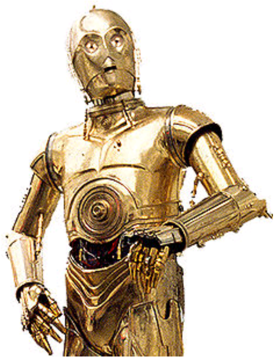
Embodied AI: Language and Perception

Russ Salakhutdinov

Machine Learning Department
Carnegie Mellon University

Learning Behaviors

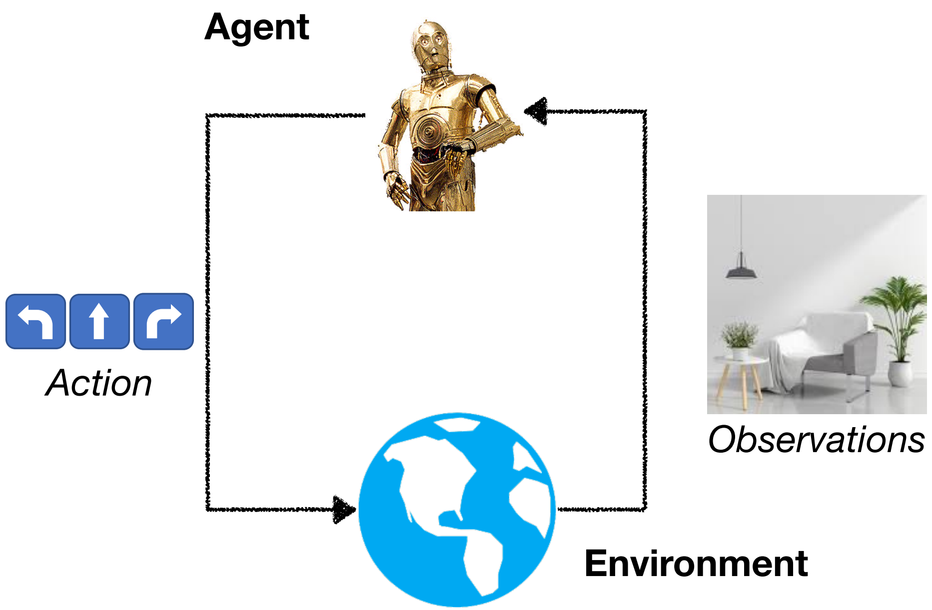
Observation $\longrightarrow$ o_t



Action $\longrightarrow$ a_t

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

Physical Intelligence

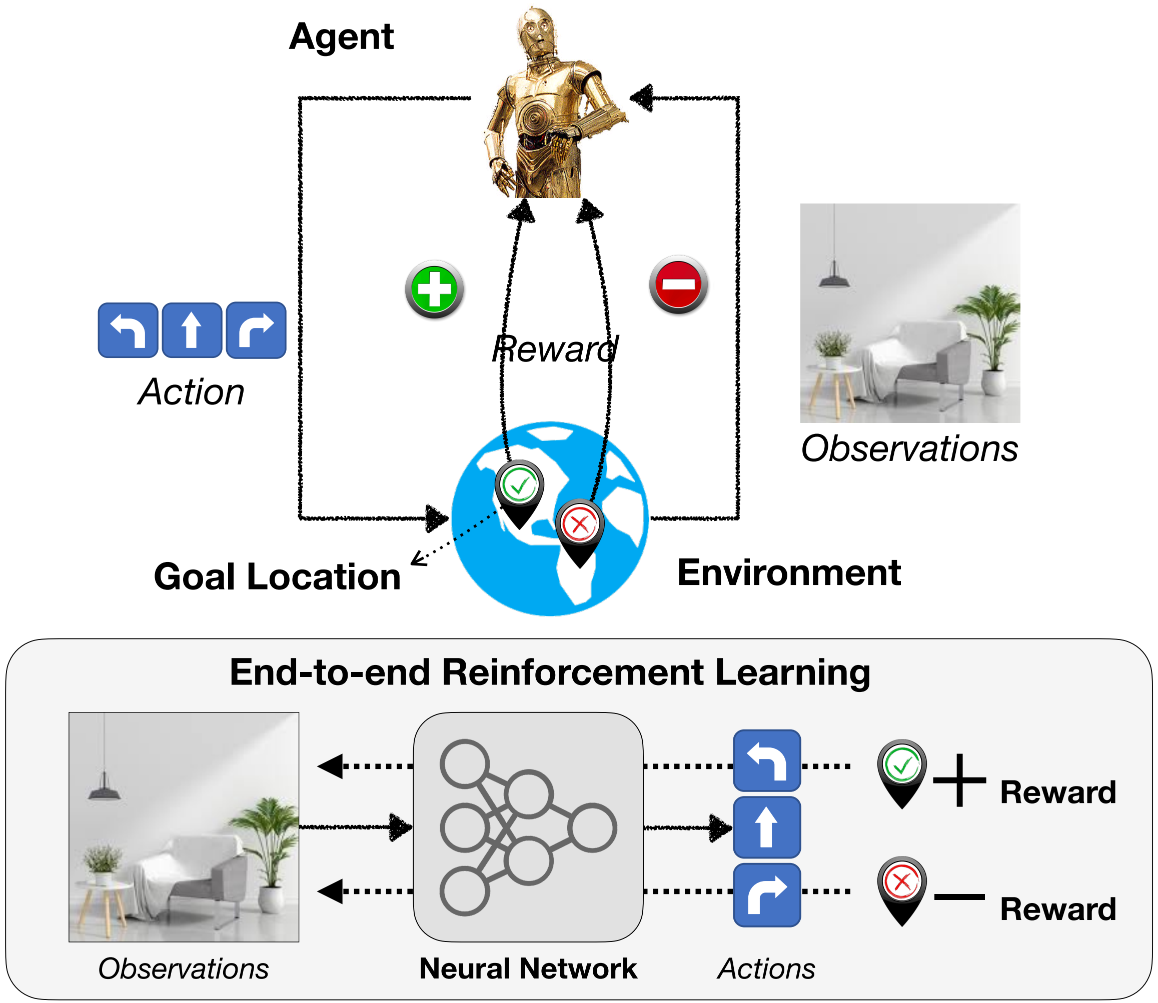


Agent needs to move in the world physically.

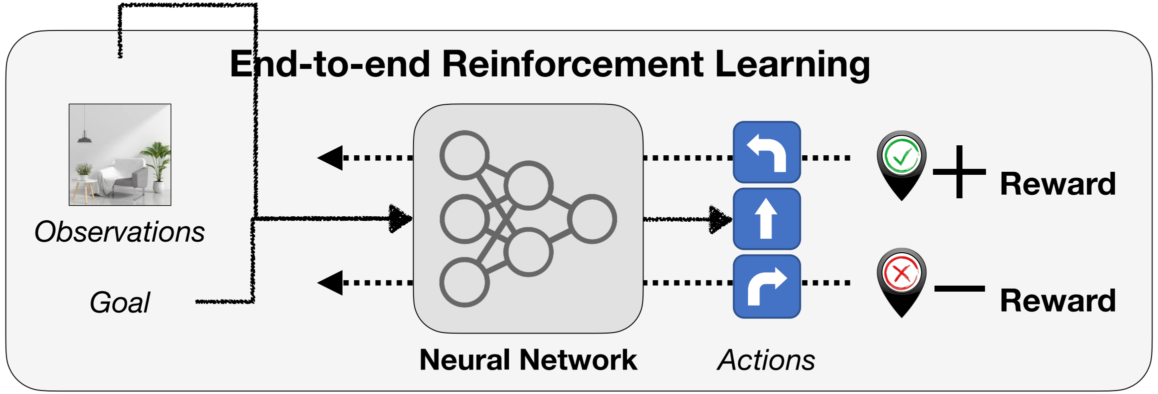
Current actions affect future observations.

Require Spatial and Semantic Understanding.

Navigation


Agent
Reward
Action
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

Navigation Tasks

Point Goal

(x, y) *coordinate*

Image Goal

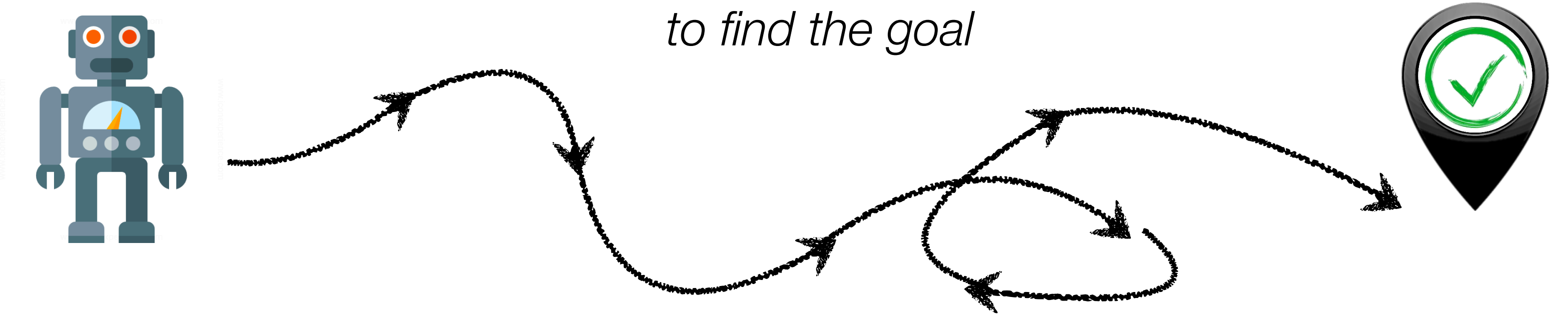
Object Goal

Chair
TV
Sofa

Language Goal

Blue Chair
Largest TV
White Sofa

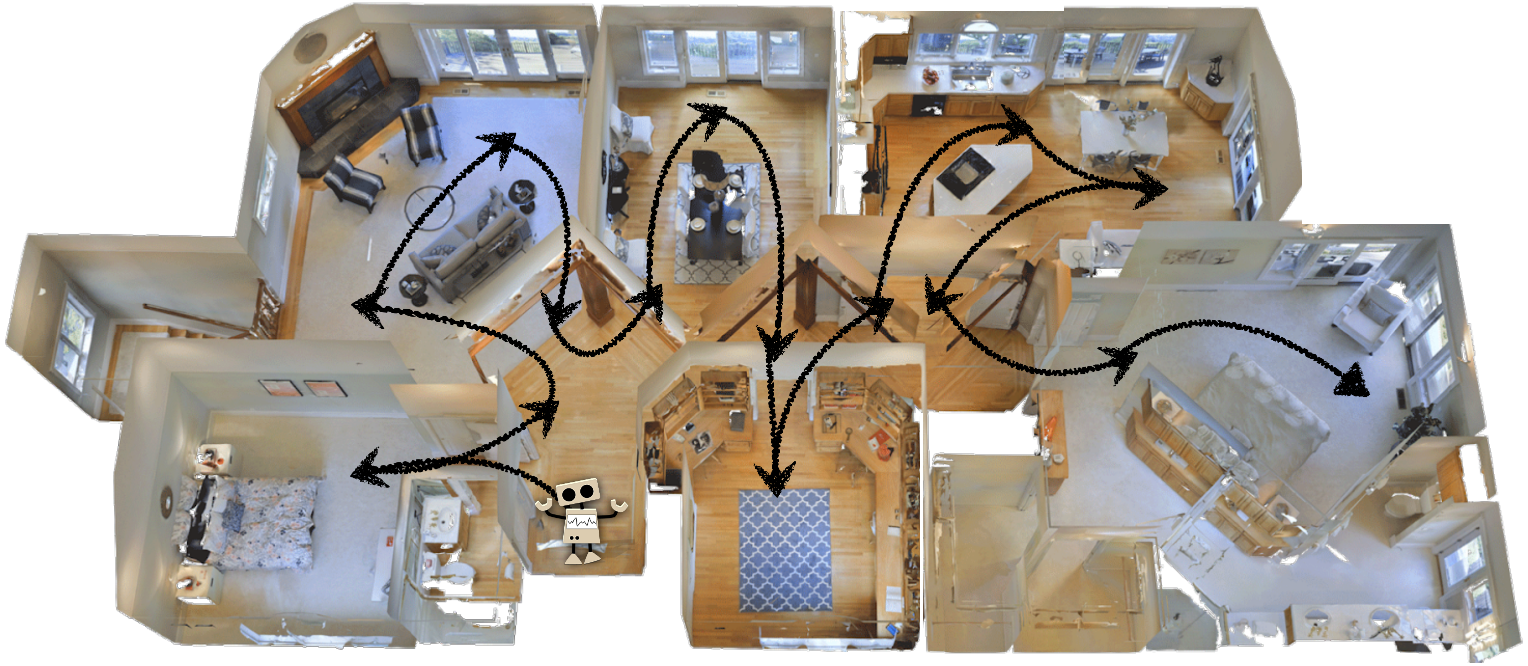
Require exploring the environment to find the goal



Real World: Object Goal Navigation

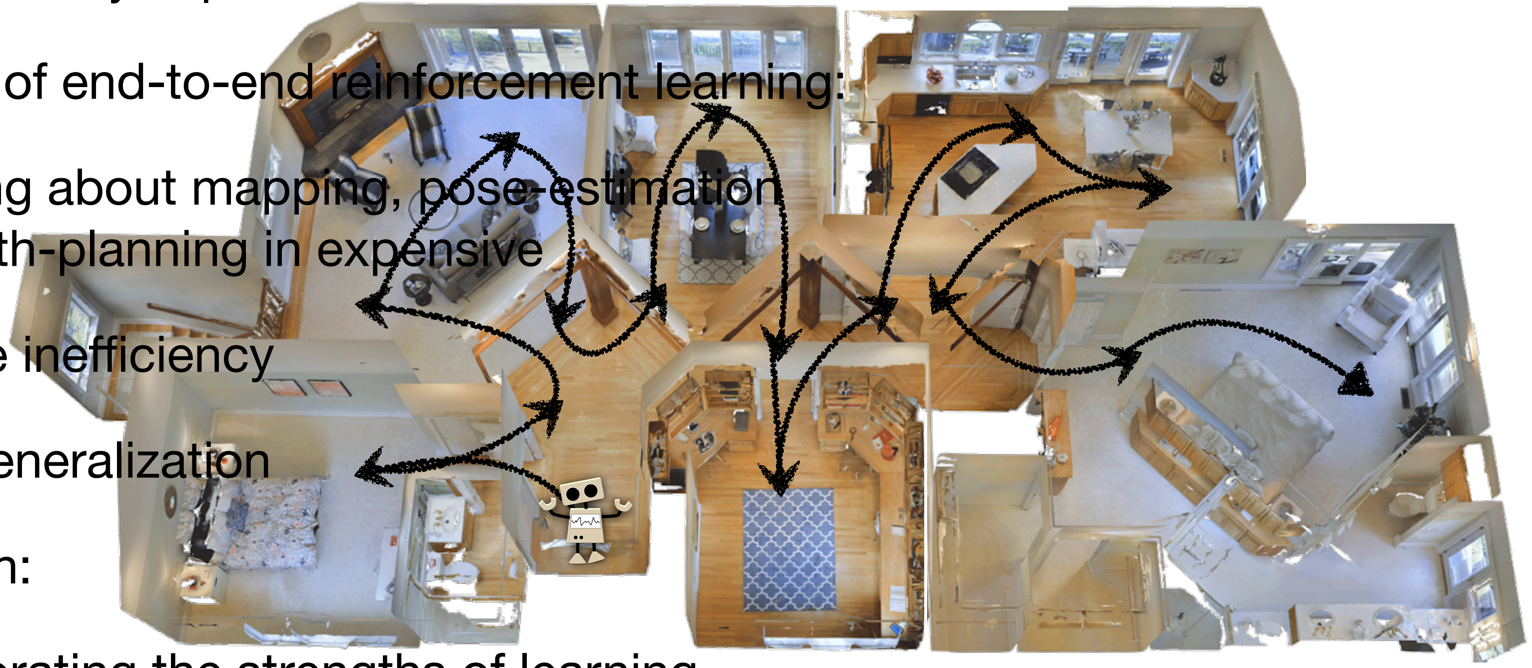
See video at: https://devendrachaplot.github.io/projects/semantic-exploration

