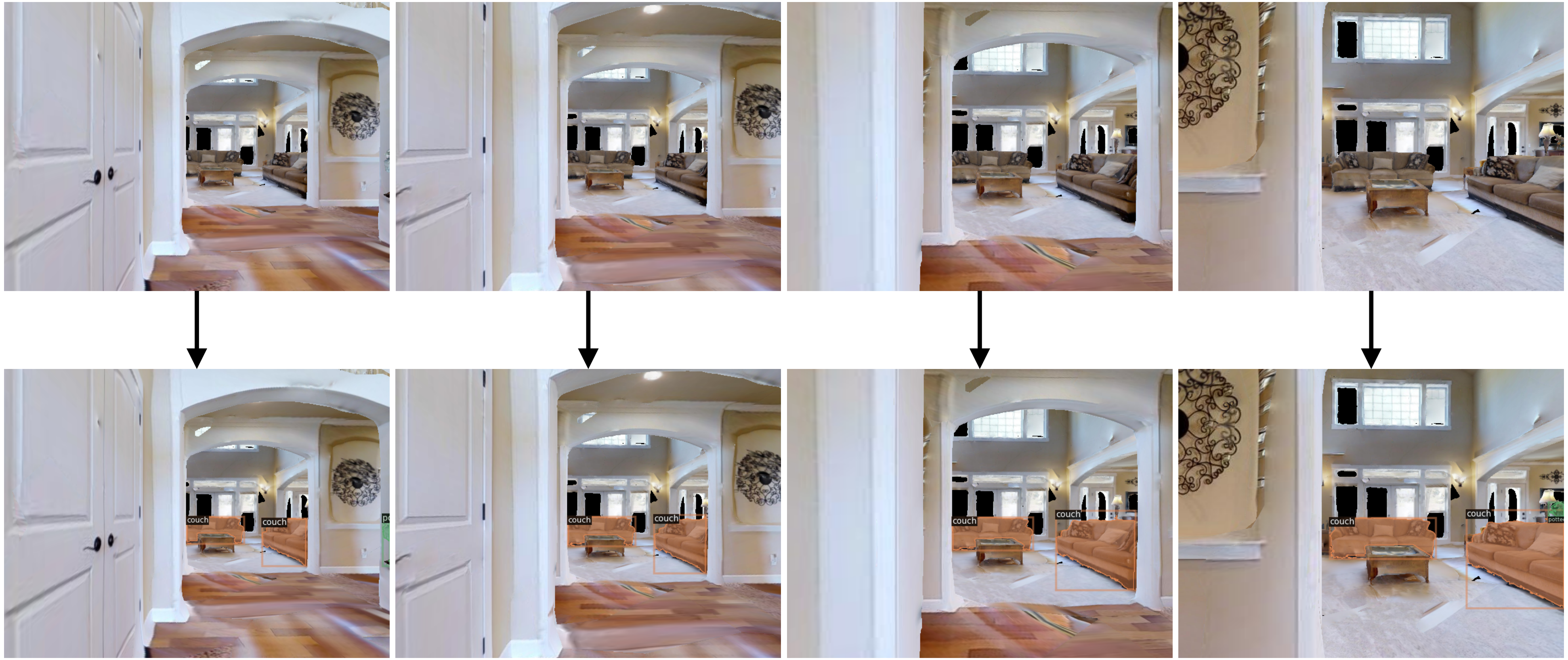
False positives

False negatives

Savva et al, Habitat: A platform for embodied AI research, ICCV 2019

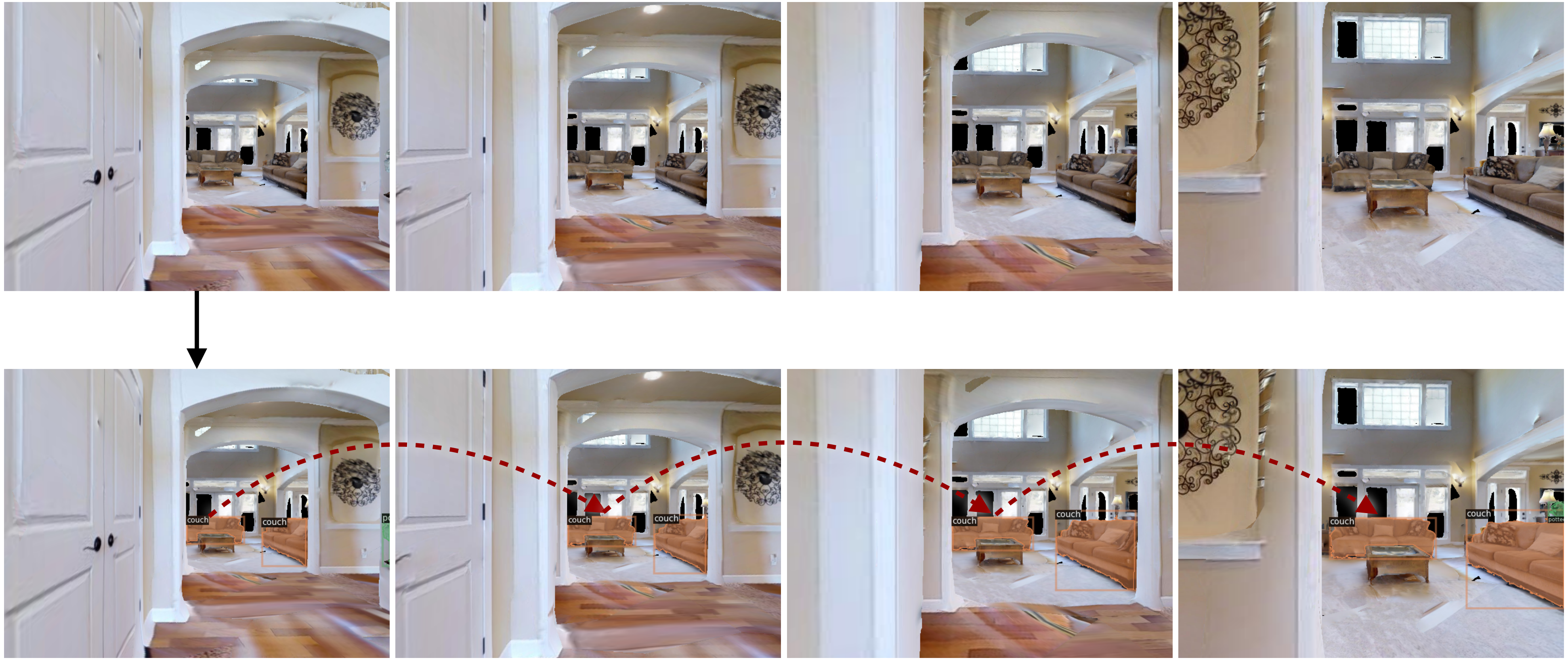
Embodied Perception

Active Embodied data

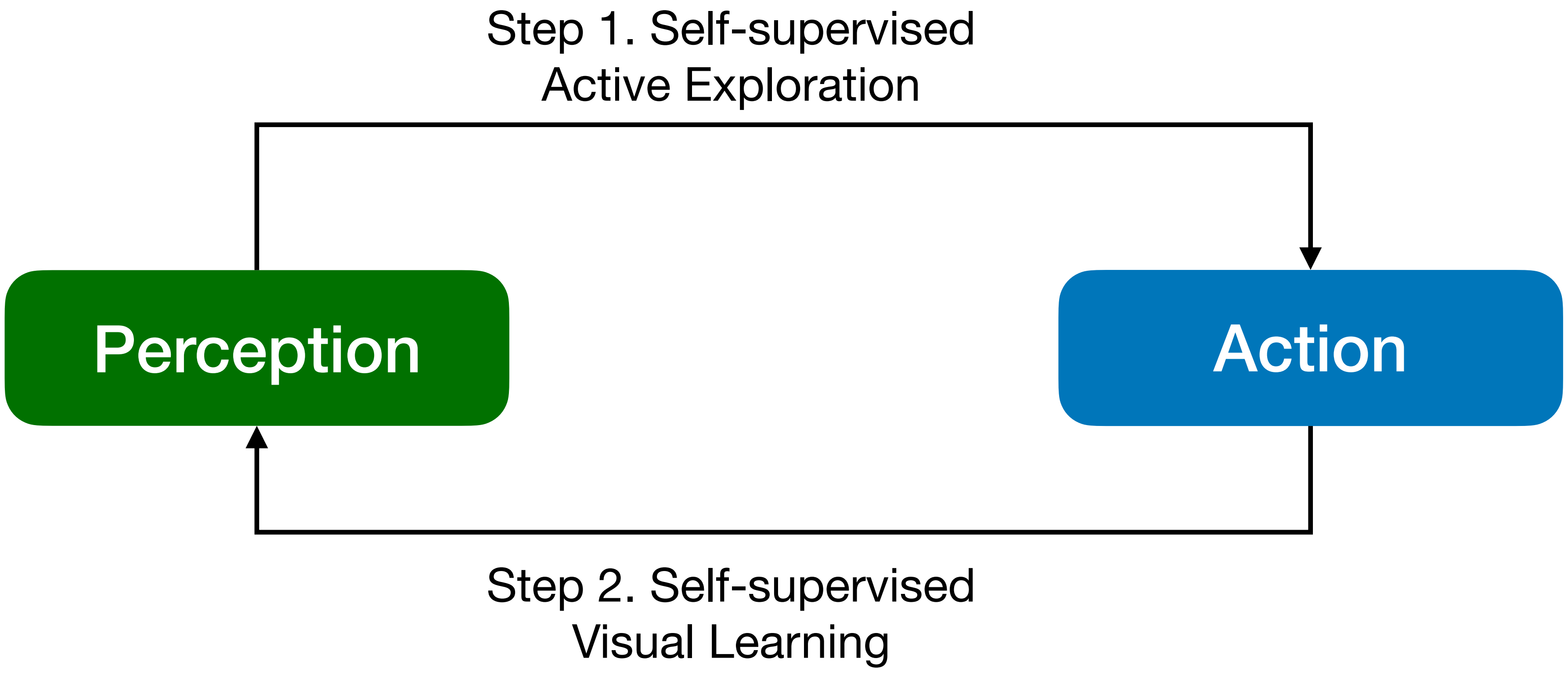


Embodied Perception

Active Embodied data

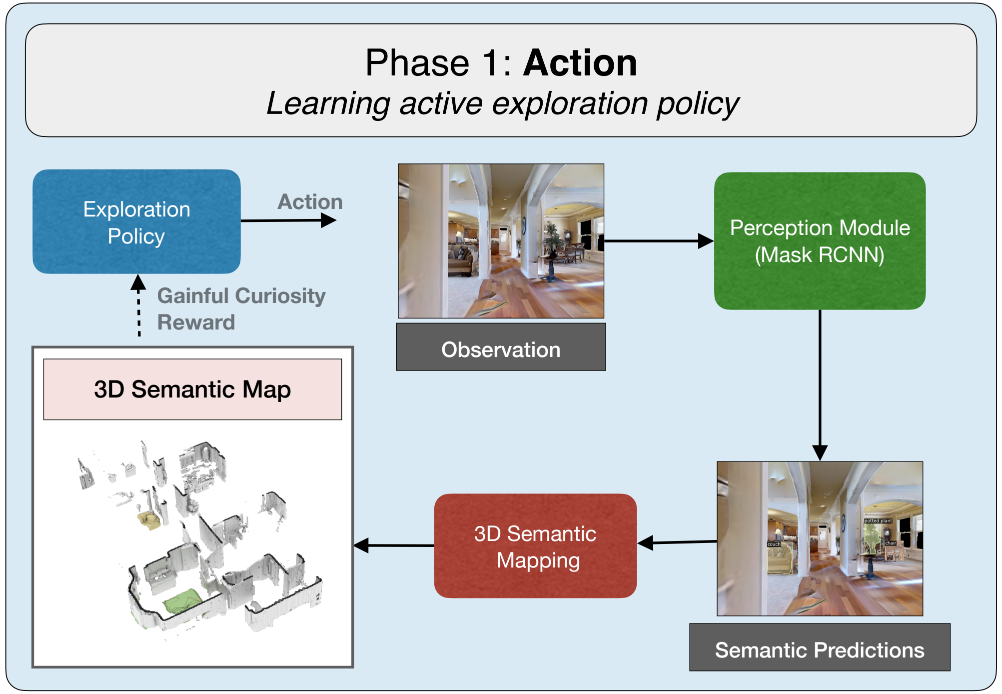


Perception-Action Loop



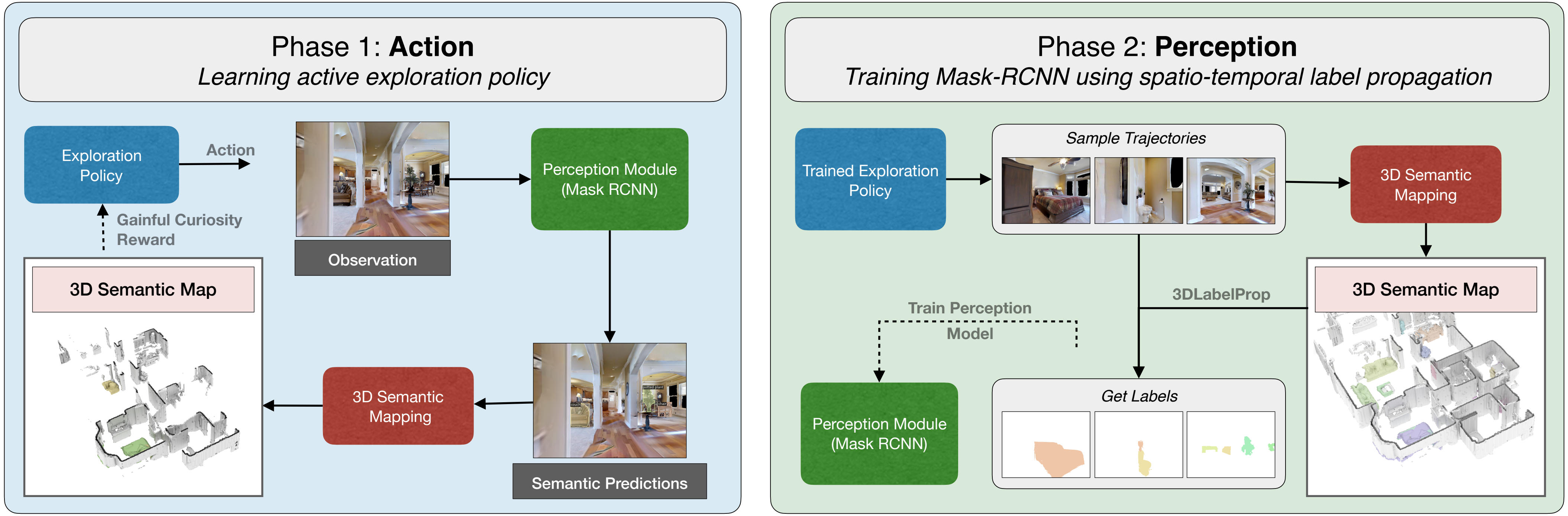
Pathak et al, Learning instance segmentation by interaction, 2018
Jang et al, Grasp2vec: Learning object representations from self-supervised grasping, 2018
Eitel et al, Self-supervised transfer learning for instance segmentation through physical interaction, 2019
Fang et al.,Move to See Better: Self-Improving Embodied Object Detection, 2021

SEAL: Self-supervised Embodied Active Learning



Chaplot, Dalal, Gupta, Malik, Salakhutdinov et al, . SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency NeurIPS-21

