Representations for Visual Navigation and How to Train Them

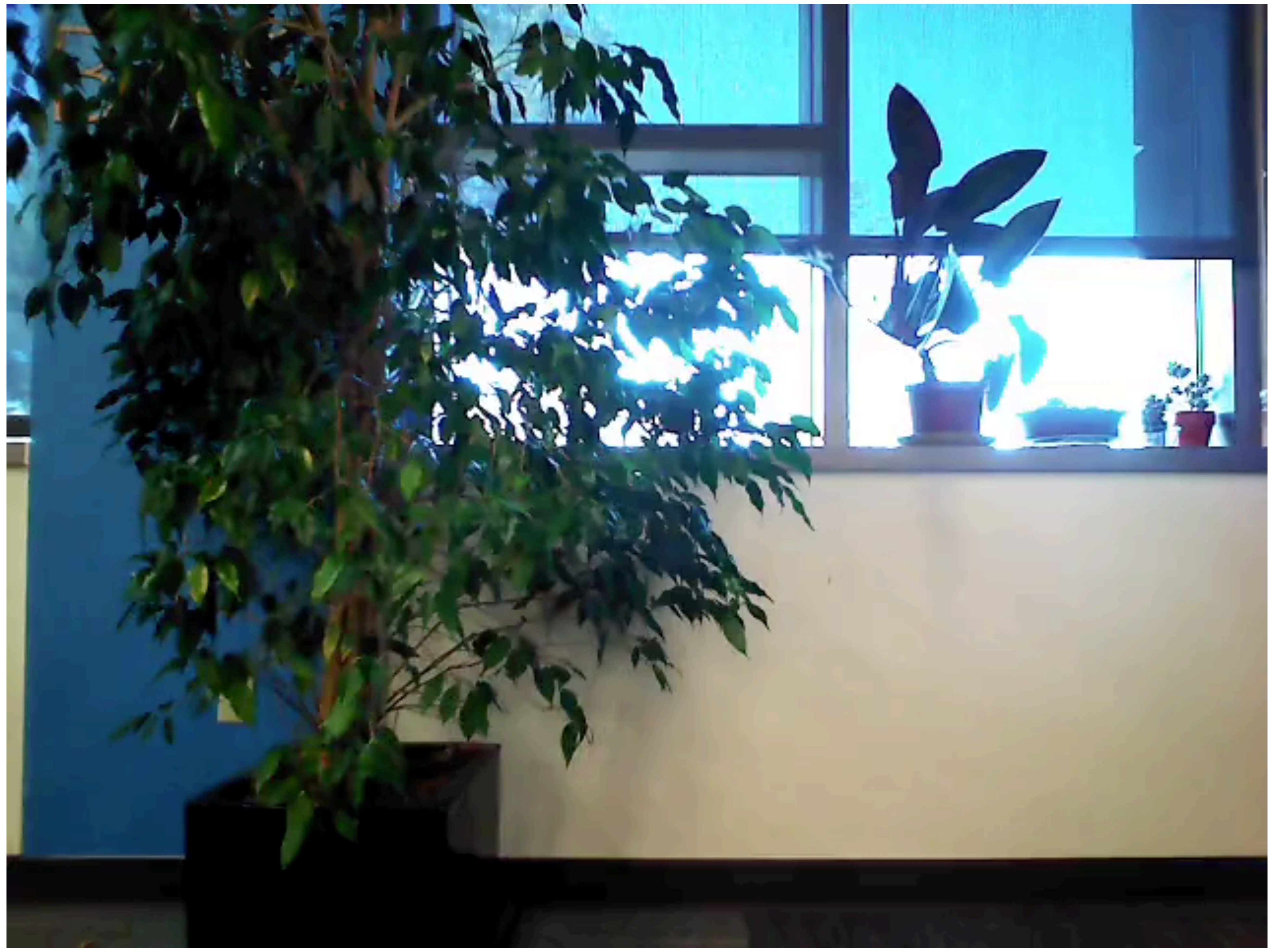
Saurabh Gupta
UIUC

In this talk,

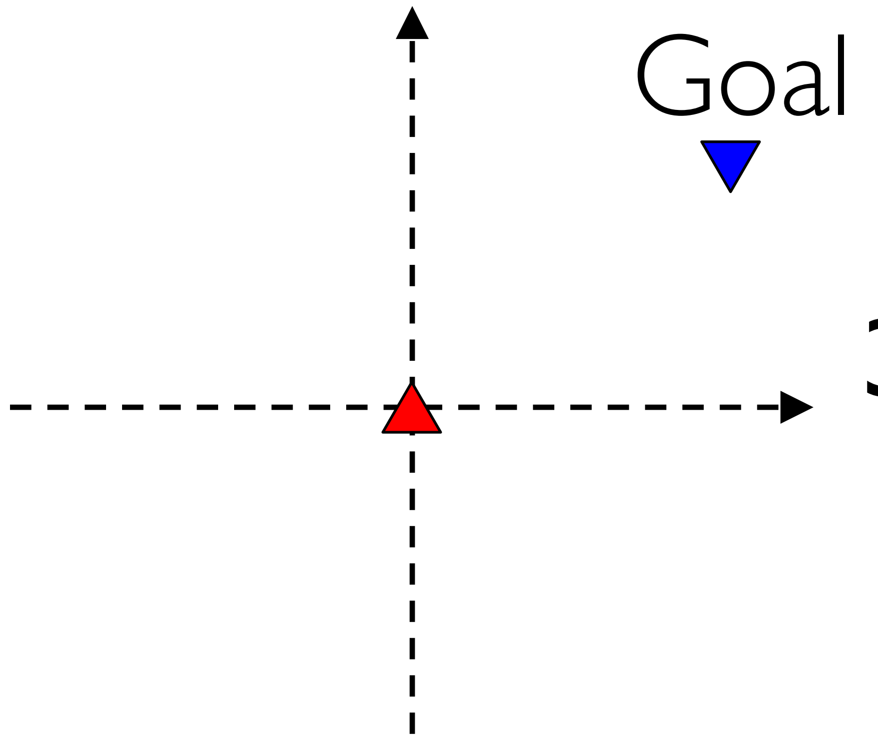
Representations for Places that