Exploration

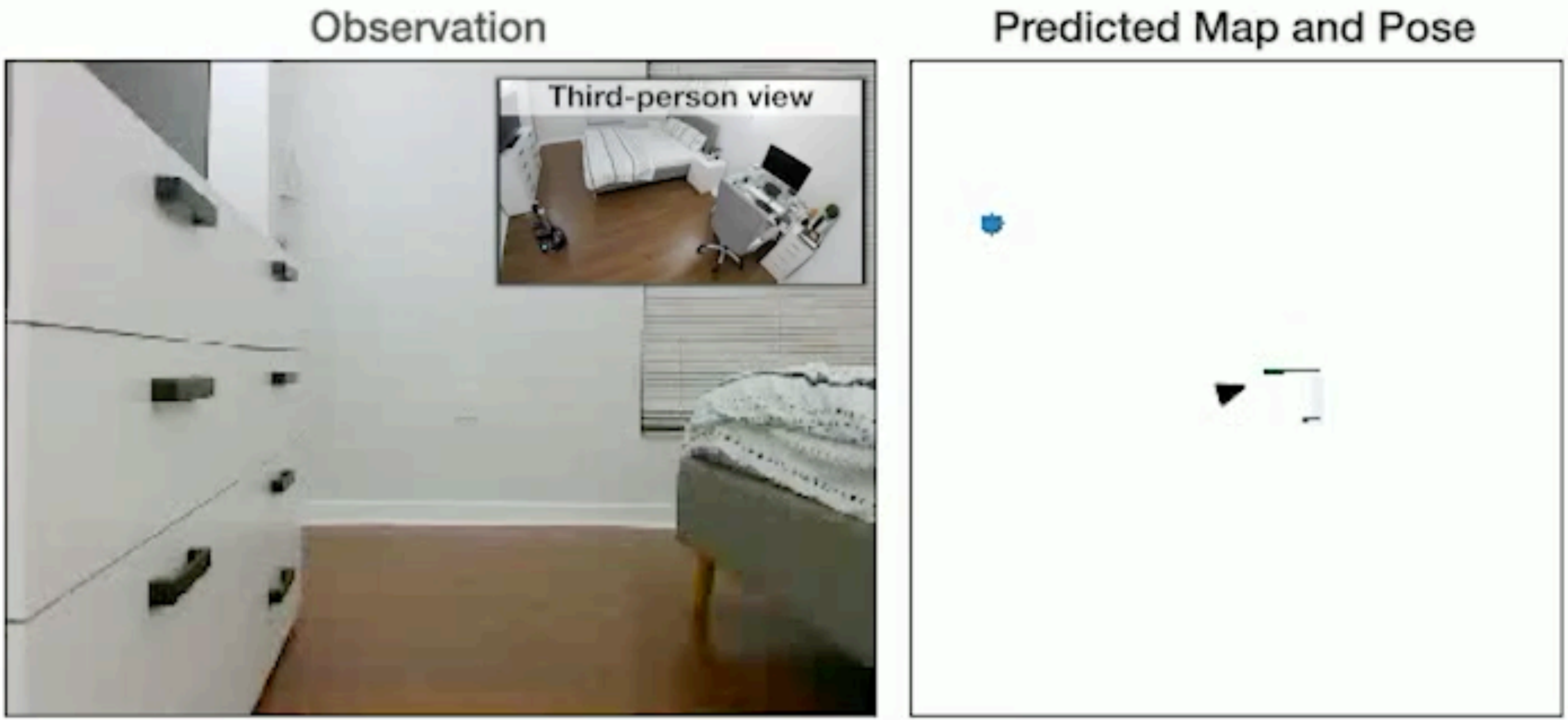


Exploration

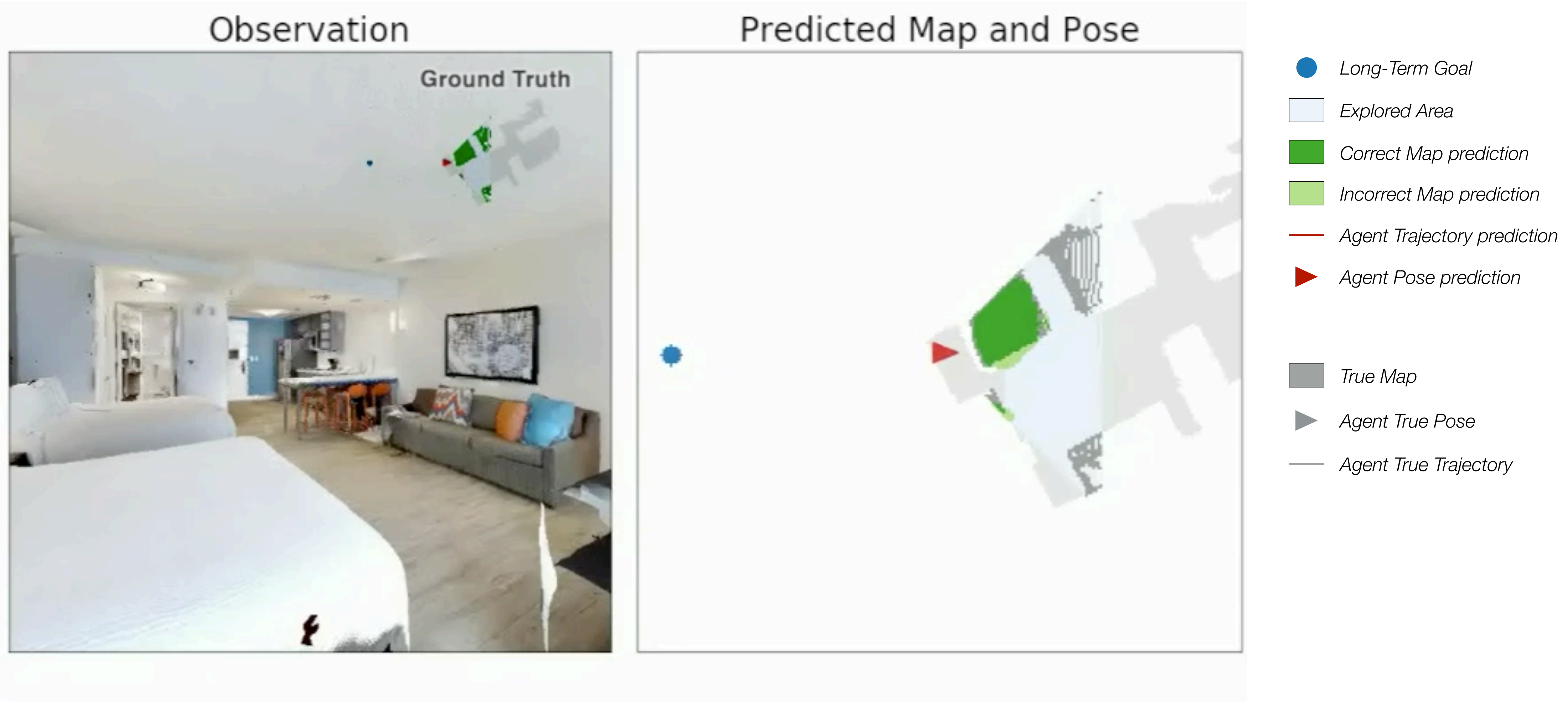
- How to efficiently explore an unseen environment?
- Limitations of end-to-end reinforcement learning:
  - Learning about mapping, pose-estimation and path-planning in expensive
  - Sample inefficiency
  - Poor generalization
- Our solution:
  - Incorporating the strengths of learning
  - Modular and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



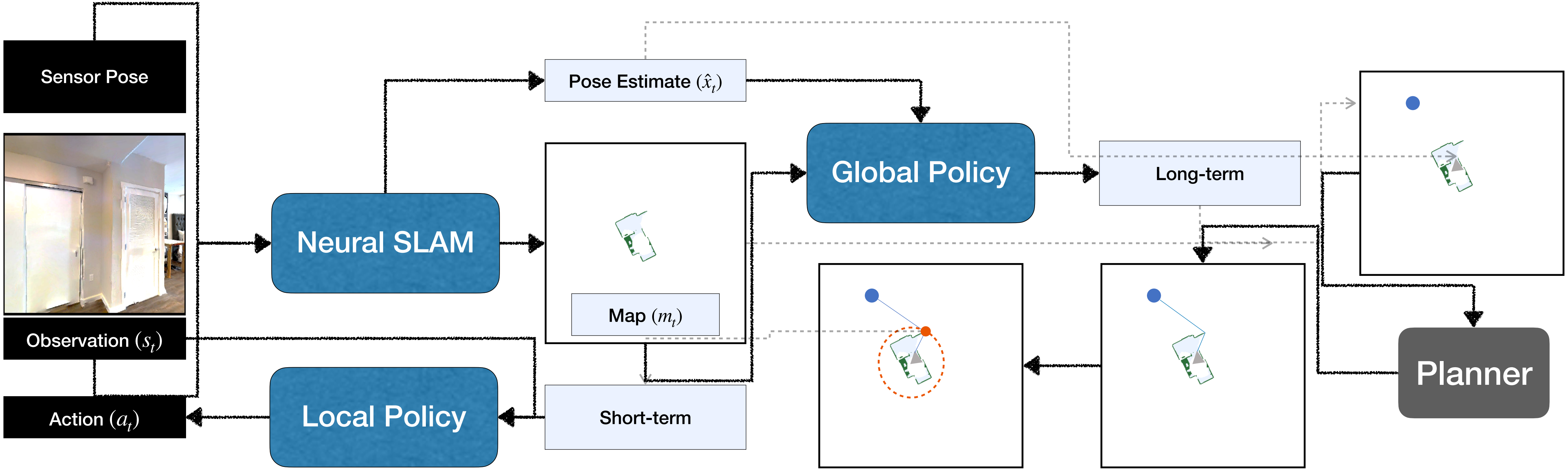
Preview: Visual Navigation in the Real World