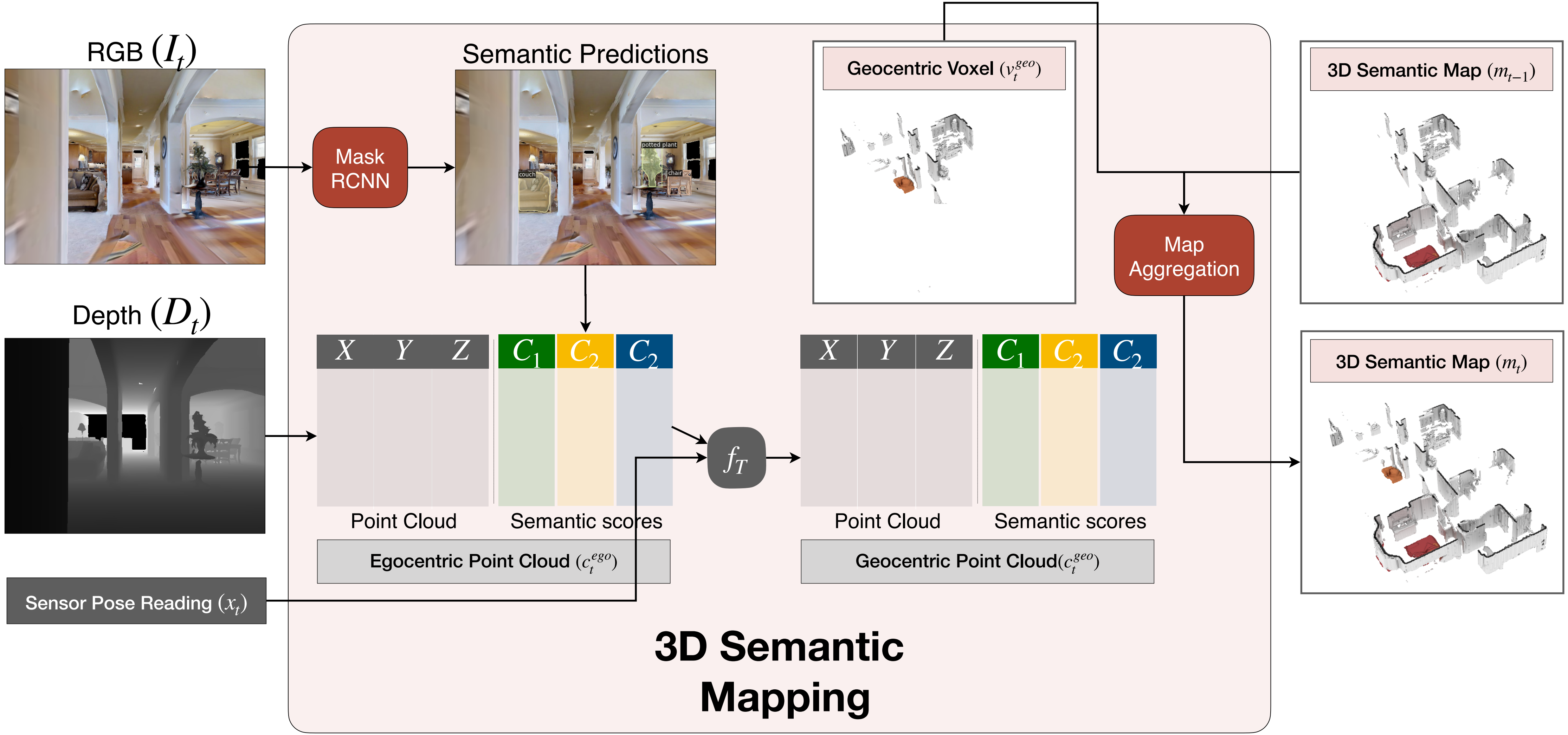
SEAL: Self-supervised Embodied Active Learning



Both phases do not require any additional labelled data

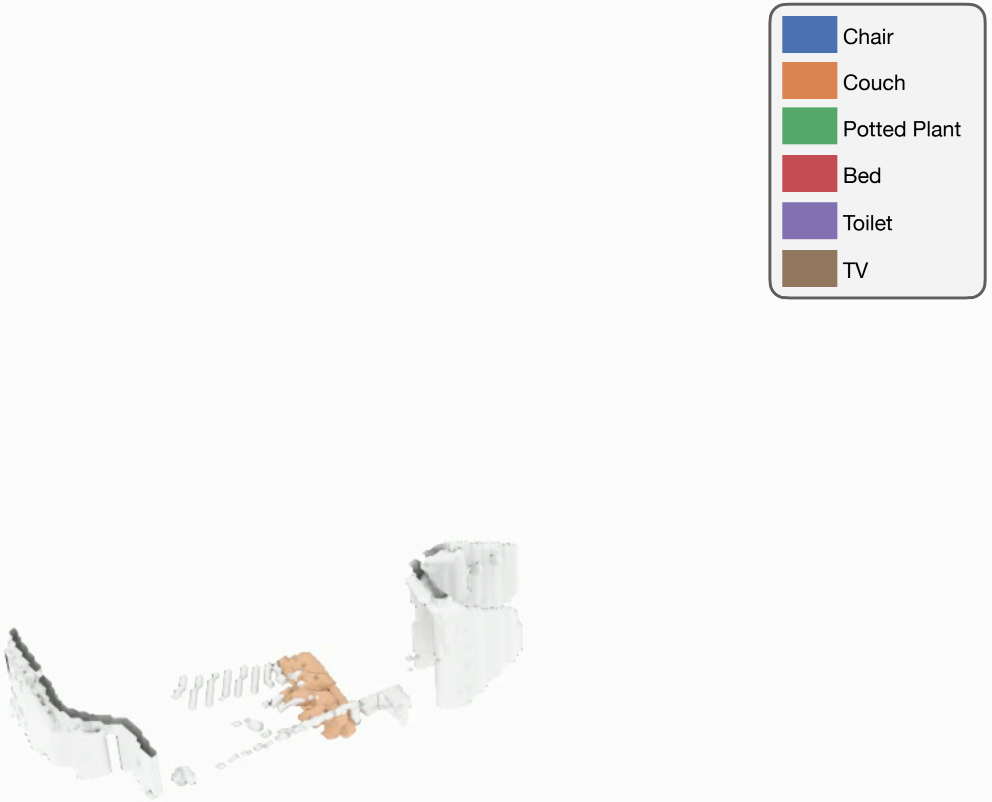
Chaplot, Dalal, Gupta, Malik, Salakhutdinov et al, . SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency NeurIPS-21