Afford Navigation in Novel Environments

Basic Navigation Problems

Robot with a first person camera



Dropped into a novel environment



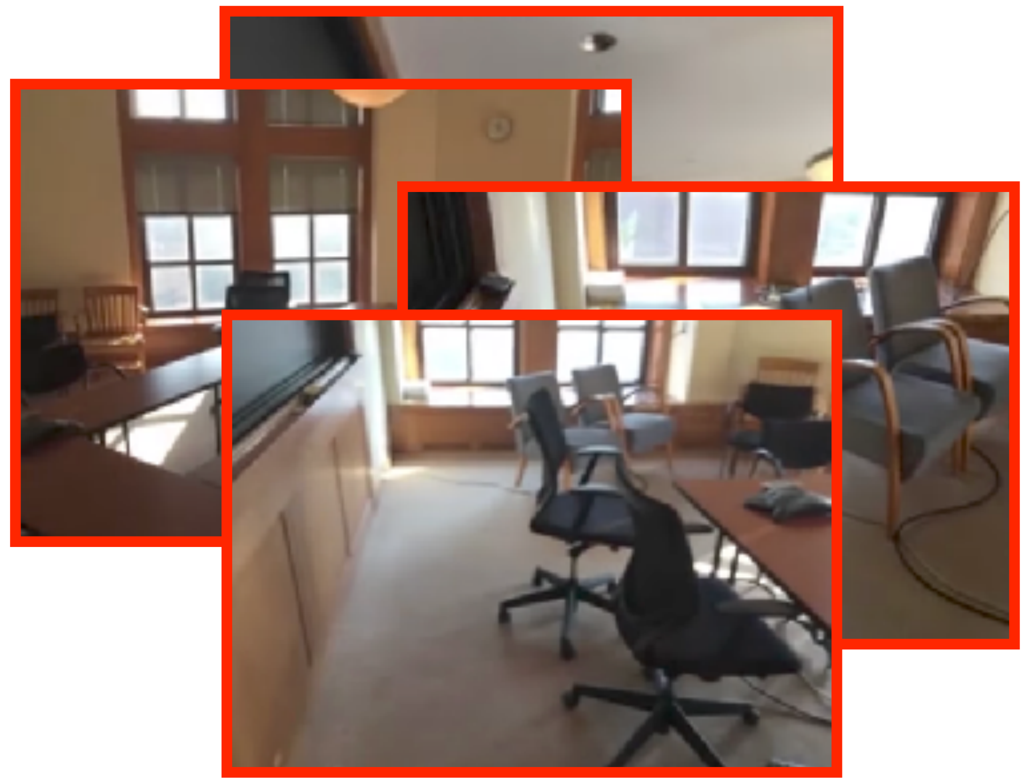
"Go
300 feet North,
400 feet East"