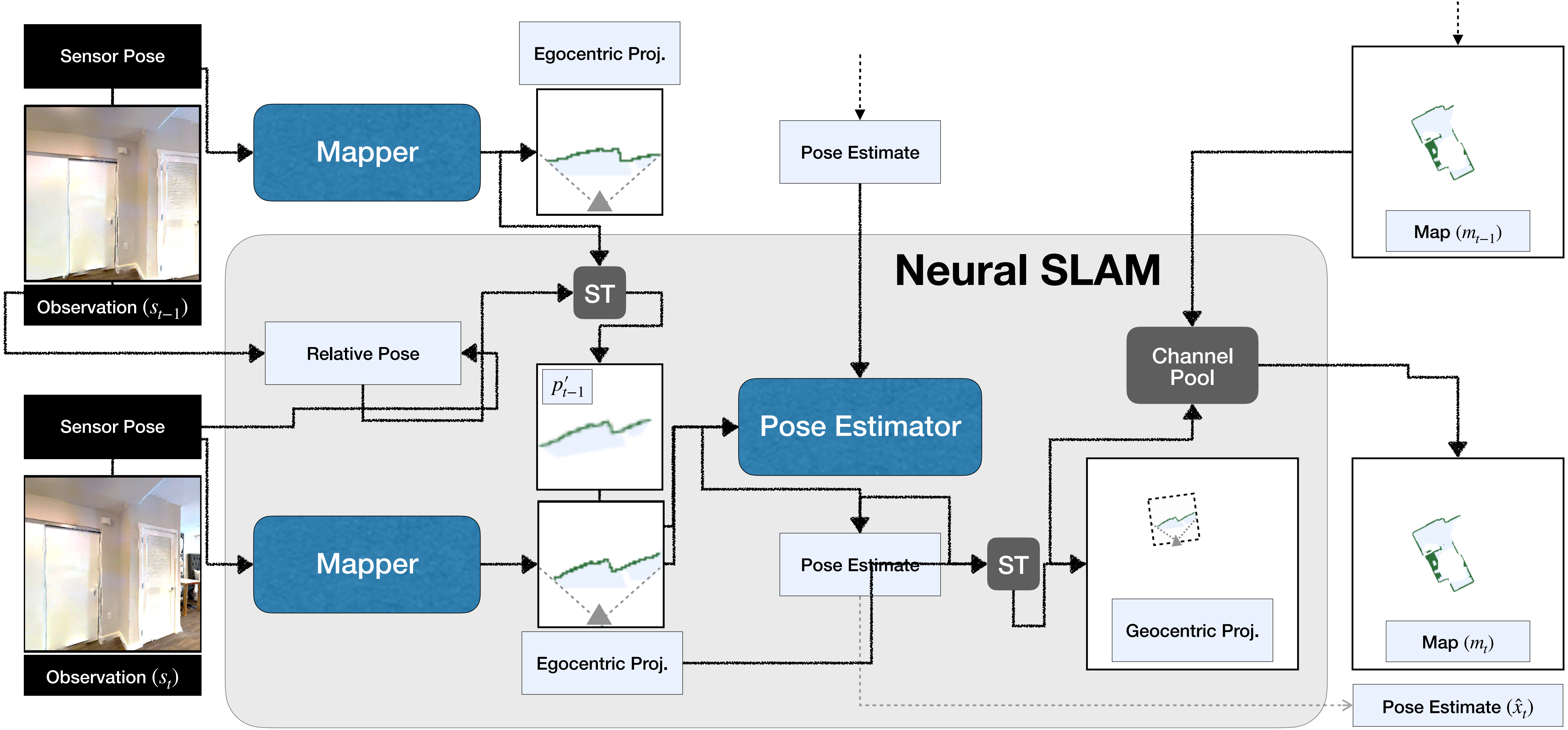
Exploration in Gibson Environment



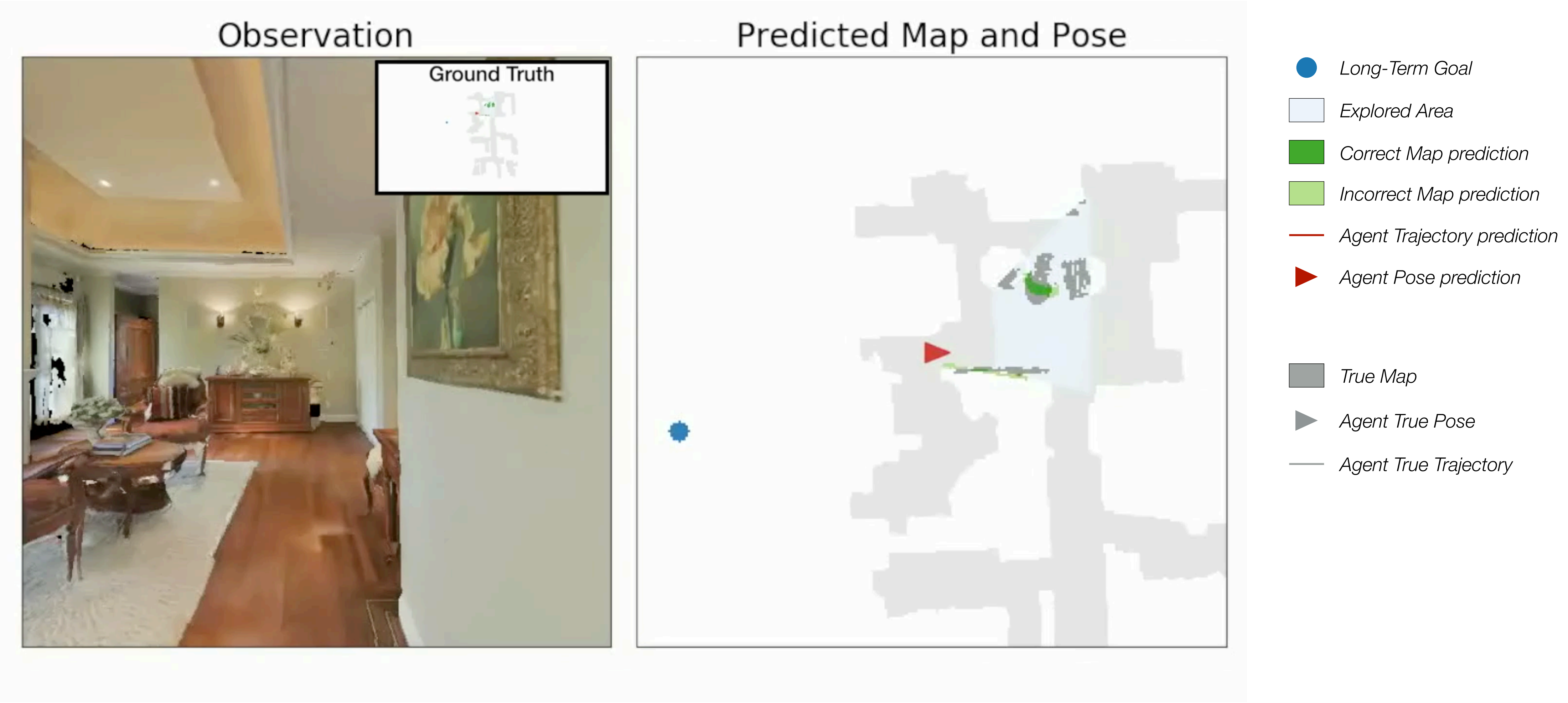
Active Neural SLAM: Overview



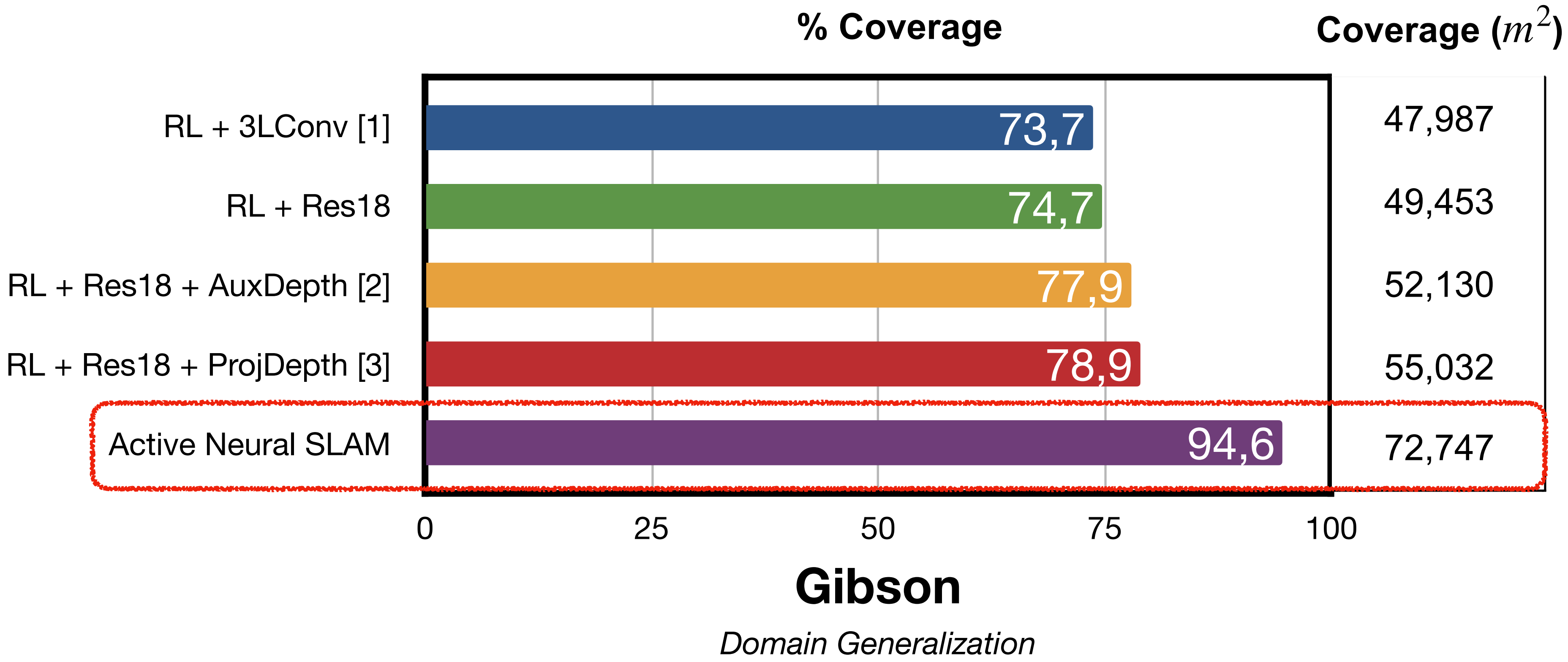
Neural SLAM Module



Domain Generalization: Matterport3D



Exploration Results



**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

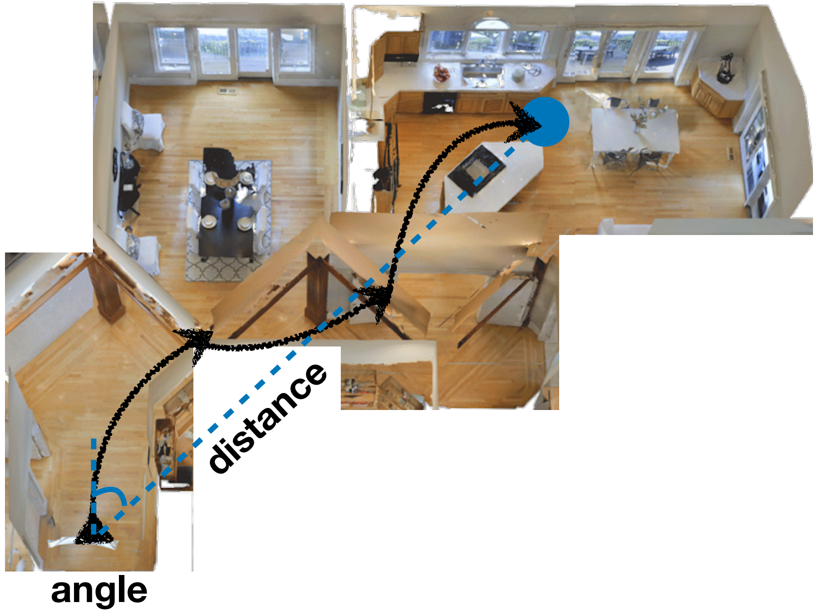
Object Goal

Chair
TV
Sofa

Language Goal

Blue Chair
Largest TV
White Sofa

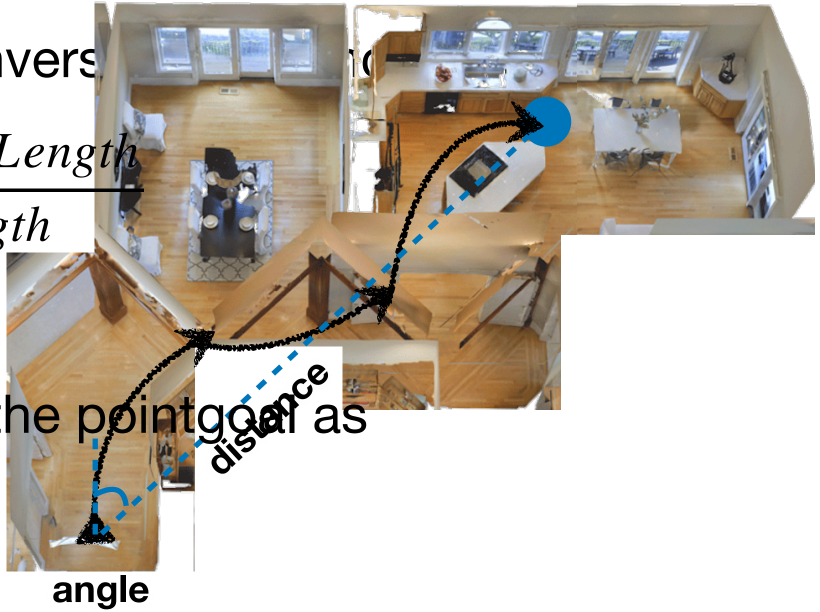
Point-Goal Navigation



Point-Goal Navigation

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

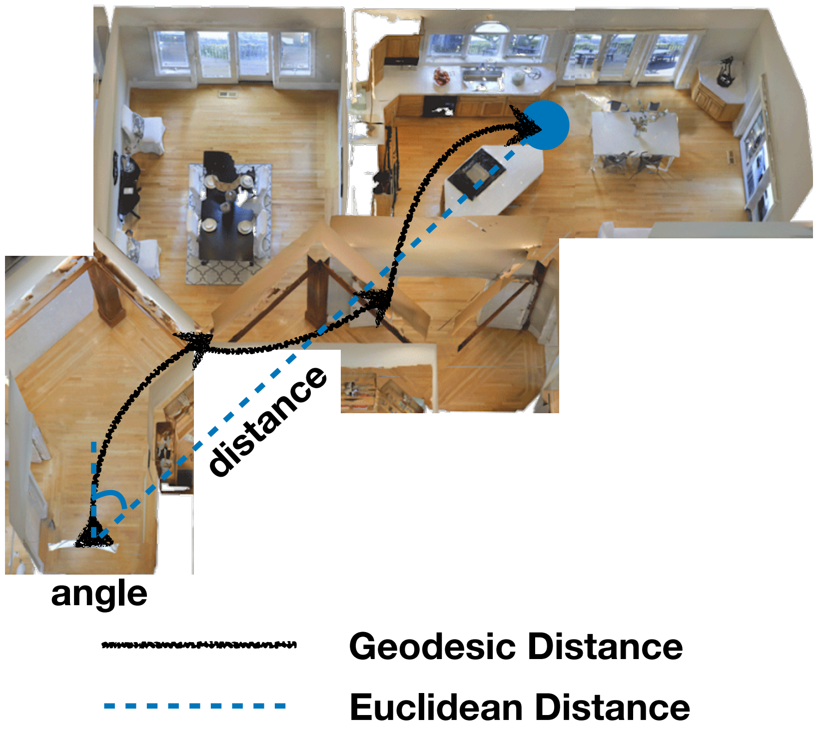
$$\frac{1}{N}\sum_{i=1}^{N} Success * \frac{ShortestPathLength}{PathLength}$$