3D Semantic Mapping



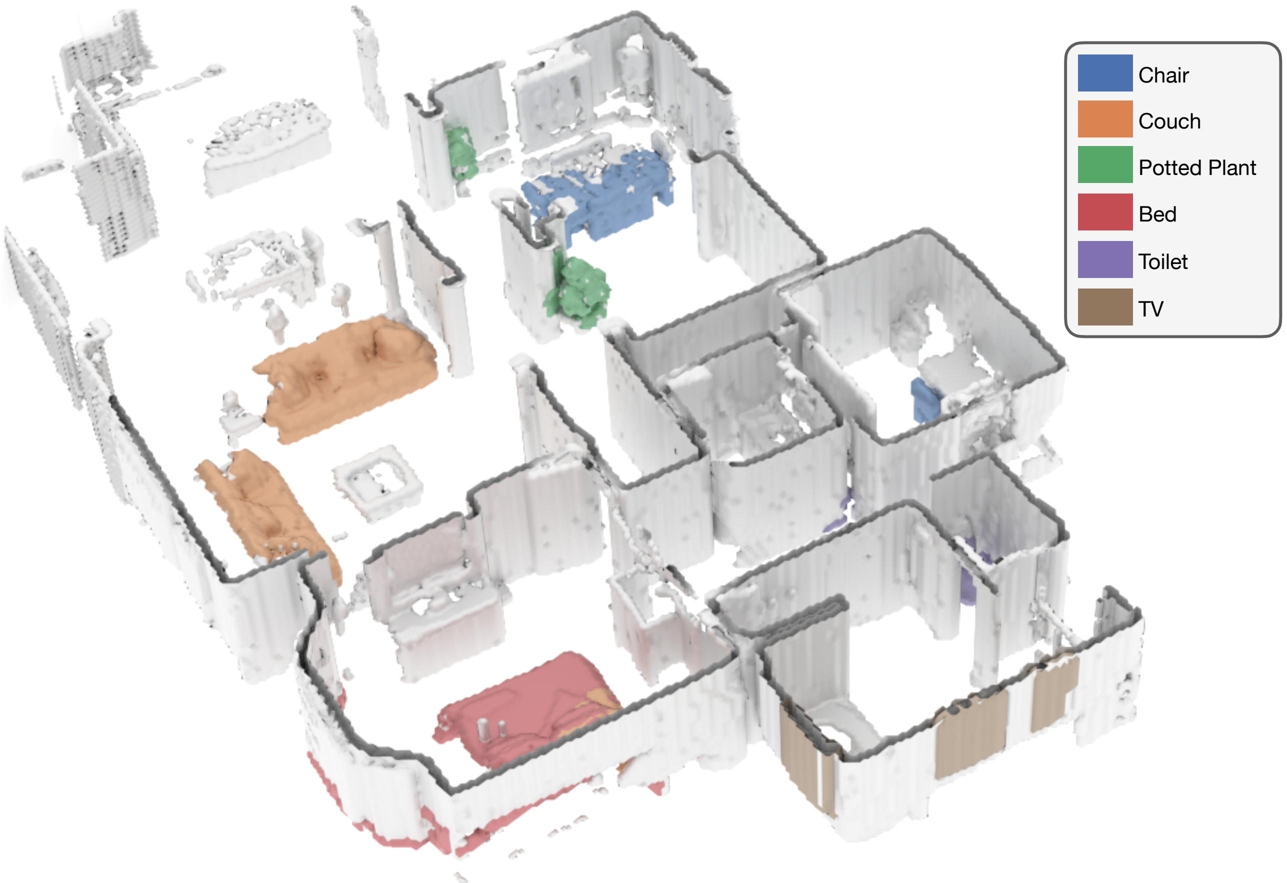
3D Semantic Mapping

3D Semantic Map $M = K \times L \times W \times H$

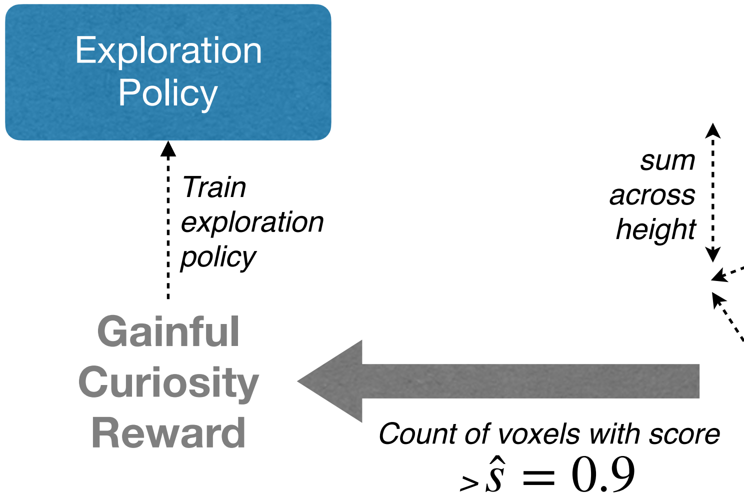


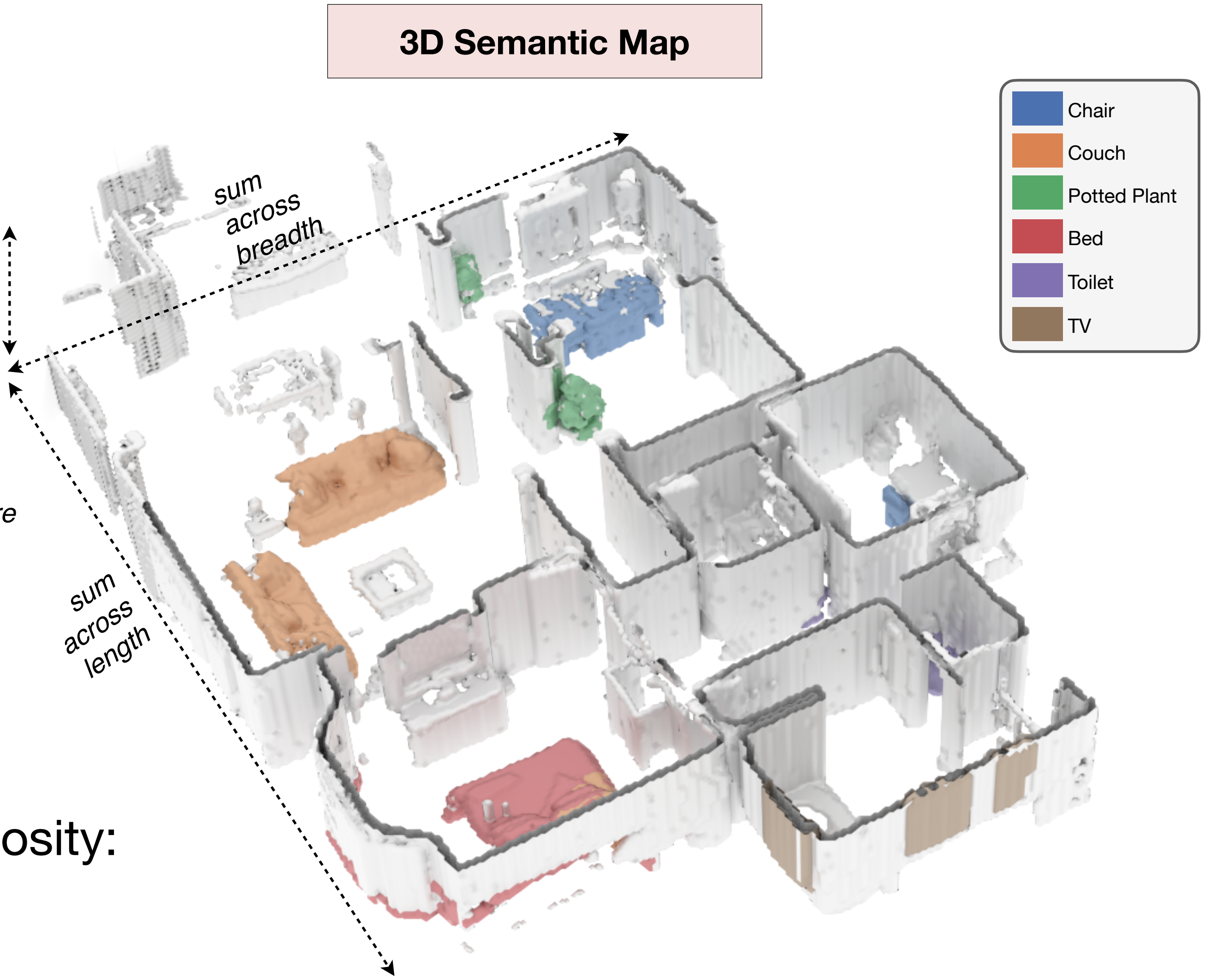
3D Semantic Mapping

3D Semantic Map $M = K \times L \times W \times H$



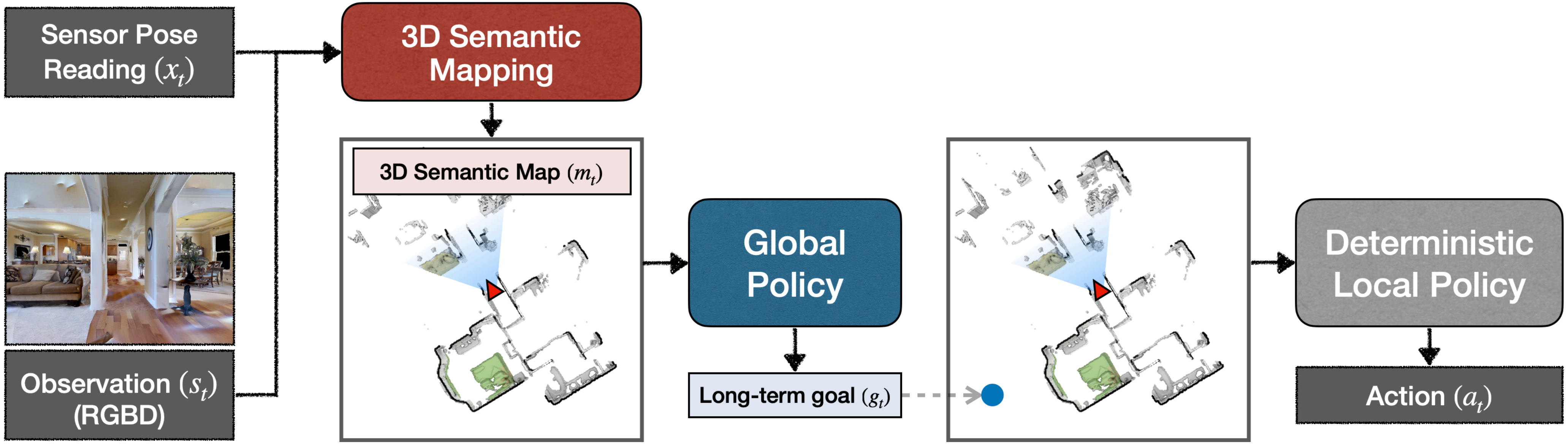