"Go Find a Chair"

"Explore the Environment"

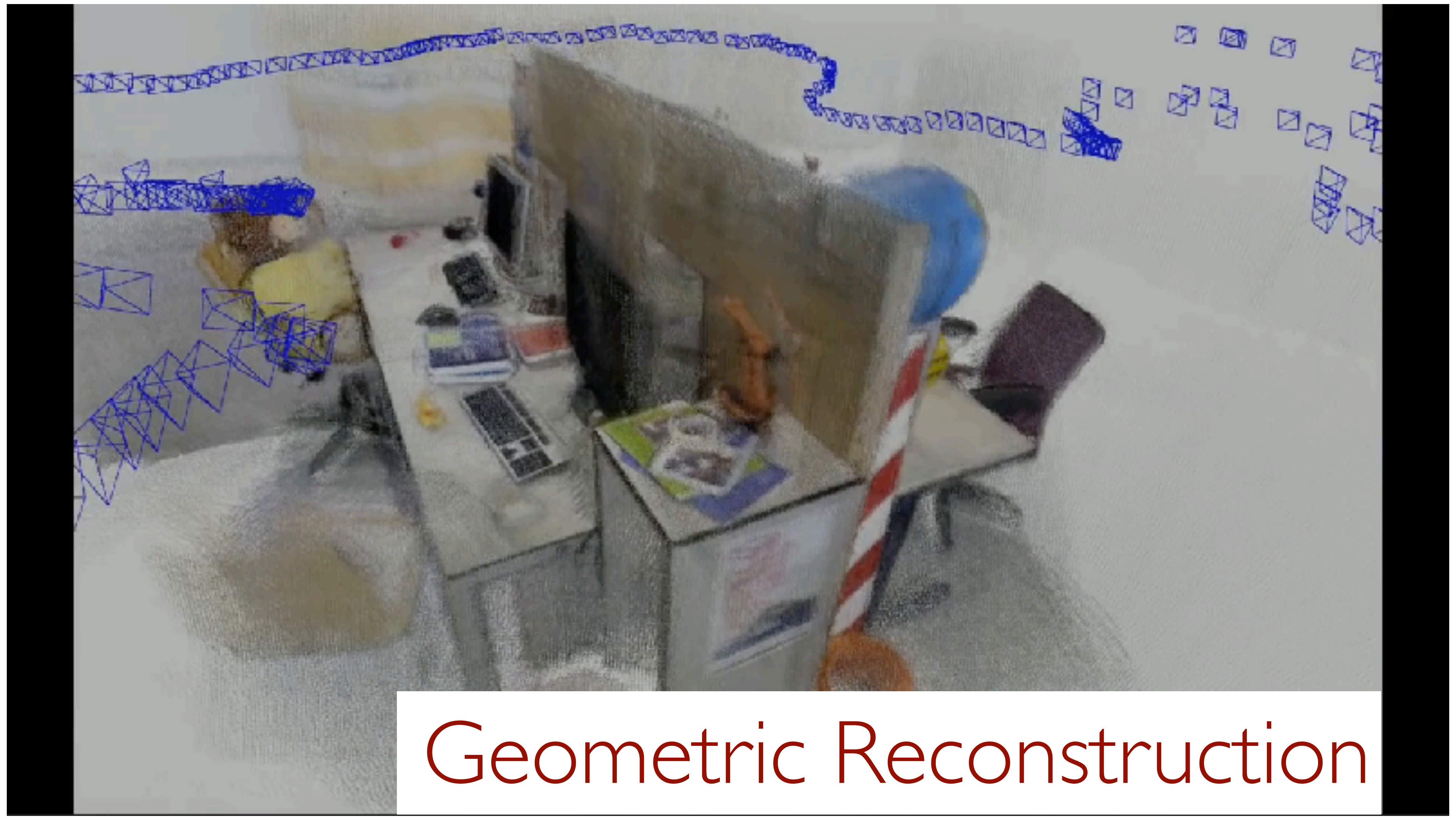
Discover paths or Explore

Classical Solution

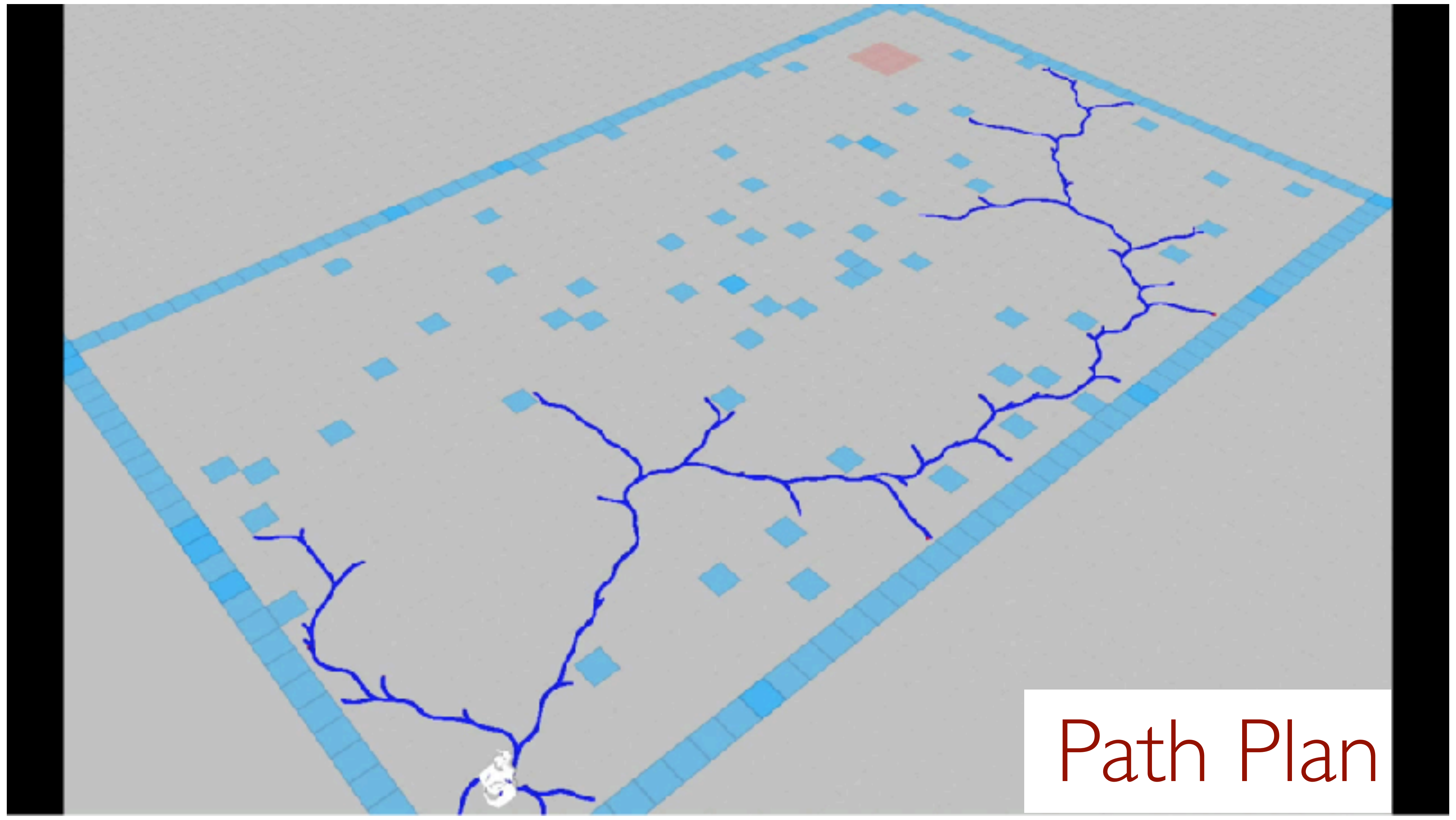


Observed Images

Mapping



Planning



Hartley and Zisserman. 2000. Multiple View Geometry in Computer Vision
Thrun, Burgard, Fox. 2005. Probabilistic Robotics

Canny. 1988. The complexity of robot motion planning.
Kavraki et al. RA1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces.
Lavalle and Kuffner. 2000. Rapidly-exploring random trees: Progress and prospects.