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

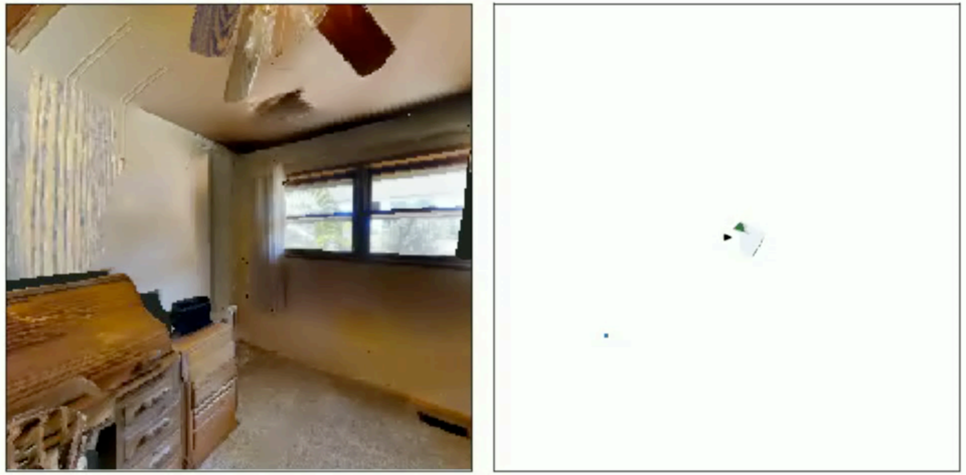
Harder Datasets

- **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



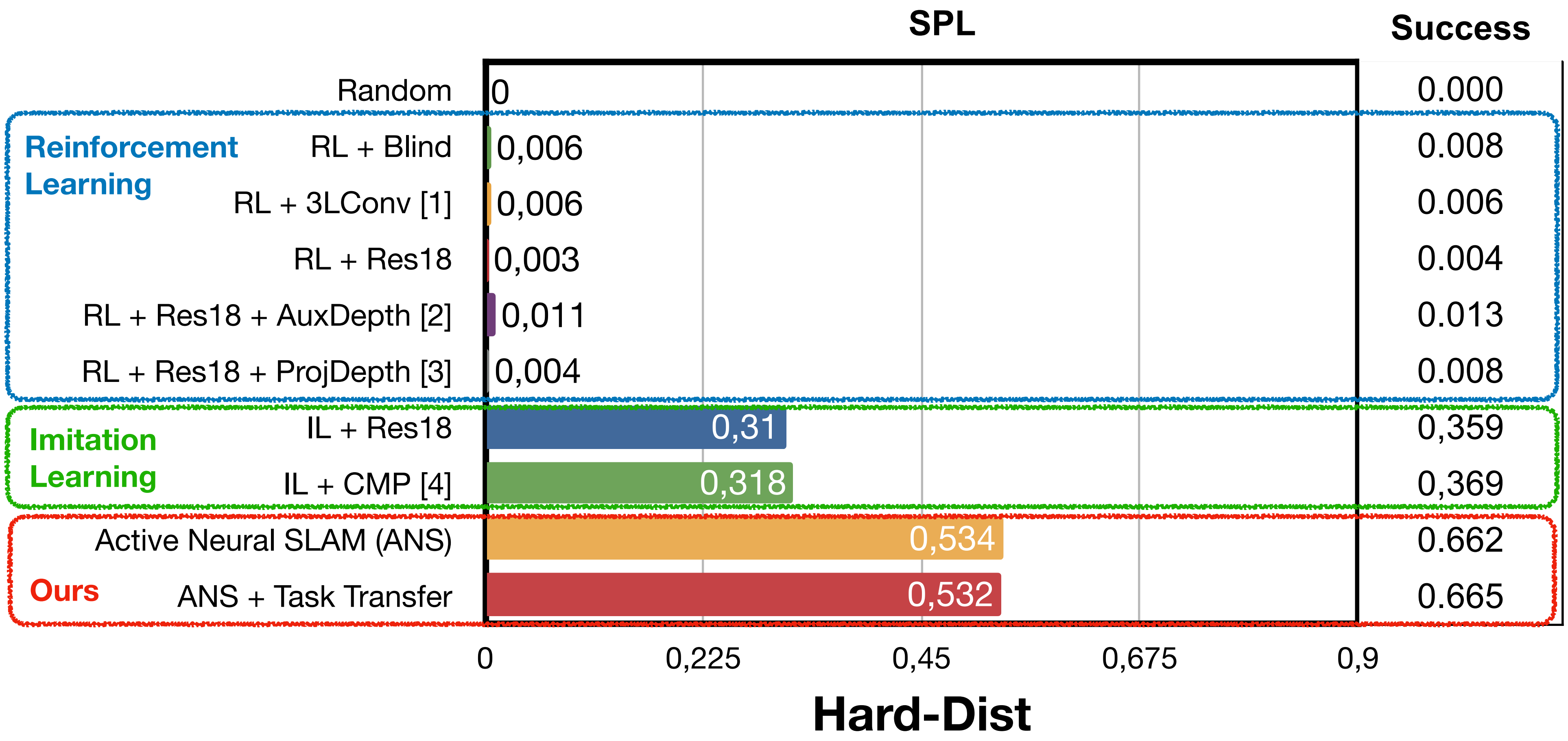
Point-Goal Navigation

Gibson



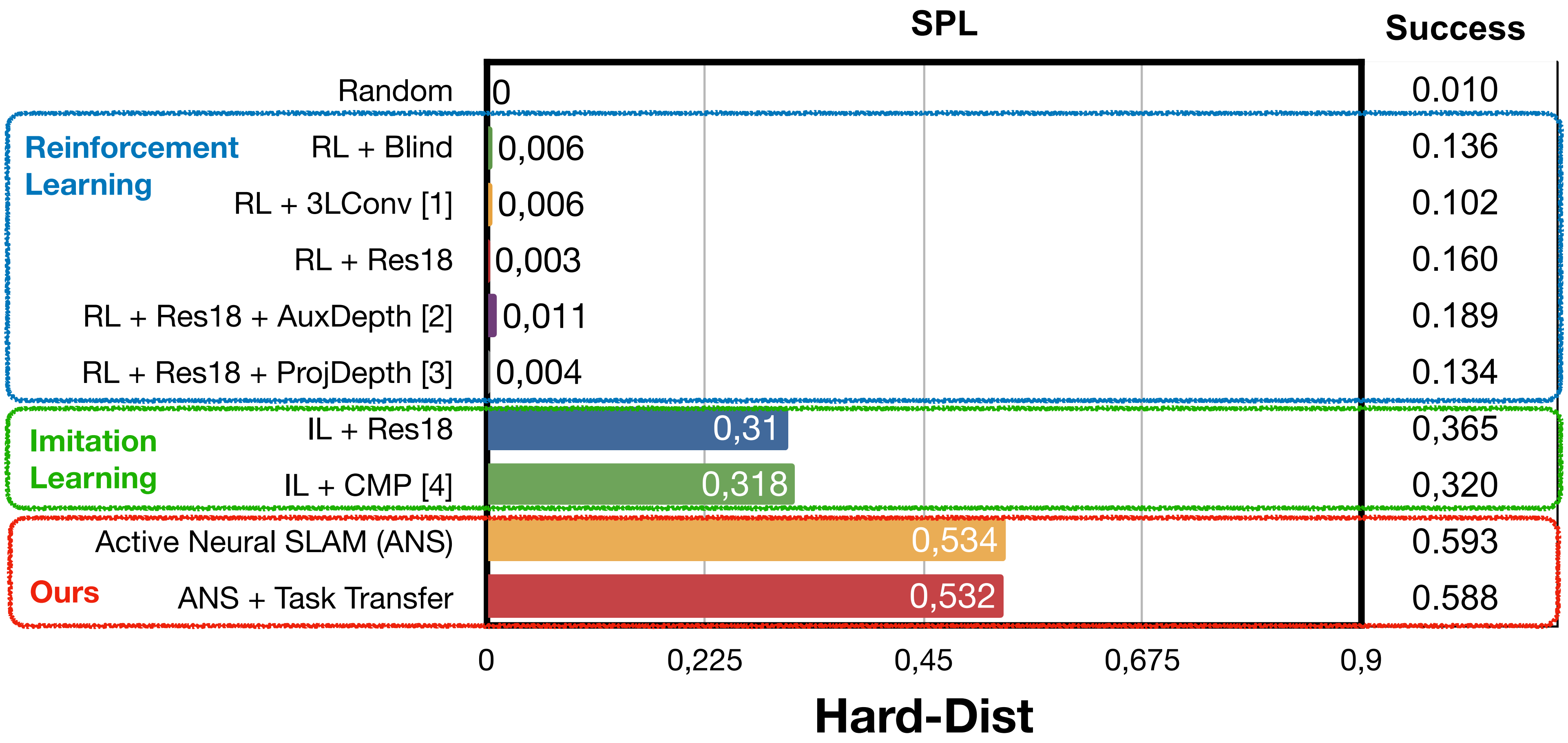
MP3D

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

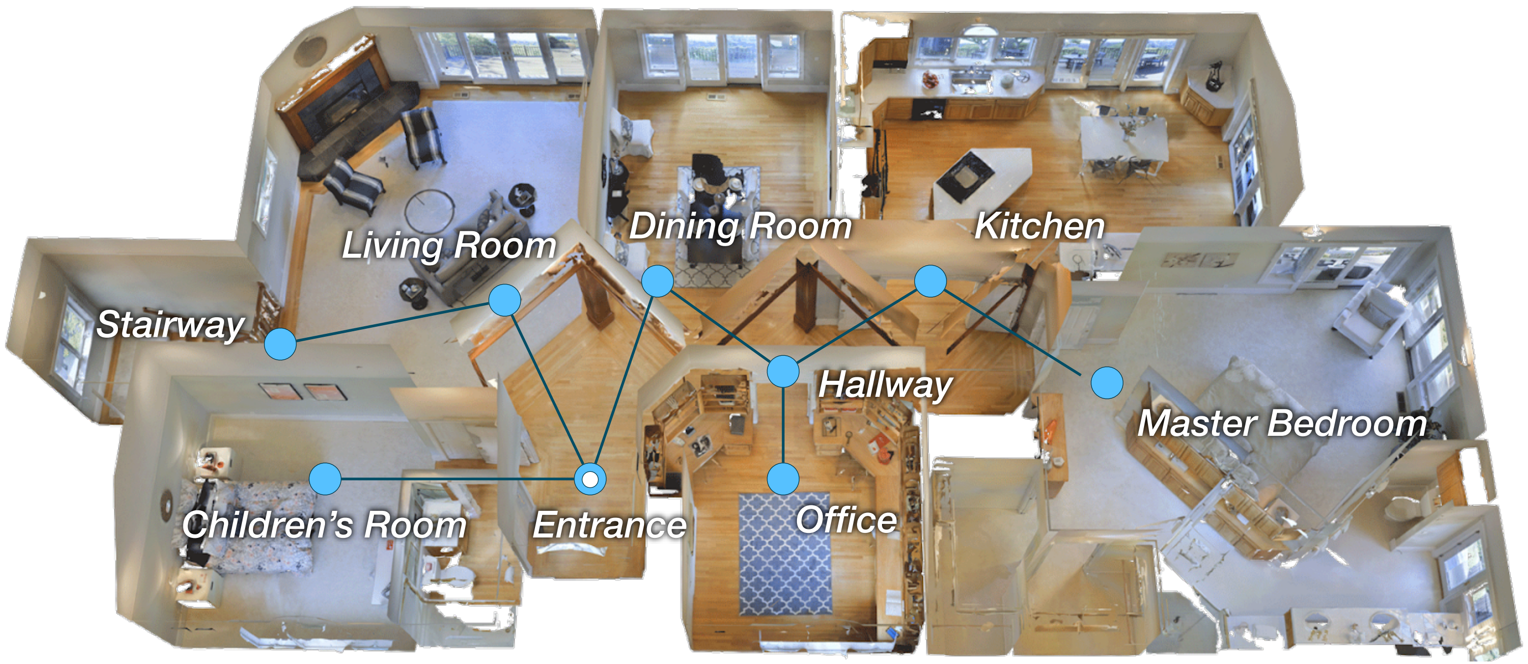
Language Goal

Blue Chair
Largest TV
White Sofa

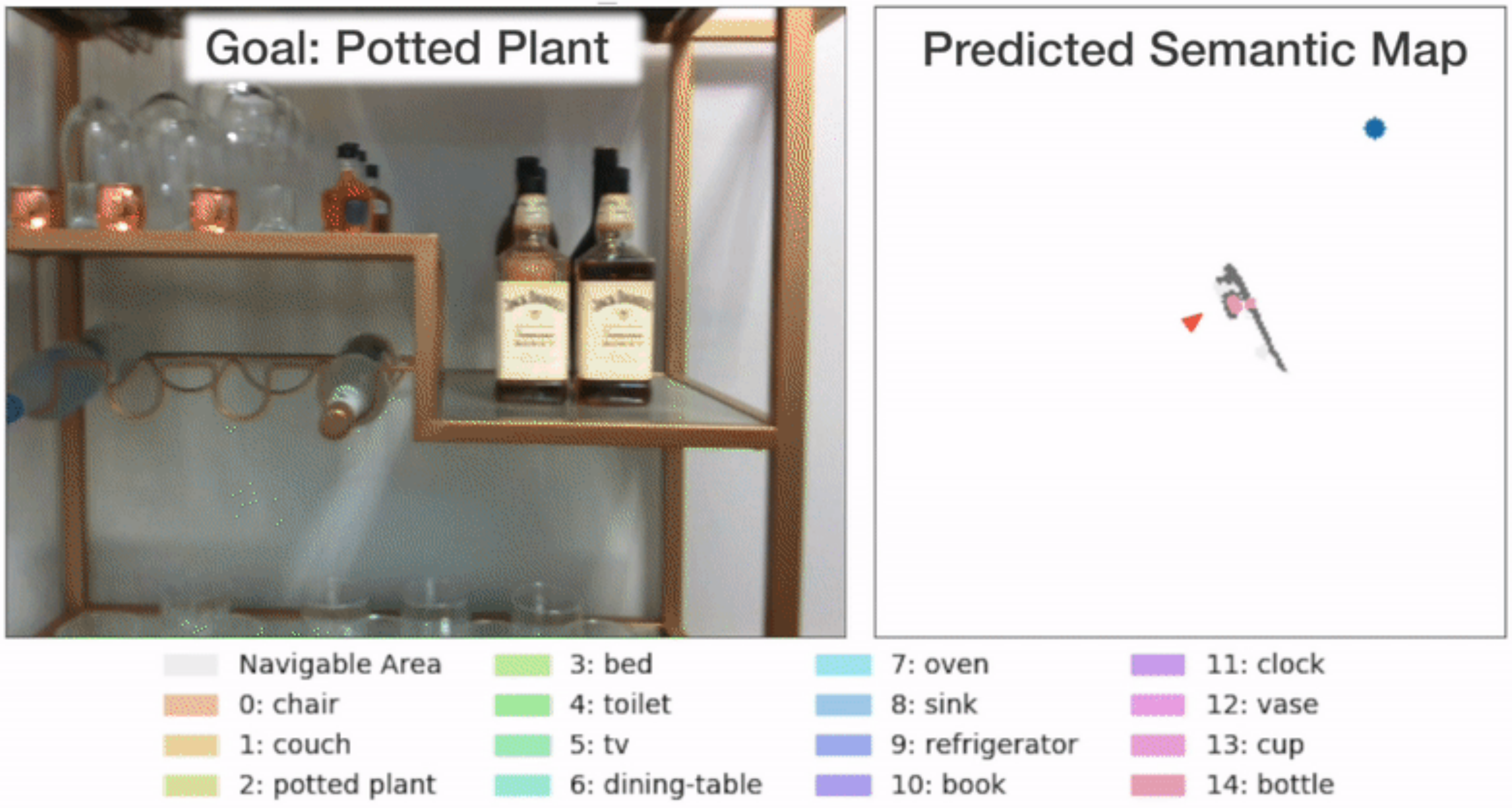
Semantic Priors and Common-Sense

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

Topological Maps



Explicit Semantic Mapping



Internet vs Embodied Data

Static Internet Data

Active Embodied Data

Using Internet models for Embodied Agents

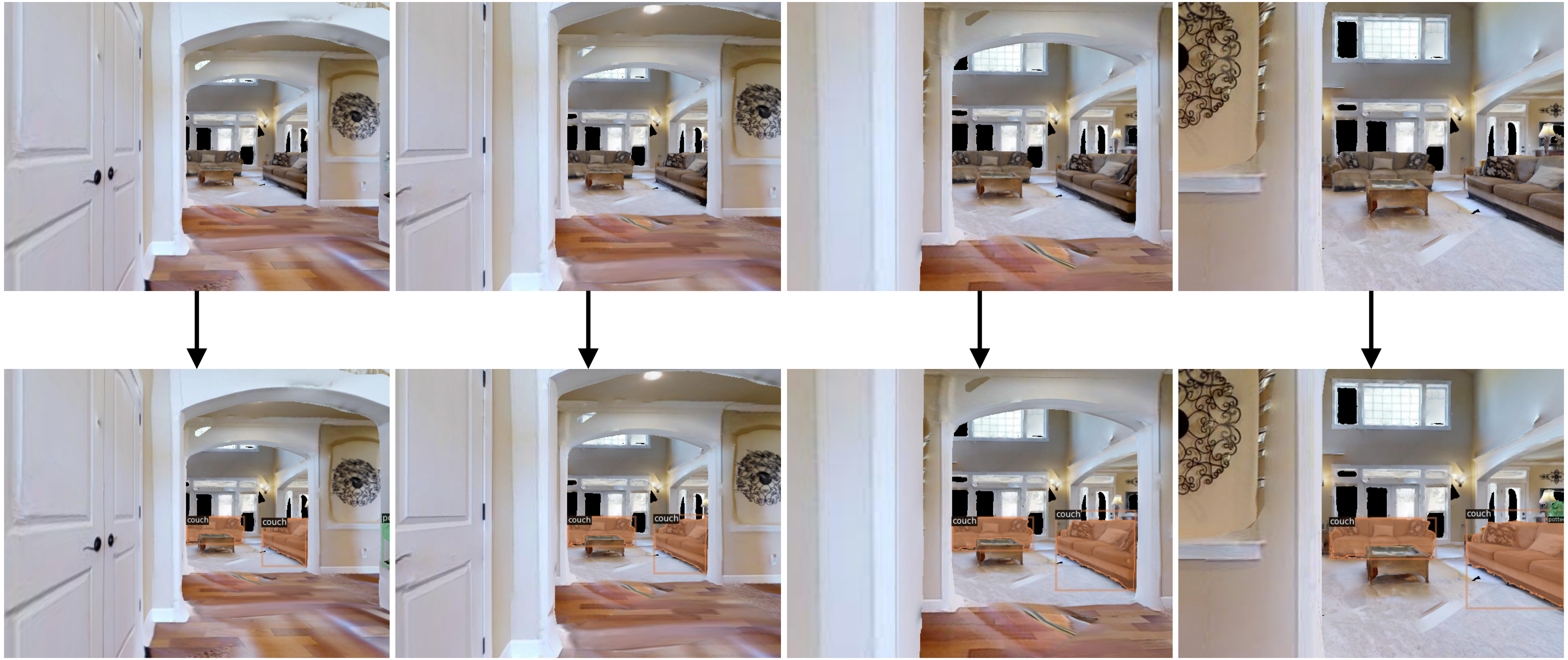
False positives

False negatives

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

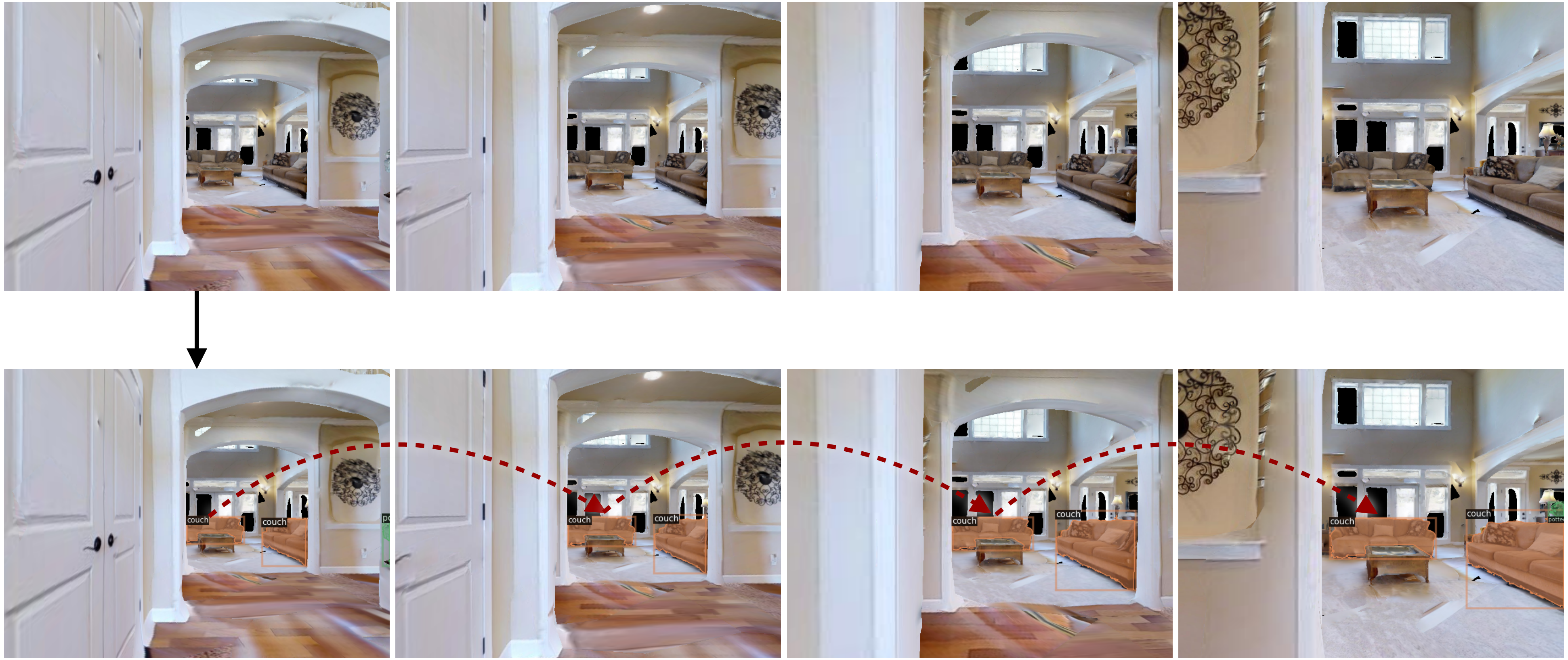