Gainful Curiosity





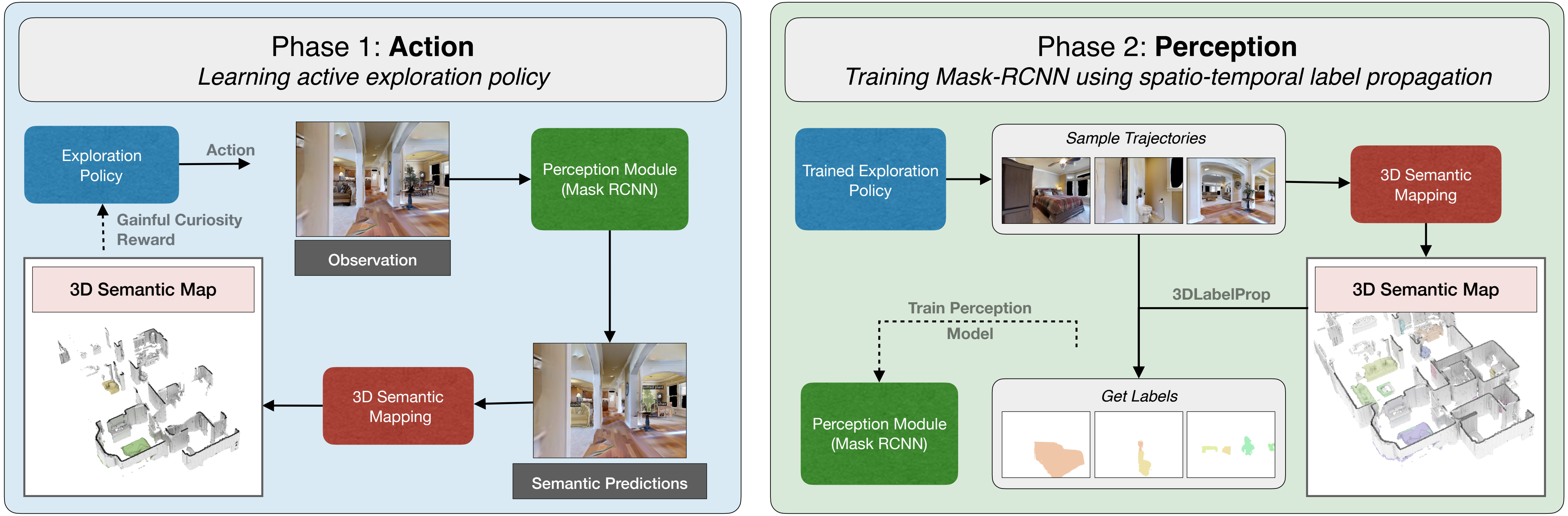
- Trained to maximise Gainful Curiosity: gaining definitive knowledge

Policy Learning



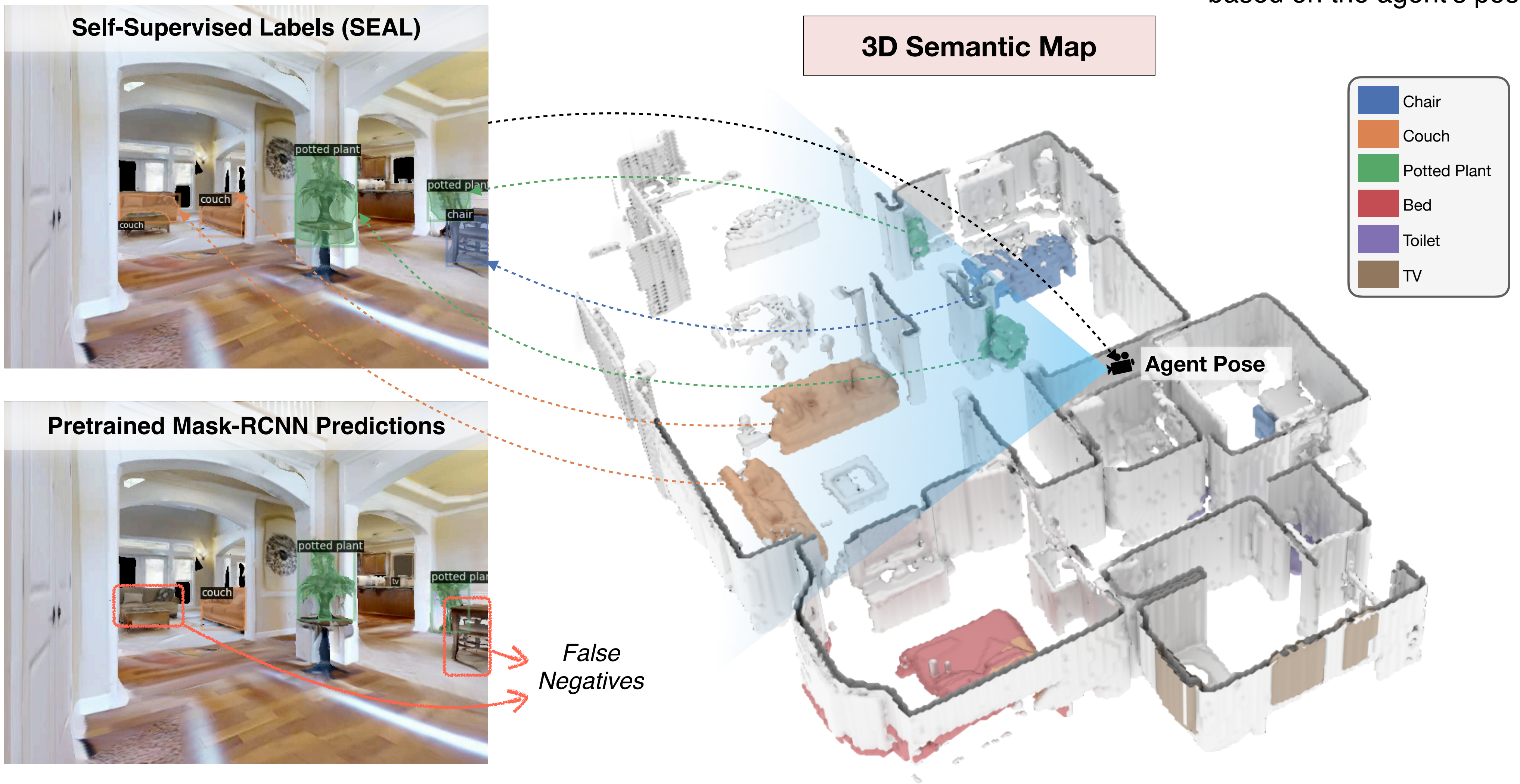
- Global Policy: samples a goal every 25 local steps
- Action Space: move forward (25cm), turn left or right (30 degrees)

SEAL: Self-supervised Embodied Active Learning

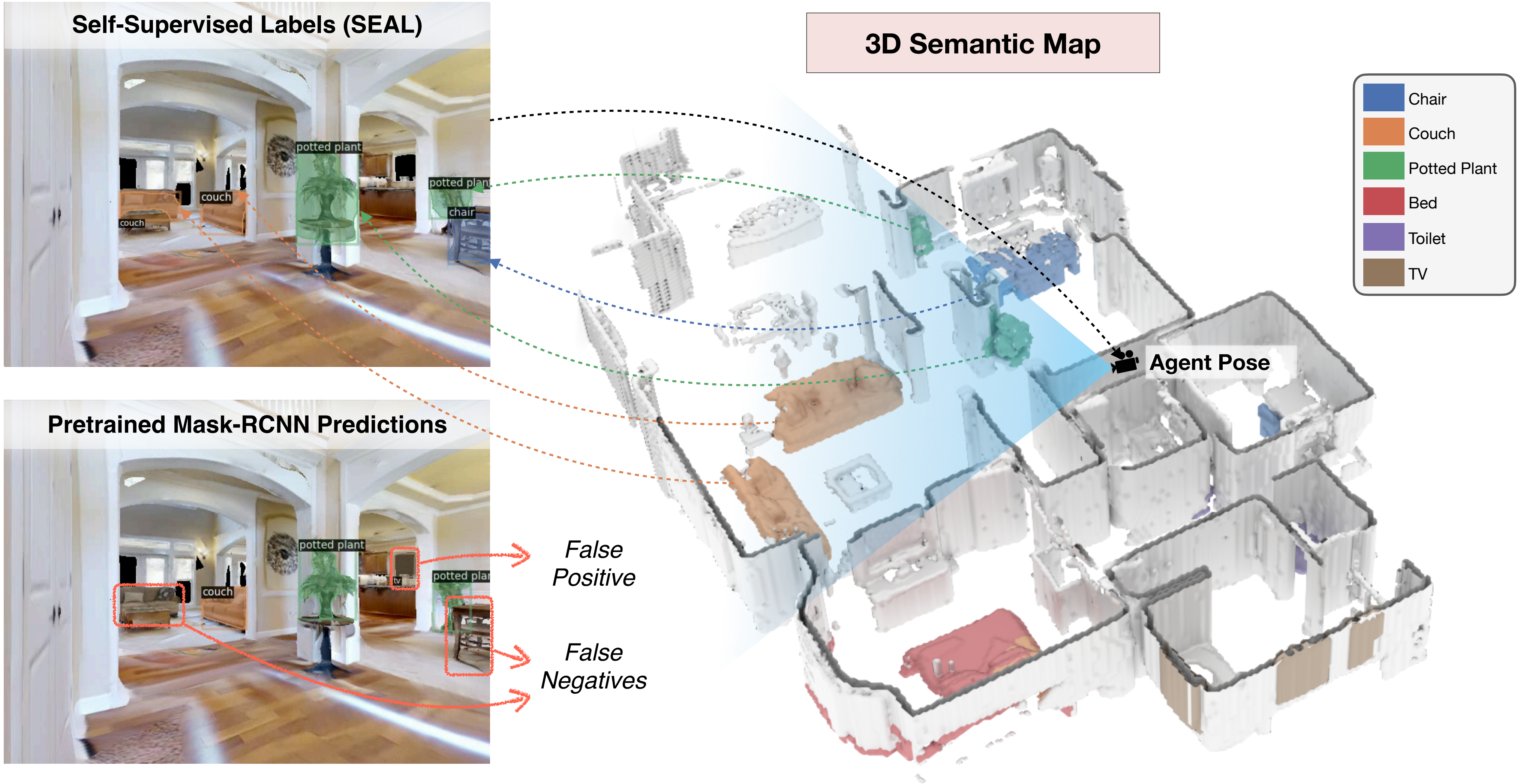


3D Label Propagation

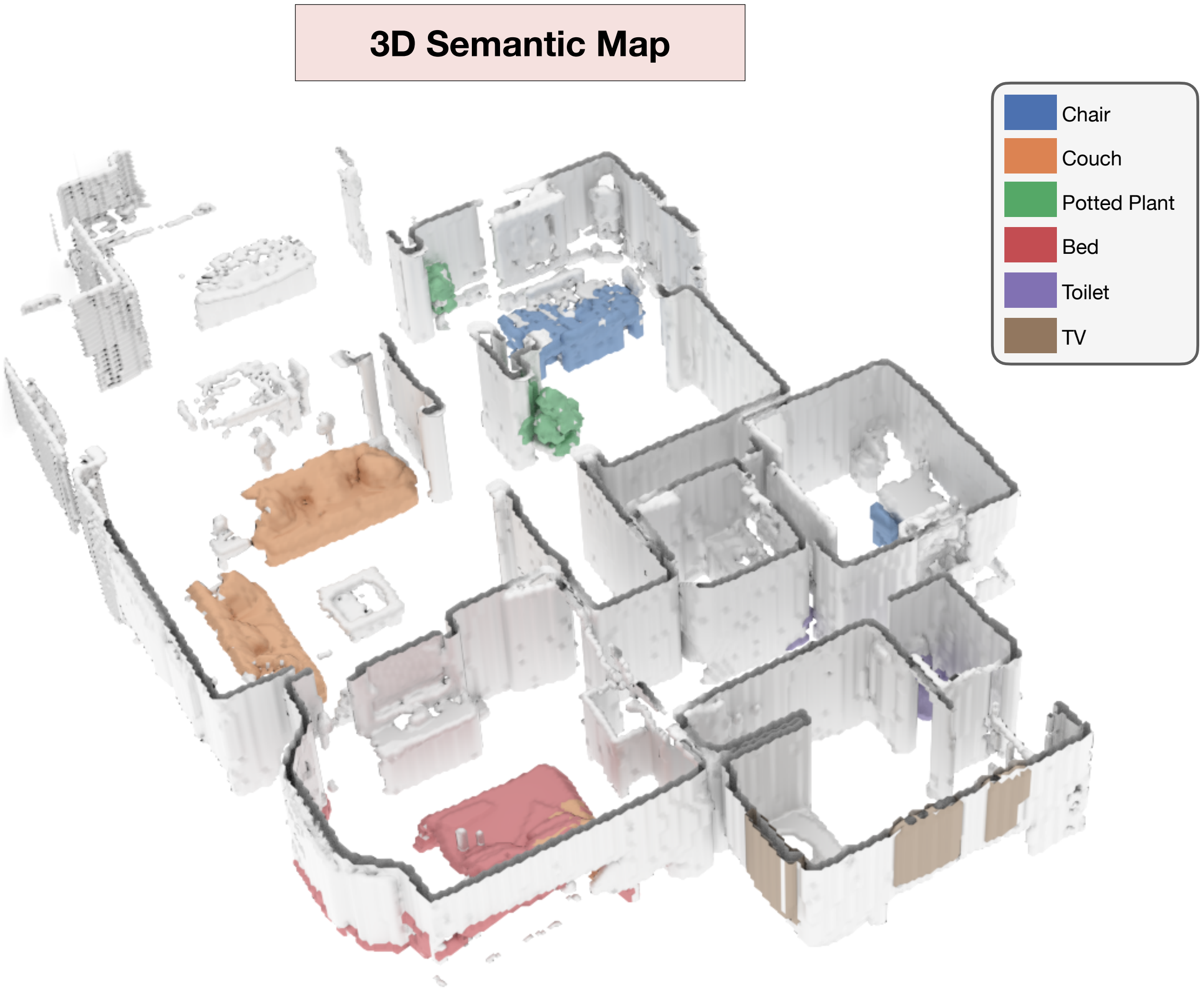
Instance label for each pixel is obtained using ray tracing based on the agent's pose