Embodied Perception

Active Embodied data

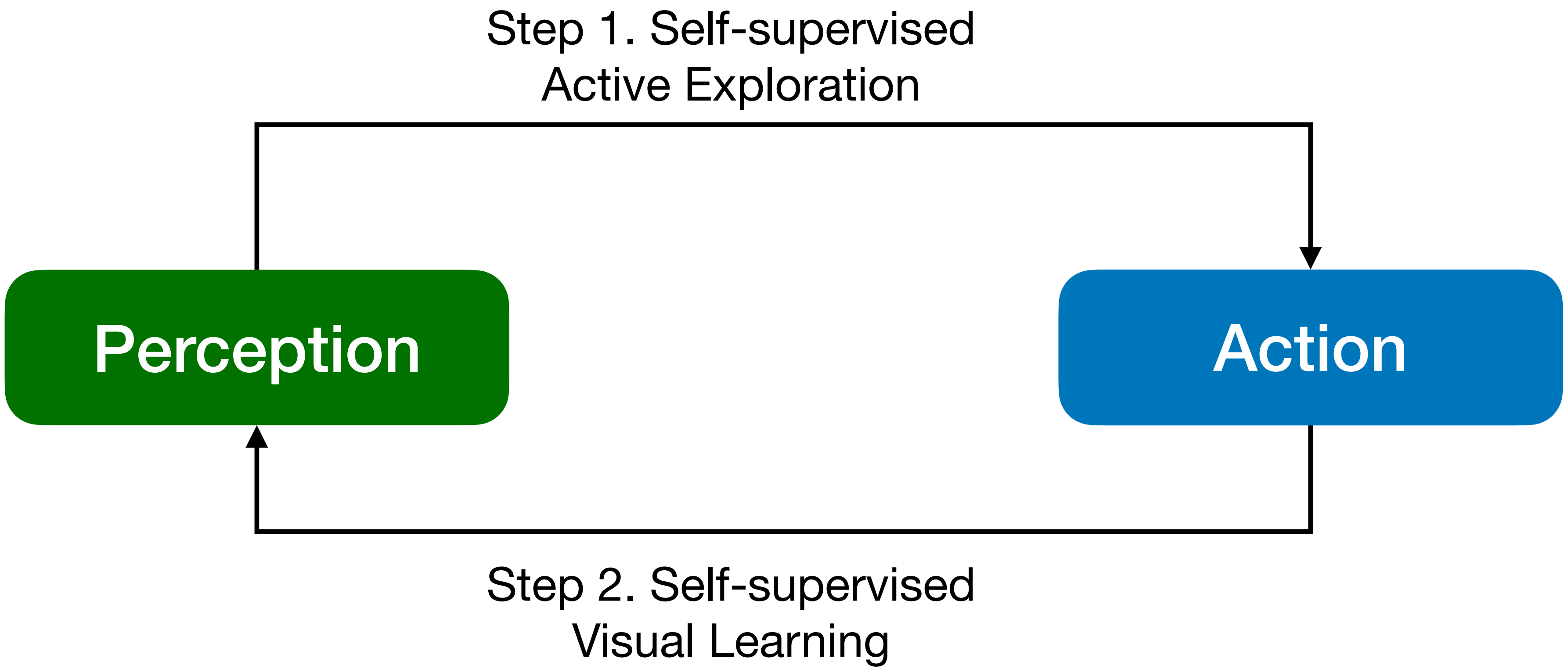


Embodied Perception

Active Embodied data

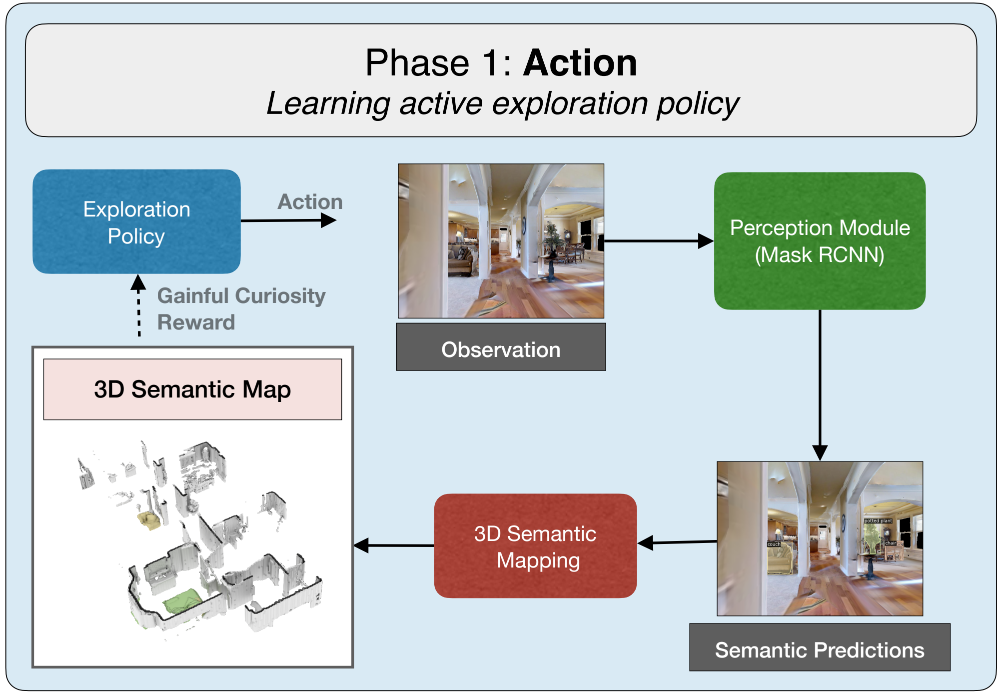


Perception-Action Loop



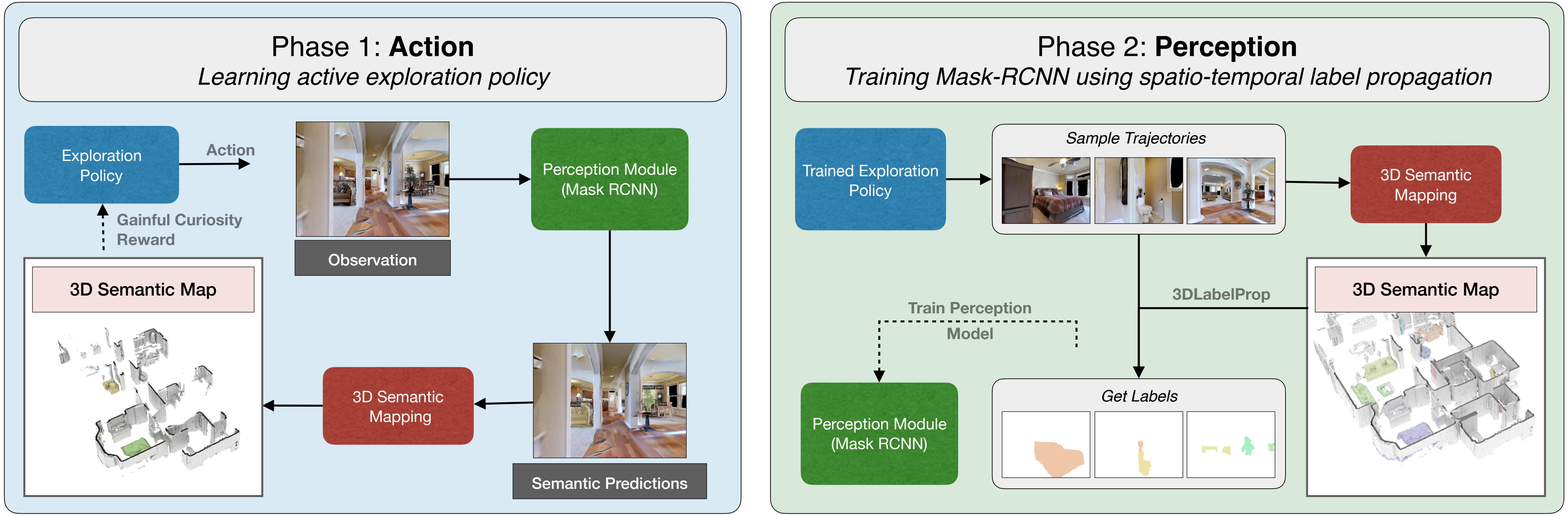
Pathak et al, Learning instance segmentation by interaction, 2018
Jang et al, Grasp2vec: Learning object representations from self-supervised grasping, 2018
Eitel et al, Self-supervised transfer learning for instance segmentation through physical interaction, 2019
Fang et al.,Move to See Better: Self-Improving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

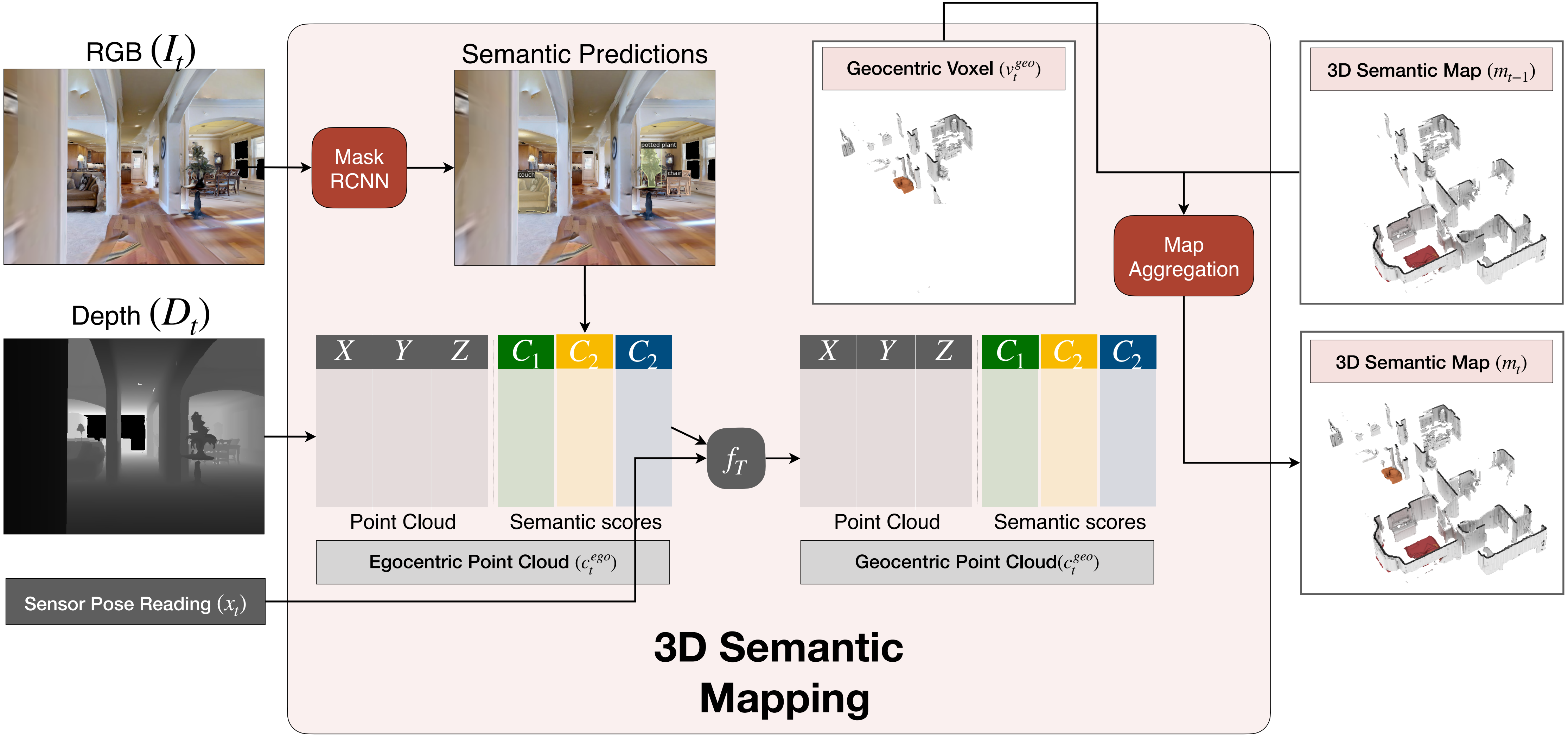
SEAL: Self-supervised Embodied Active Learning



Both phases do not require any additional labelled data

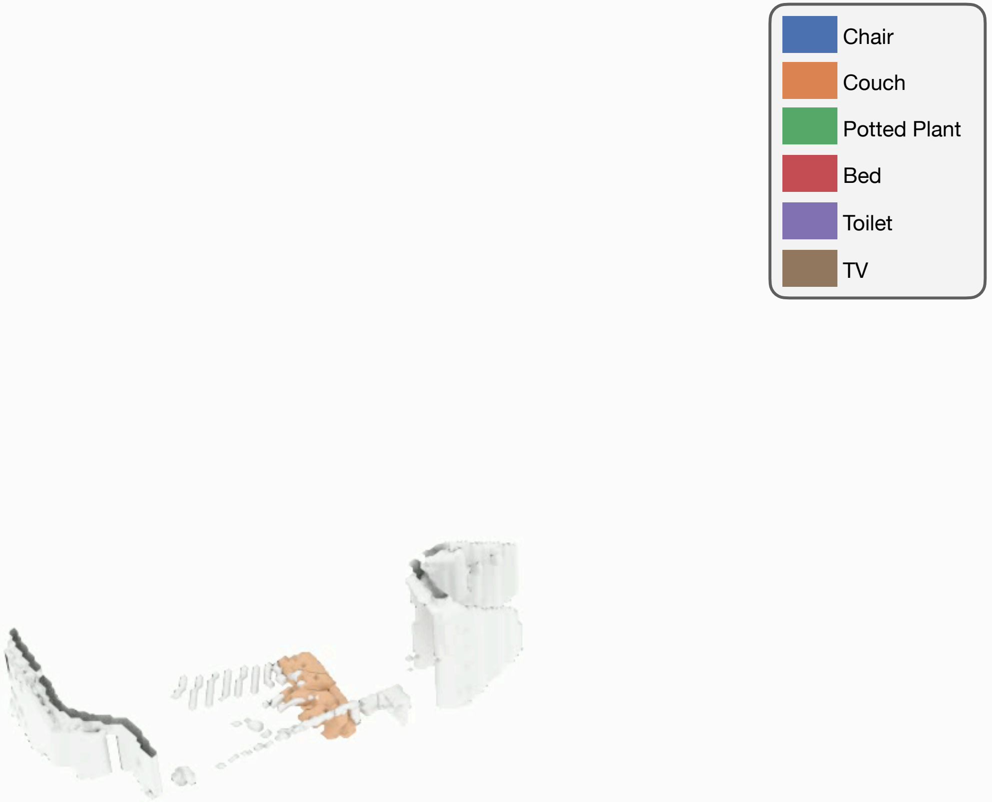
Chaplot, Dalal, Gupta, Malik, Salakhutdinov et al, . SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency NeurIPS-21