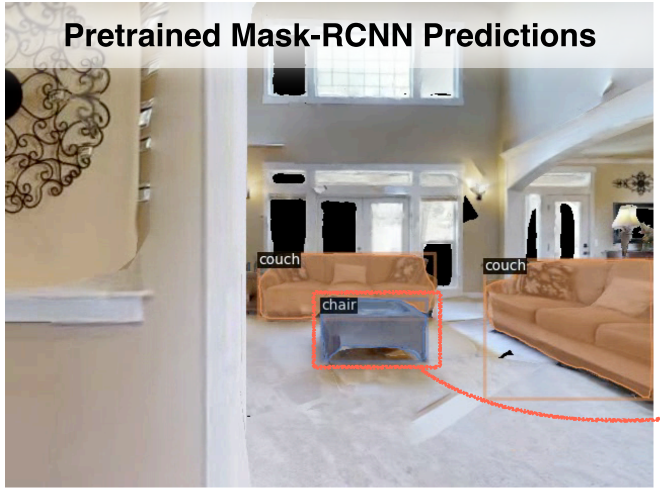
3D Label Propagation



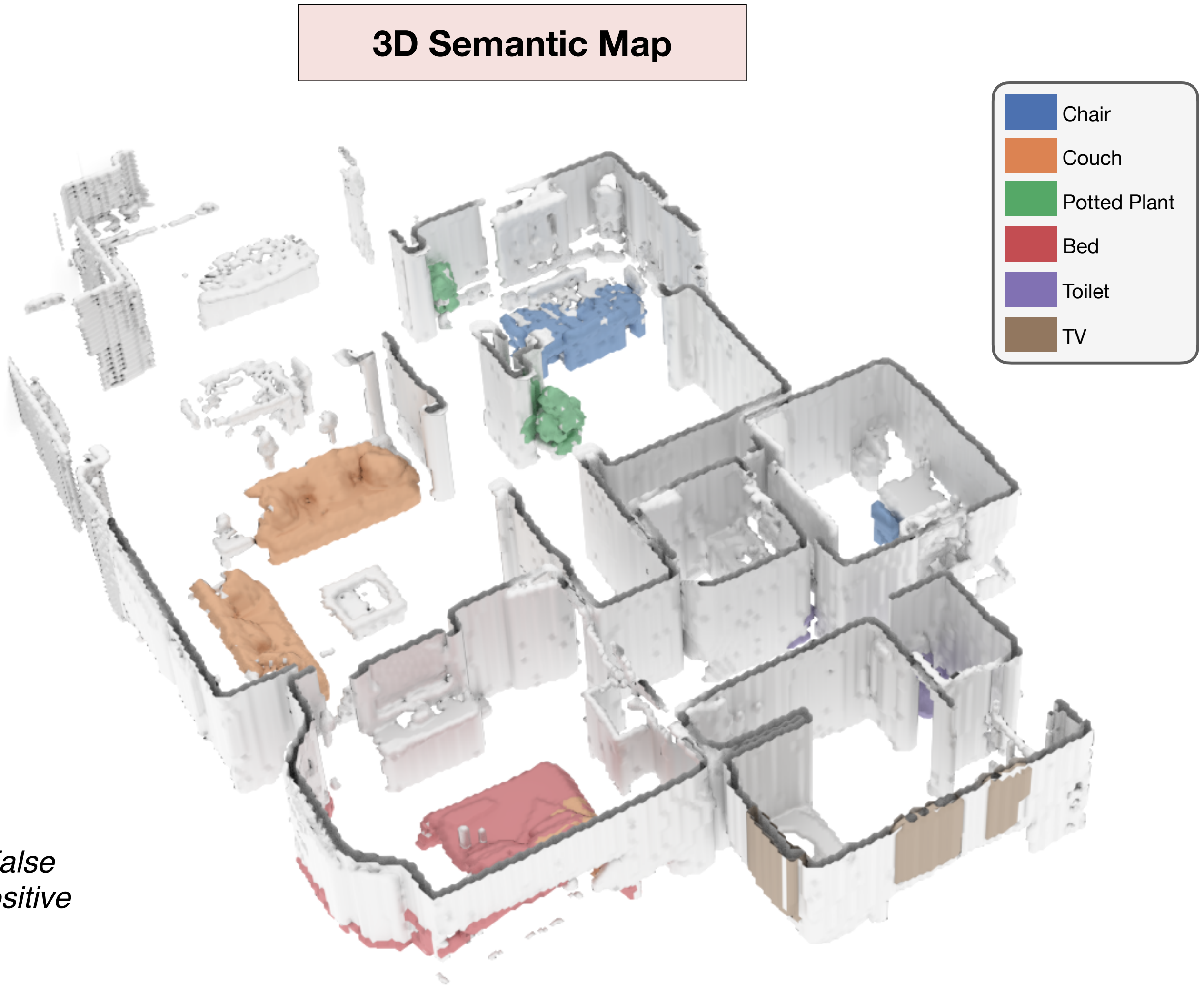
3D Label Propagation



3D Label Propagation

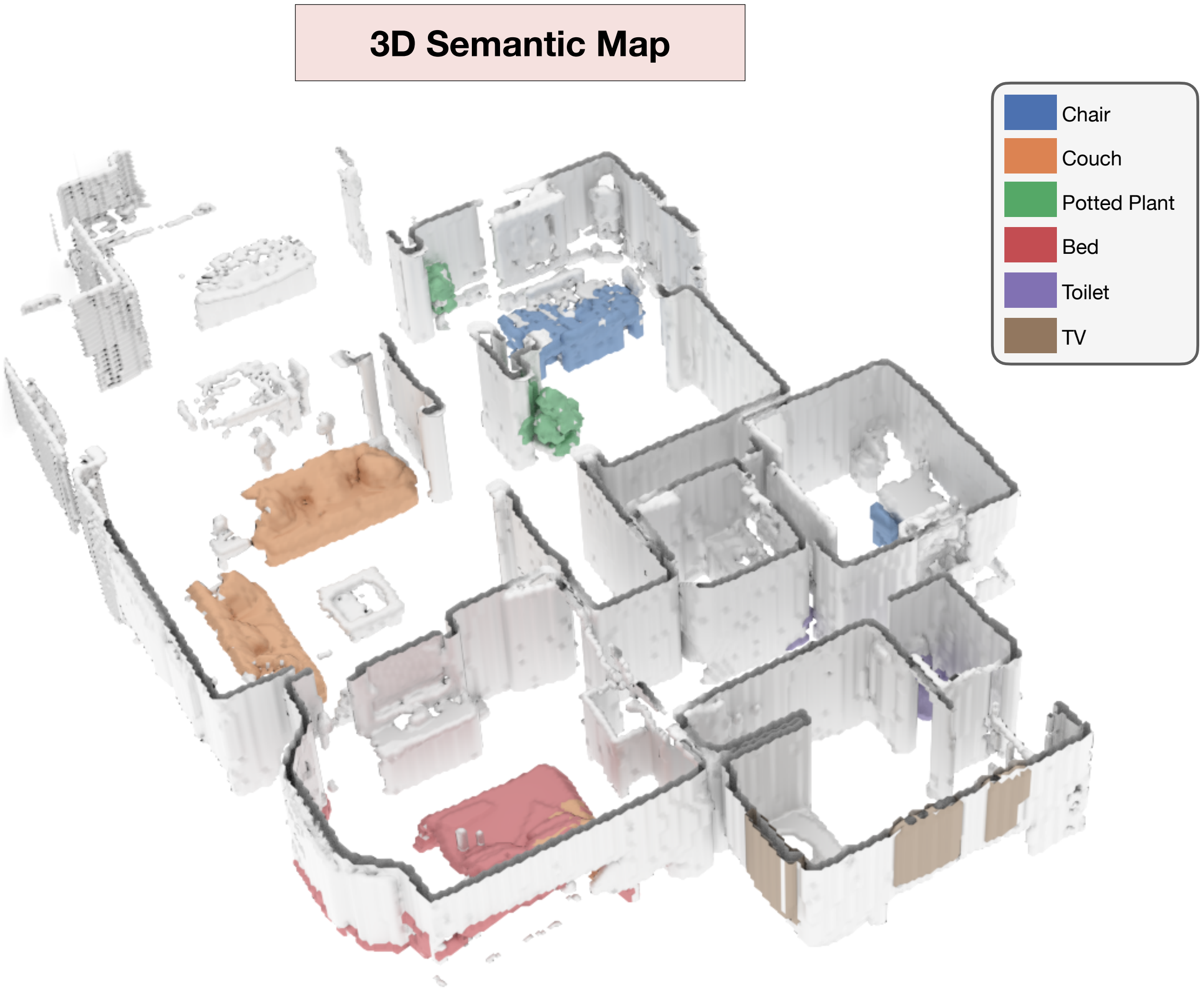


False Positive

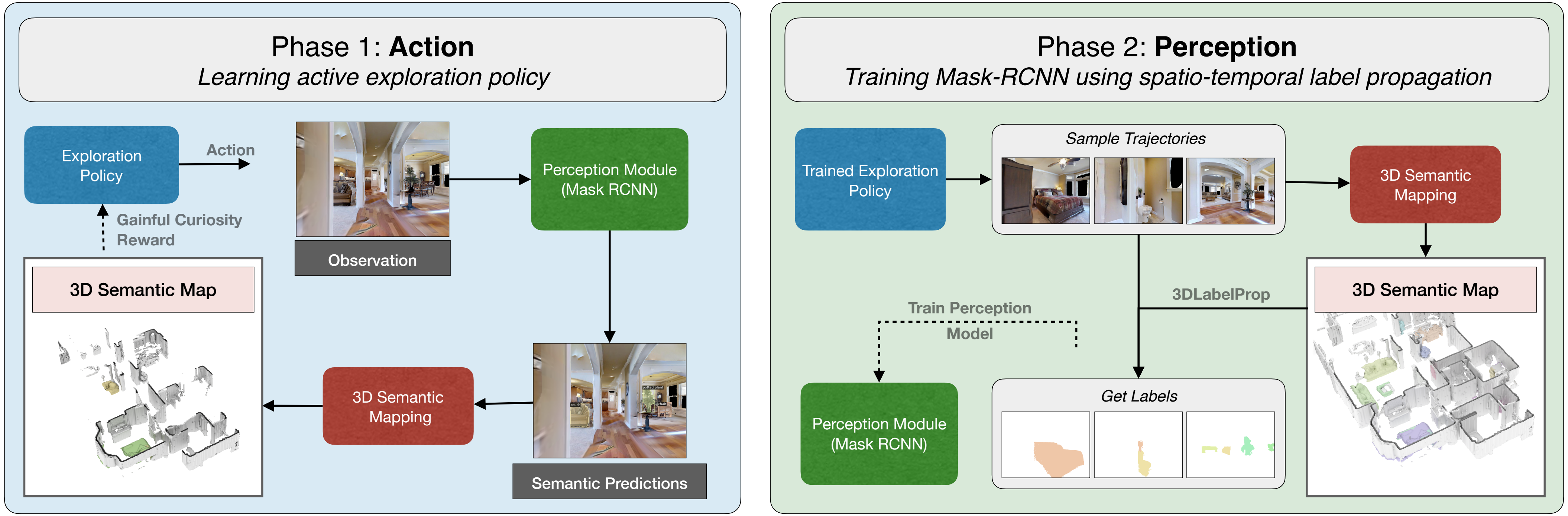


3D Label Propagation

Train Perception Model



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

Armeni et al , 3d scene graph: A structure for unified semantics, 3d space, and camera, ICCV 2019