3D Semantic Mapping



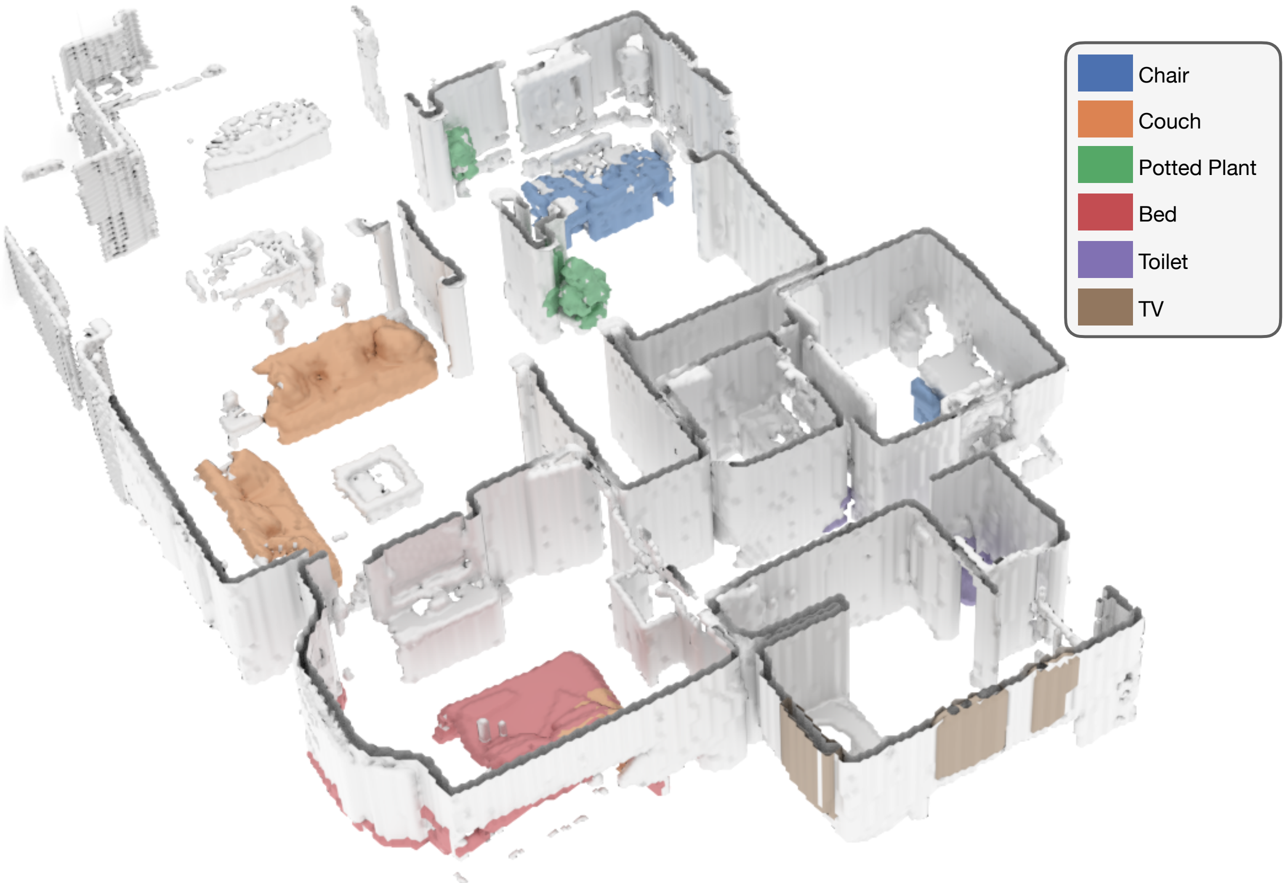
3D Semantic Mapping

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

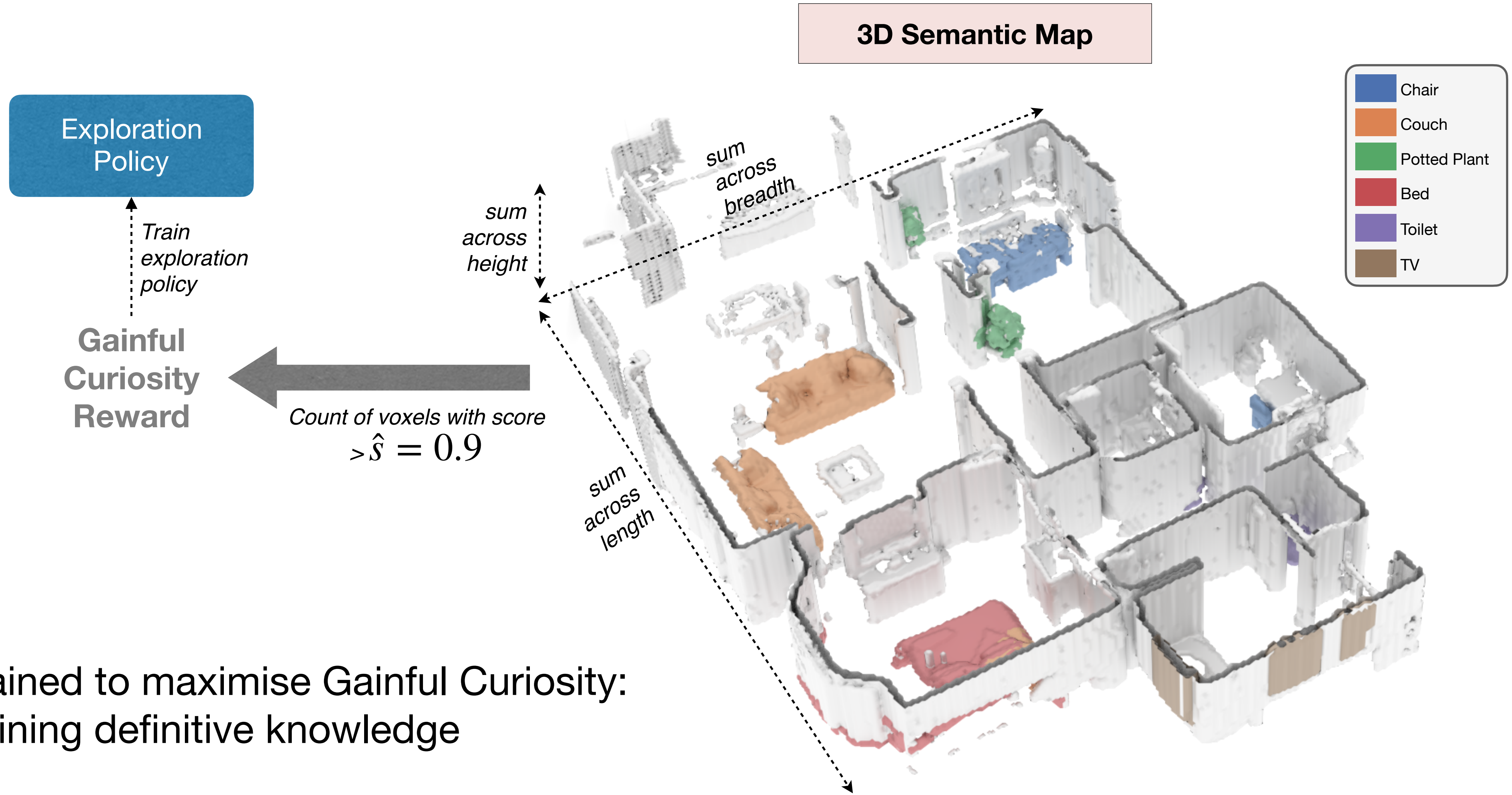


3D Semantic Mapping

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

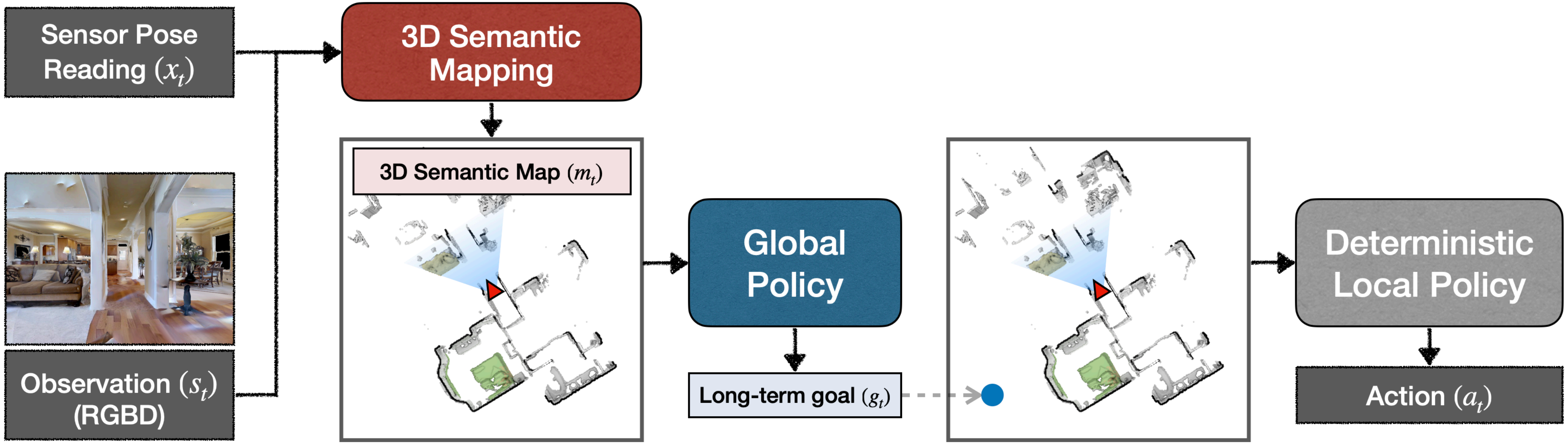


Gainful Curiosity



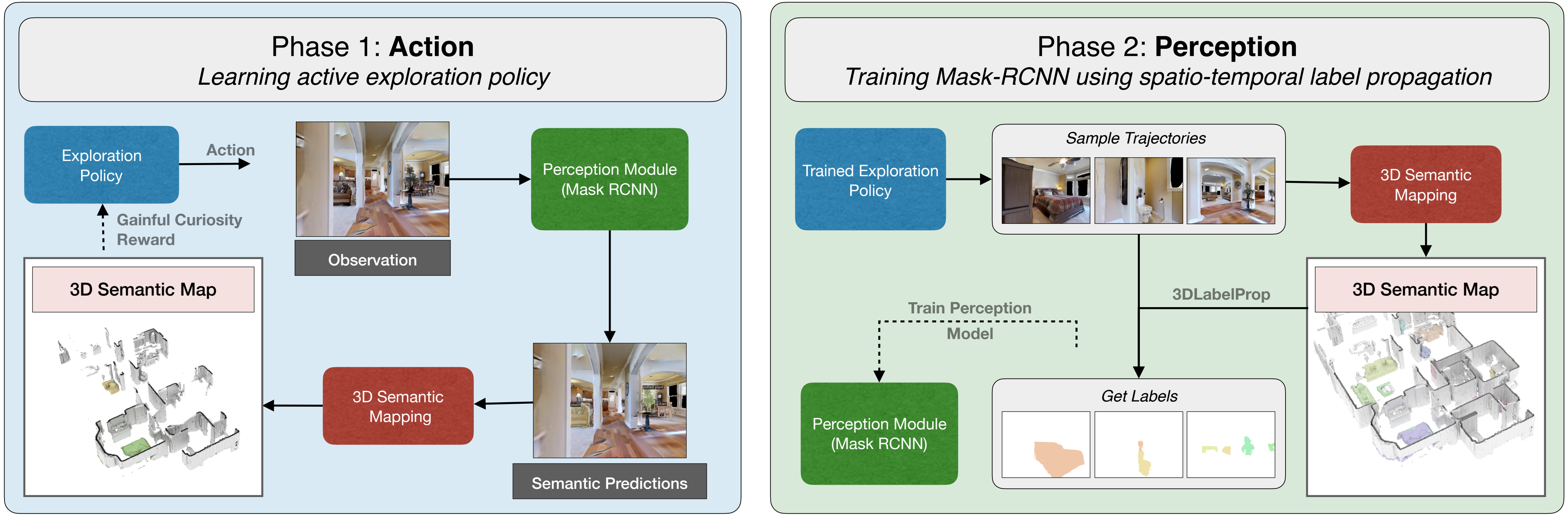
- Trained to maximise Gainful Curiosity: gaining definitive knowledge

Policy Learning



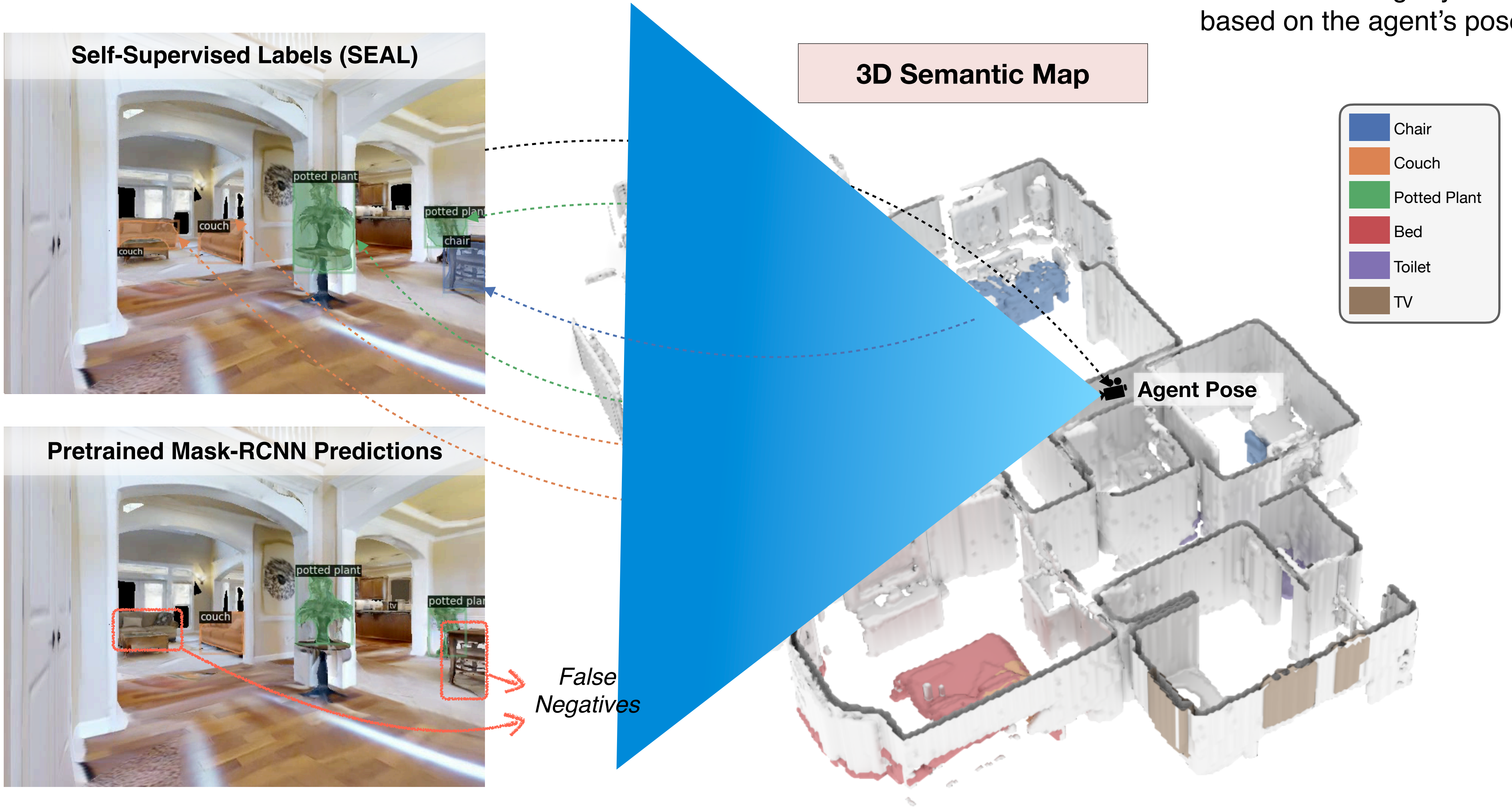
- Global Policy: samples a goal every 25 local steps
- Action Space: move forward (25cm), turn left or right (30 degrees)

SEAL: Self-supervised Embodied Active Learning

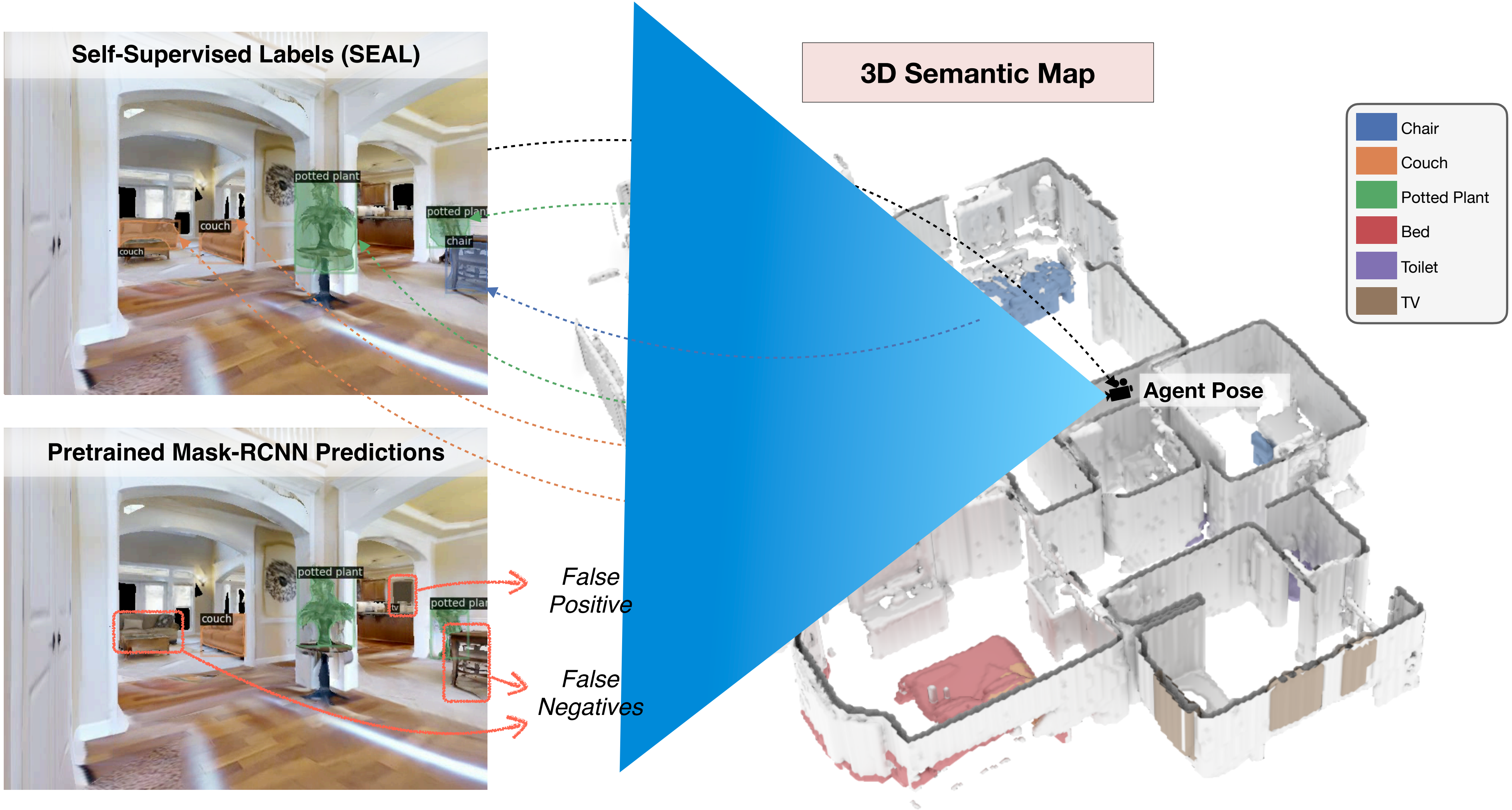


3D Label Propagation

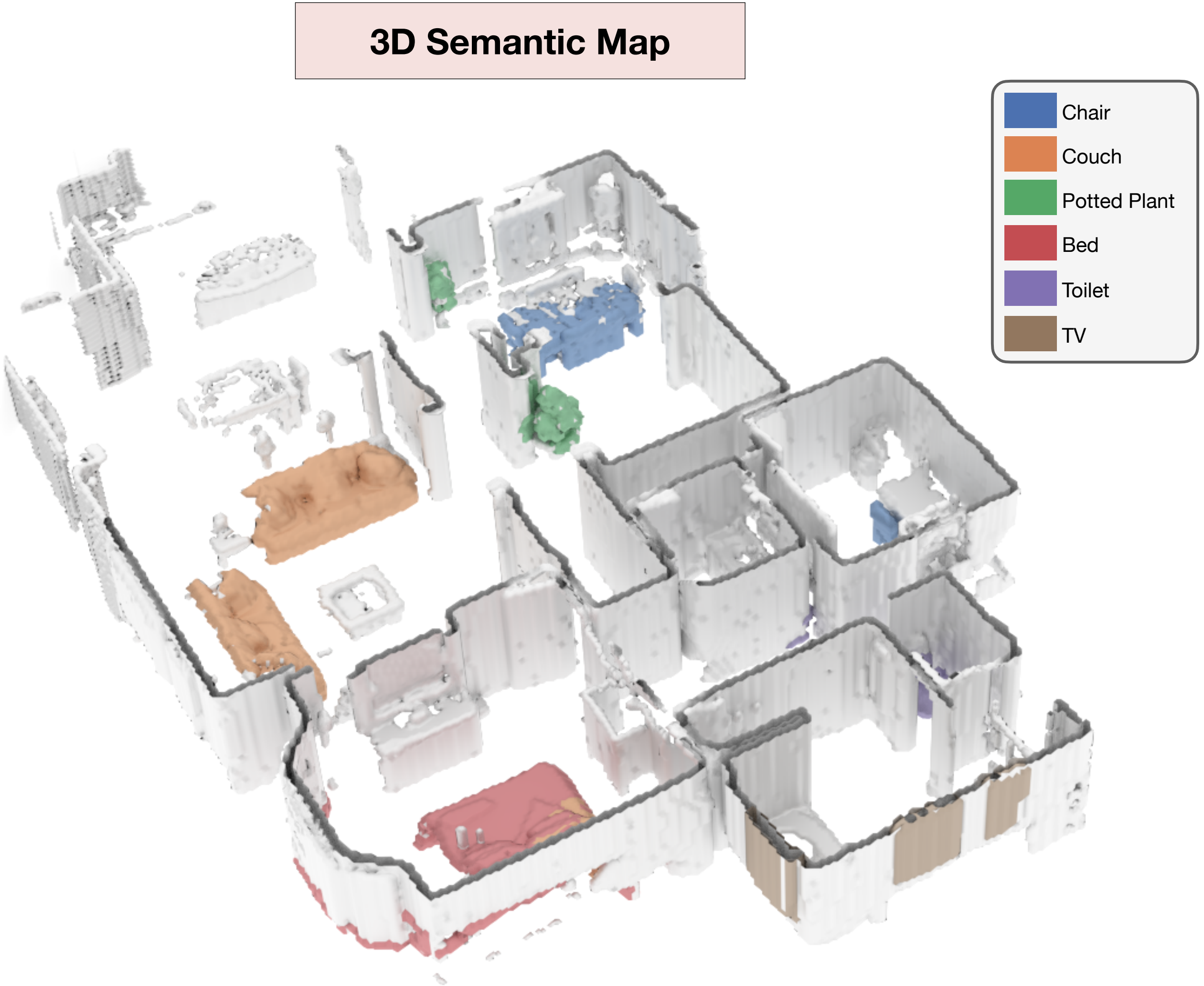
Instance label for each pixel is obtained using ray tracing based on the agent's pose