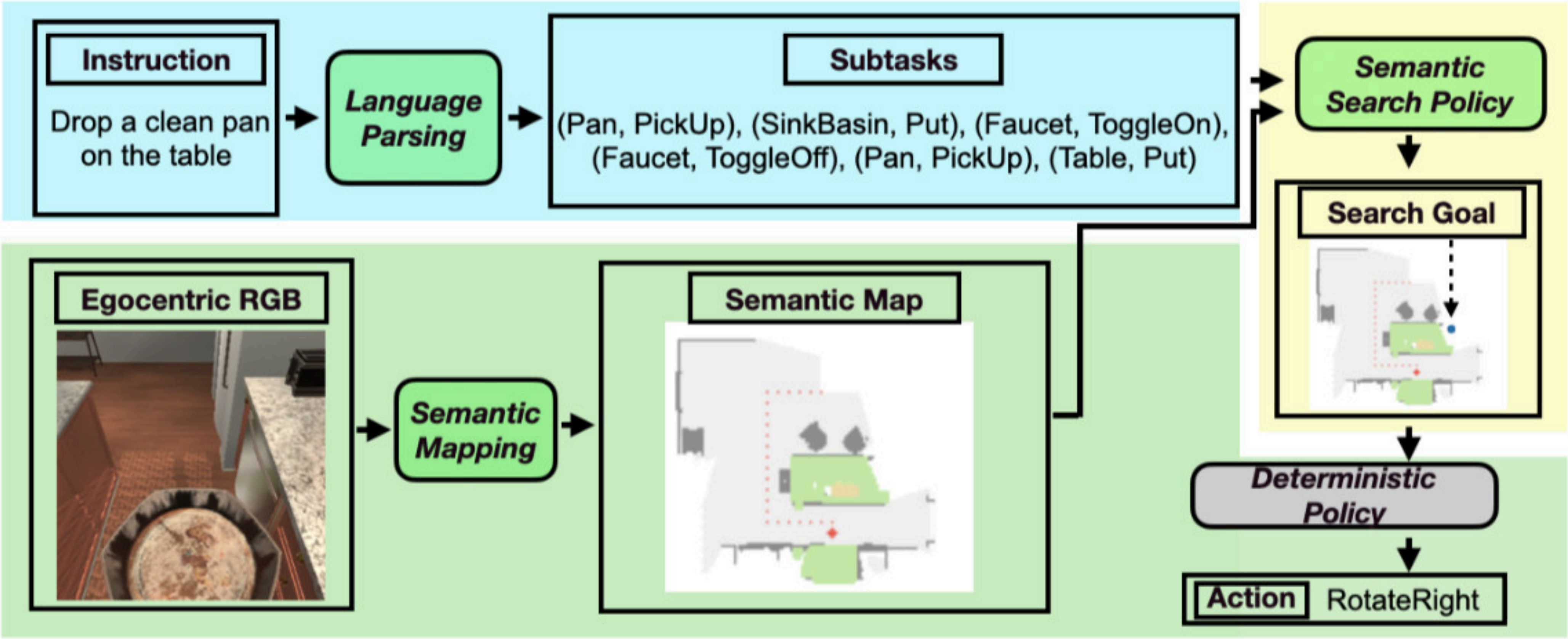
Results

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

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks

FILM: Following Instructions in Language with Modular Methods



FILM: Following Instructions in Language with Modular Methods
So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, Ruslan Salakhutdinov, ICLR 2022

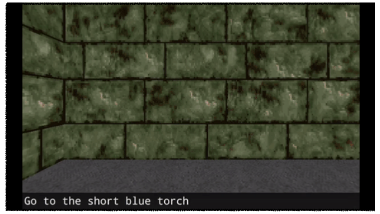
FILM: Following Instructions in Language with Modular Methods

Simulation to Real

Games

ViZDoom

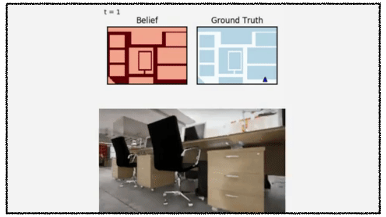
[CL *AAAI-17*]



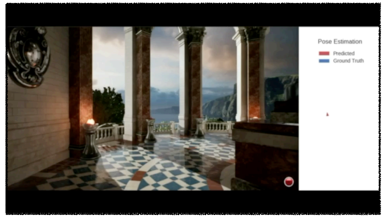
[CMPRS *AAAI-18*]

Photorealistic simulation

Unreal



[CPS *ICLR-18*]

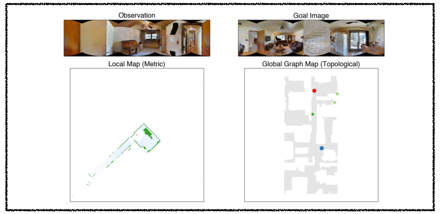


[PCZS *CVPR-18 (w)*]

Reconstructed simulation

Habitat (Gibson, MP3D)

[CGSGG *ICLR-20*]



[CSGG *CVPR-20*]

Physical Domain Gap

Visual Domain Gap

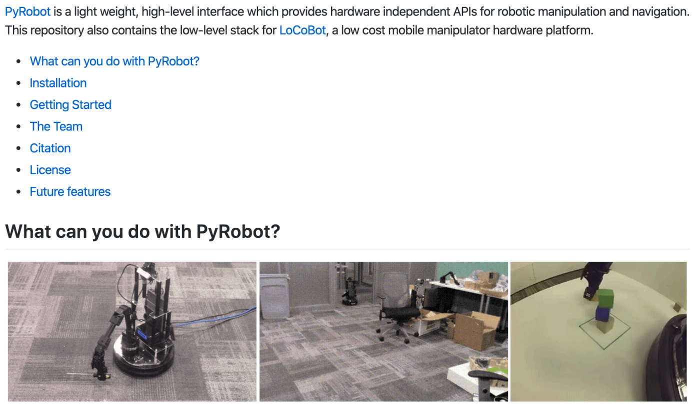
Real-world

?

Simulation to Real

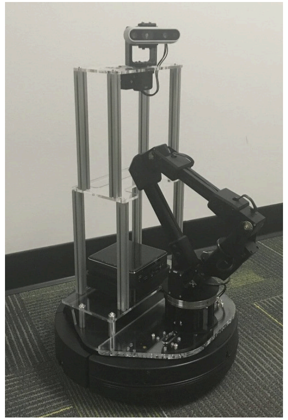
- Physical Domain Gap
  - Actuation noise models
  - Sensor noise models
- Visual Domain Gap
  - Image Translation
  - Policy-based

PyRobot



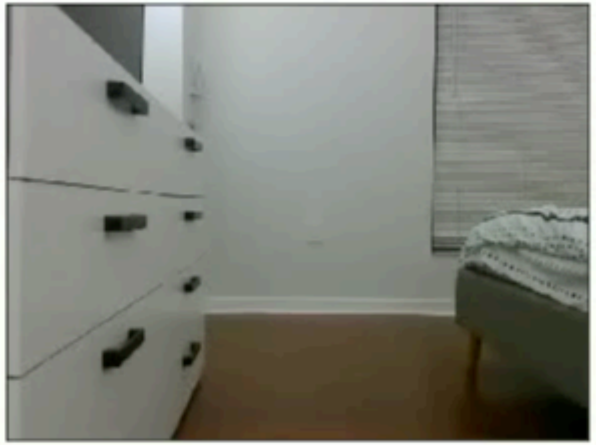
pyrobot.org

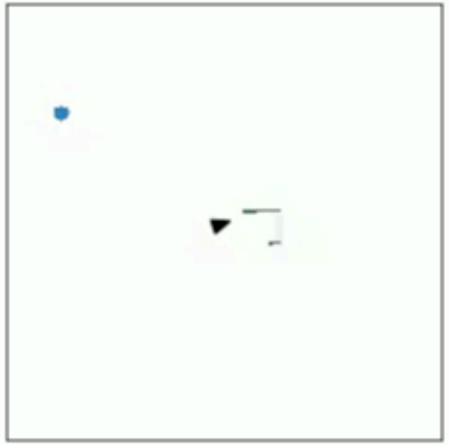
LoCoBot



locobot.org

Simulation to Real





Building Intelligent Agents

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