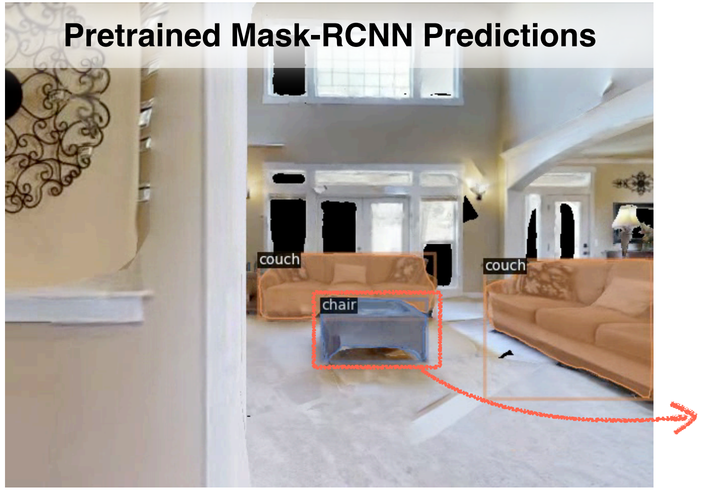
3D Label Propagation



3D Label Propagation

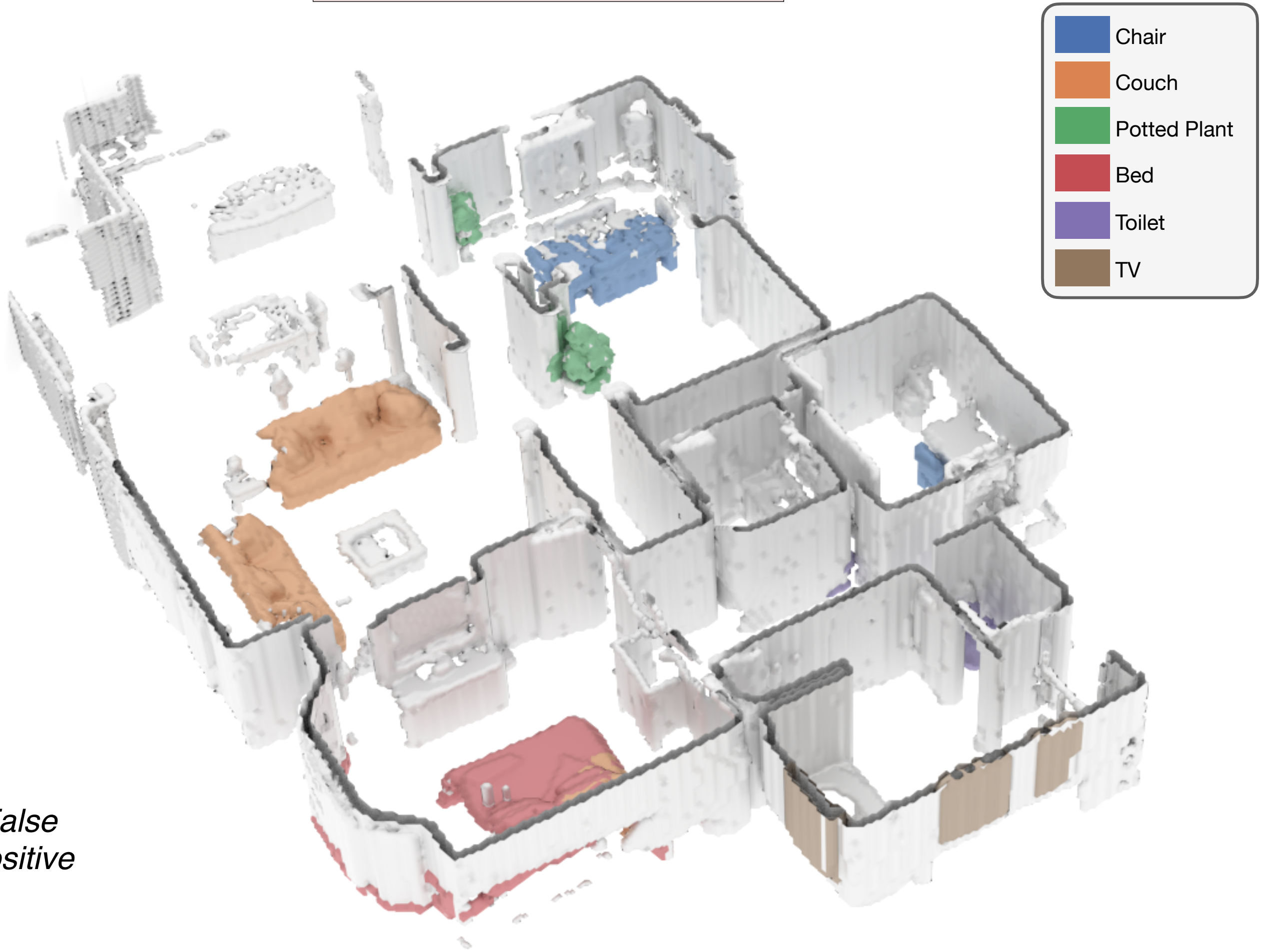


3D Label Propagation



False Positive

3D Semantic Map

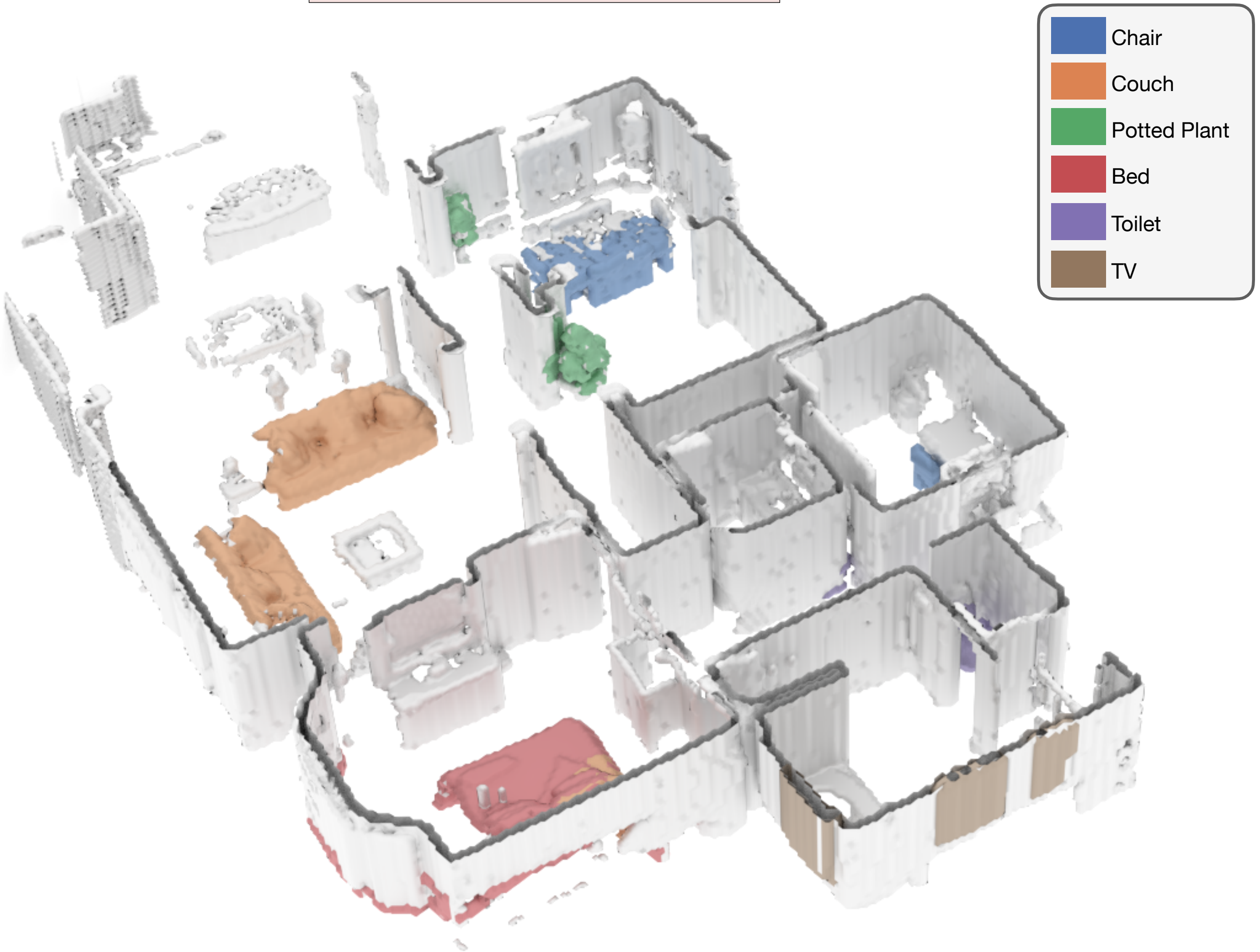


3D Label Propagation

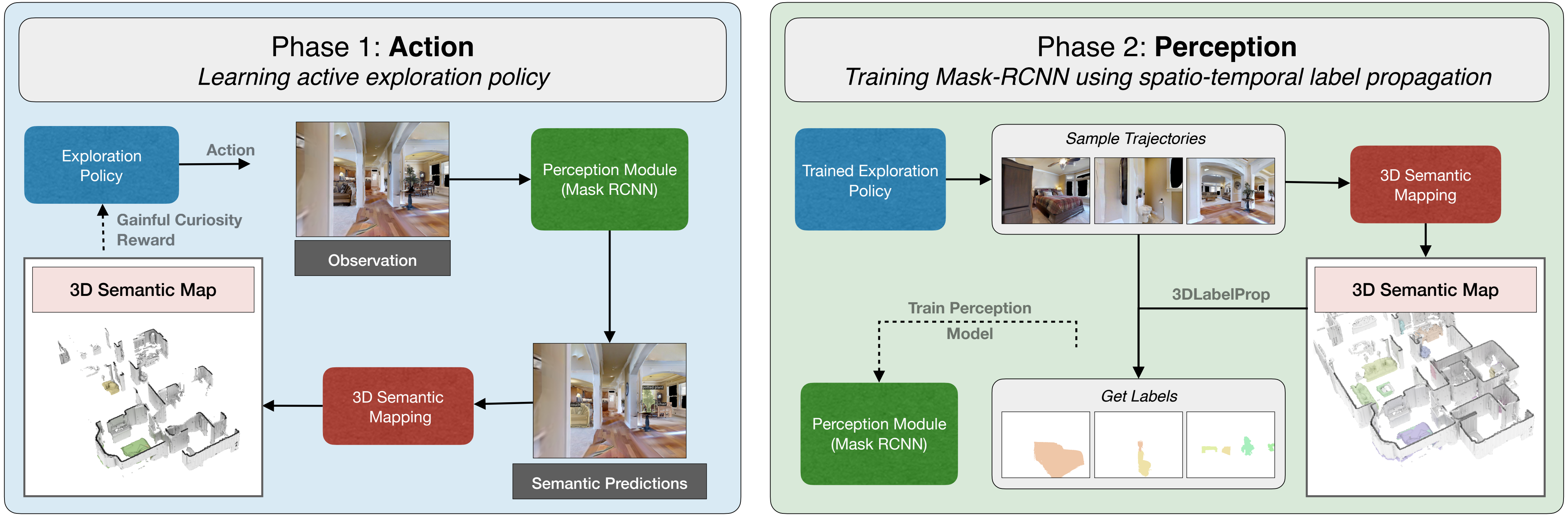
Train Perception Model

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

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

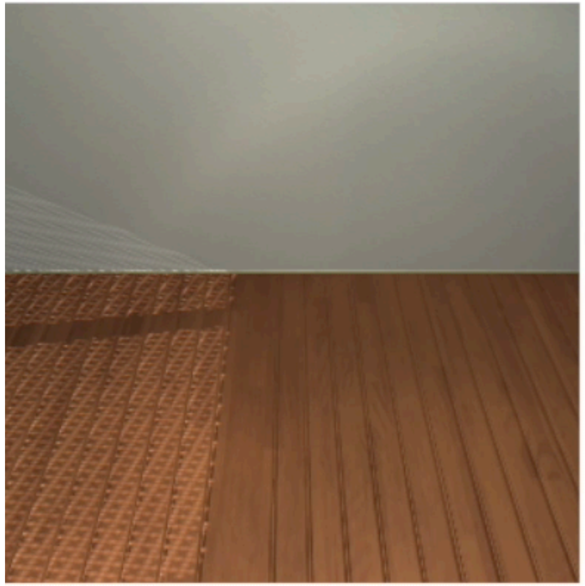
Results

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

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

RGB



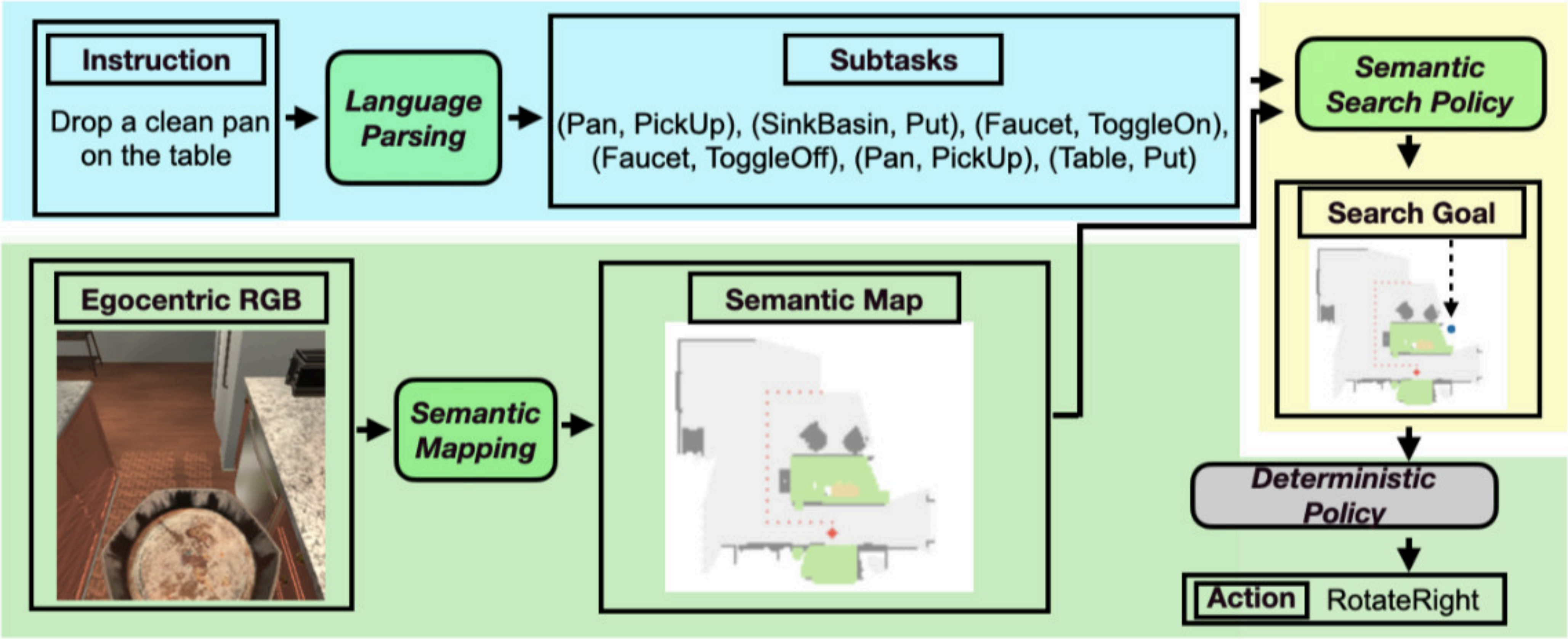
Completed Subgoals

- 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

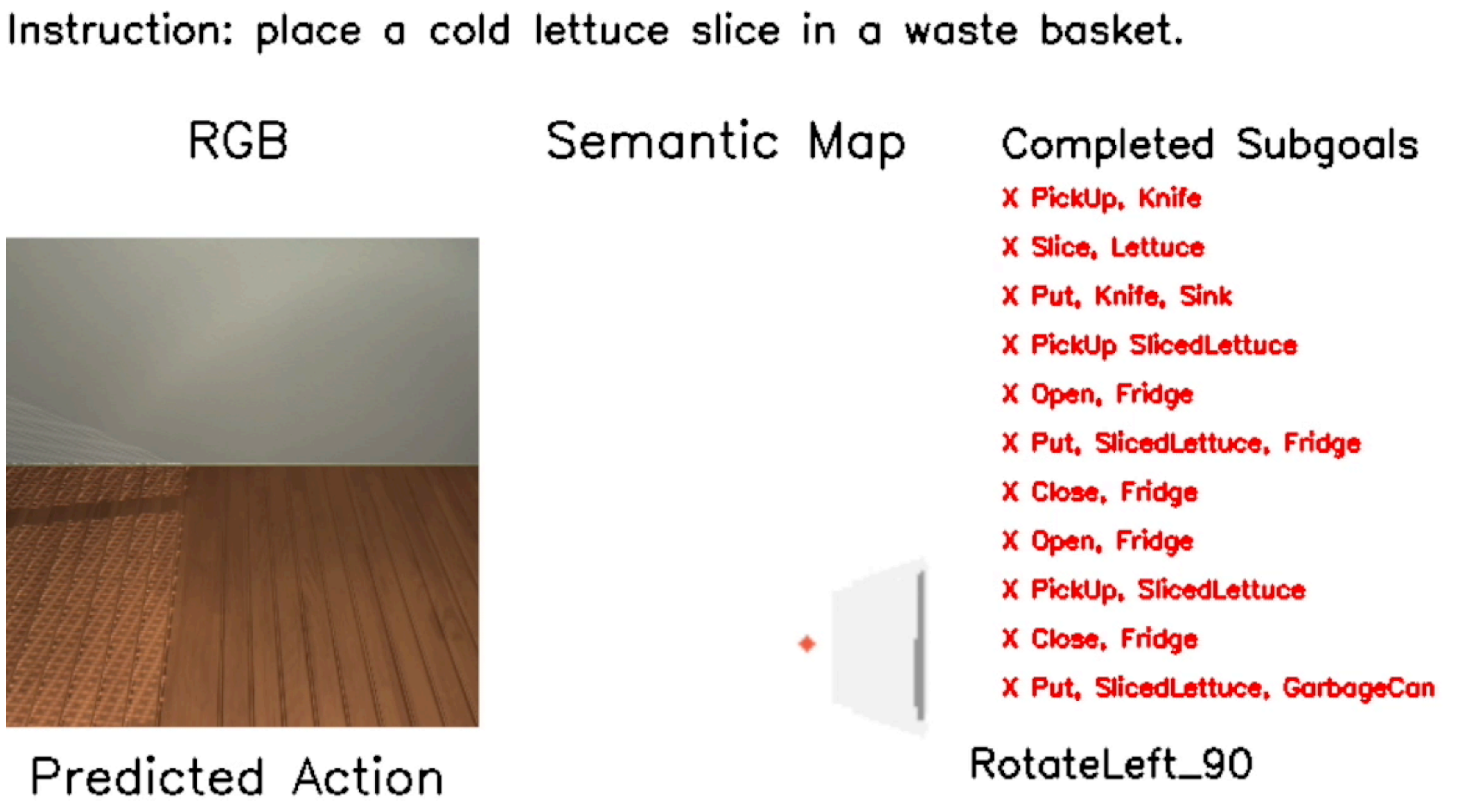
Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox.
Alfred: A benchmark for interpreting grounded instructions for everyday tasks

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



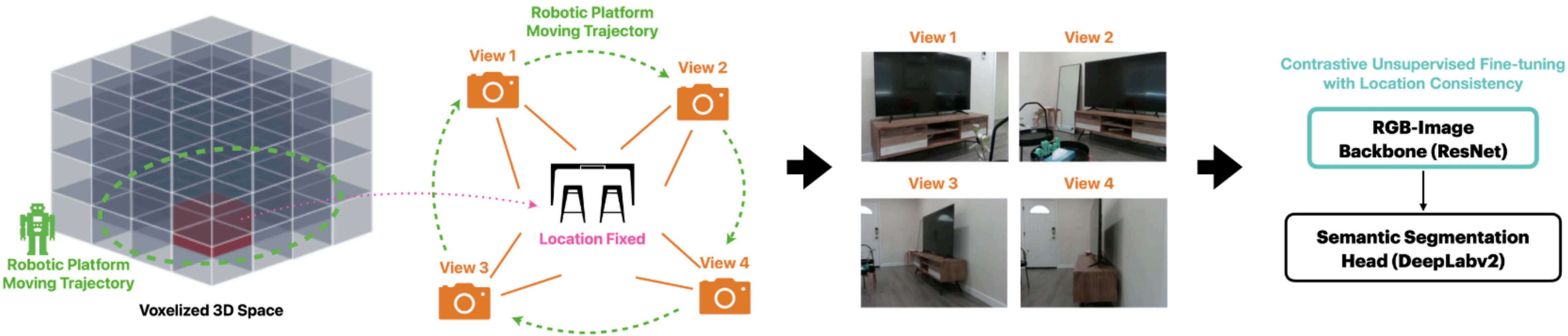
Results

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		13.10	35.59	9.43	25.90	13.37	35.51	10.17	23.94
FILM		15.06	38.51	11.23	27.67	**14.30**	**36.37**	**10.55**	**26.49**
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		12.22	34.41	8.65	24.72	12.69	34.00	9.44	22.56
FILM		**14.17**	**36.15**	**10.39**	**25.77**	**13.13**	**34.75**	**9.67**	**24.46**

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

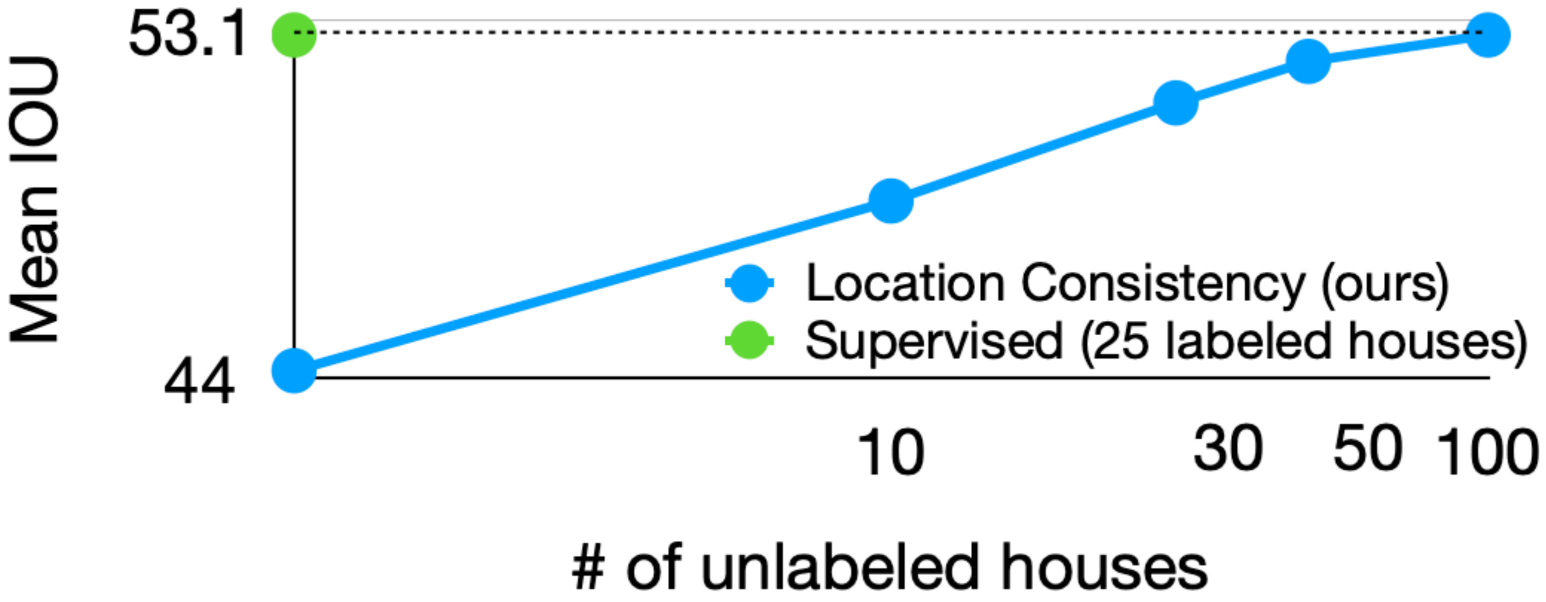


Object Goal Navigation with End-to-End Self-Supervision, S. Min, H. Tsai, W. Ding, A. Farhadi, R. Salakhutdinov, Y. Bisk, J. Zhang, 2023

Finding Bed

Object Goal Navigation with End-to-End Self-Supervision, S. Min, H. Tsai, W. Ding, A. Farhadi, R. Salakhutdinov, Y. Bisk, J. Zhang, 2023

Self-Supervision: Semantic Segmentation

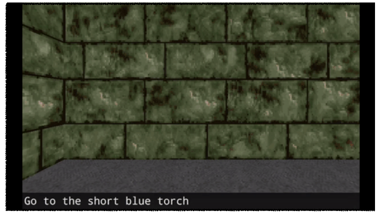


Simulation to Real

Games

ViZDoom

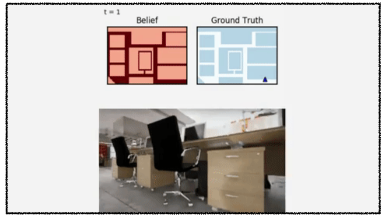
[CL *AAAI-17*]



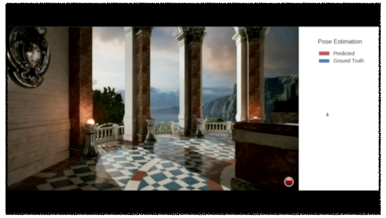
[CMPRS *AAAI-18*]

Photorealistic simulation

Unreal



[CPS *ICLR-18*]

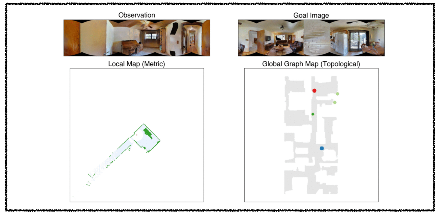


[PCZS *CVPR-18 (w)*]

Reconstructed simulation

Habitat (Gibson, MP3D)

[CGSGG *ICLR-20*]



[CSGG *CVPR-20*]

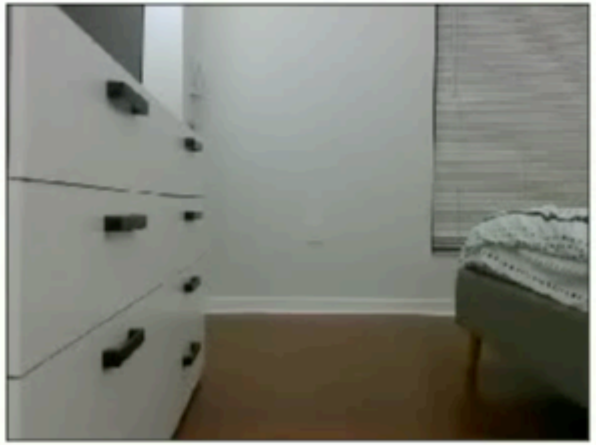
Physical Domain Gap

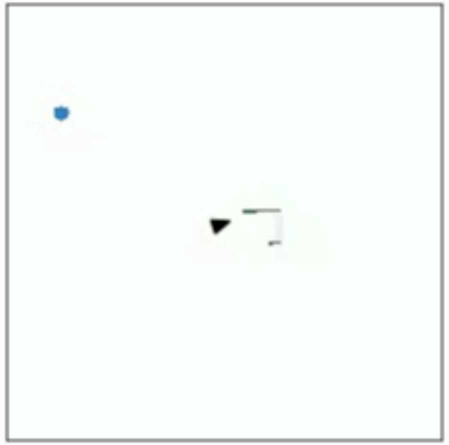
Visual Domain Gap

Real-world

?

Simulation to Real





Building Intelligent Agents

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