Video Credits: Mur-Artal et al., Palmieri et al.

Geometric 3D Reconstruction of the World

Unnecessary

Do we need to
tediously reconstruct
everything on this table?

Video Credit: Mur-Artal and Tardos, TRobotics 2016. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras.

Humans can do quite a bit without accurate metric 3D information

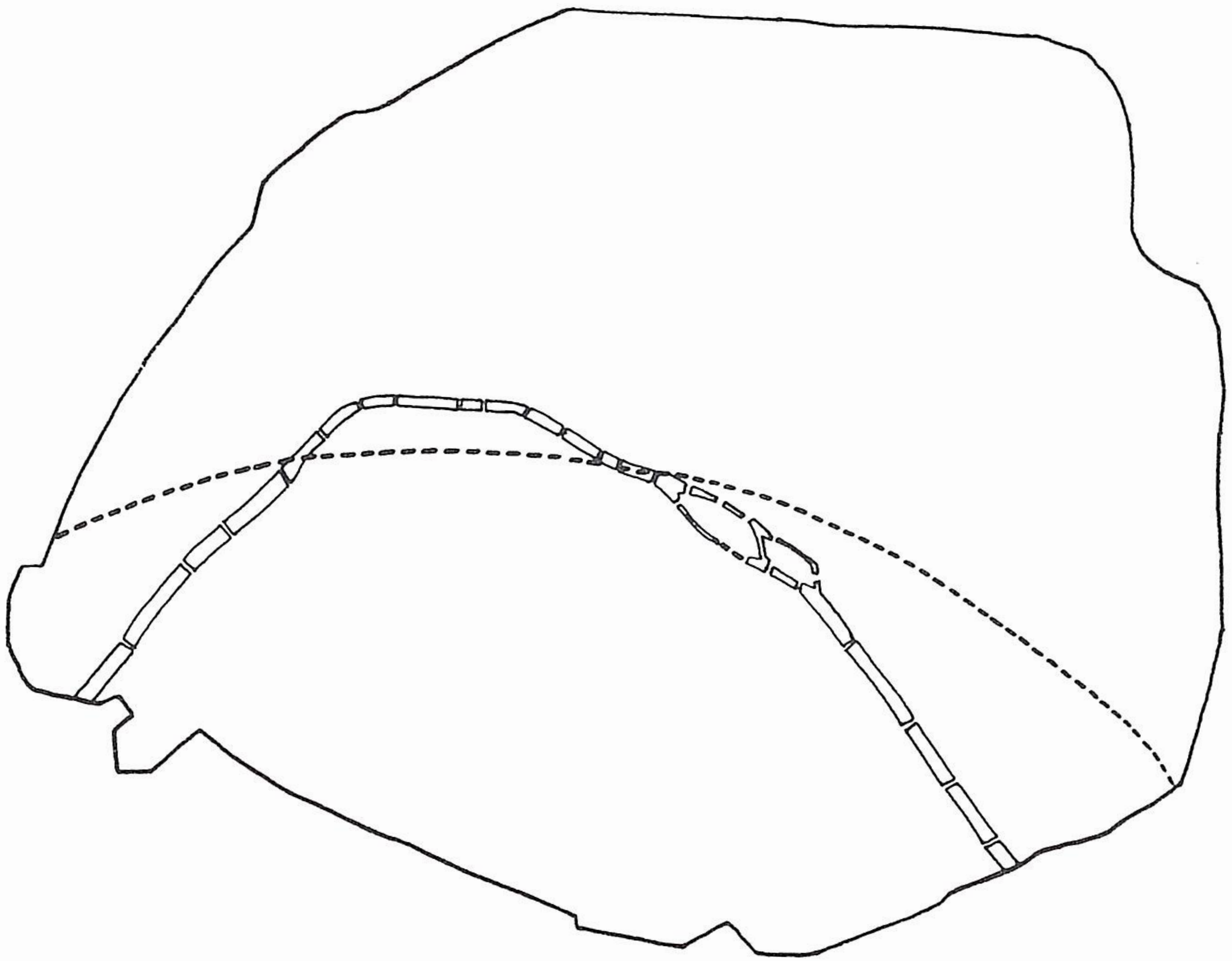


FIGURE 8.5
Perceived curvature of the Seine. The dotted line represents the median curvature imparted to the Seine in the subject's handdrawn maps. It is superimposed on the actual course of the river.

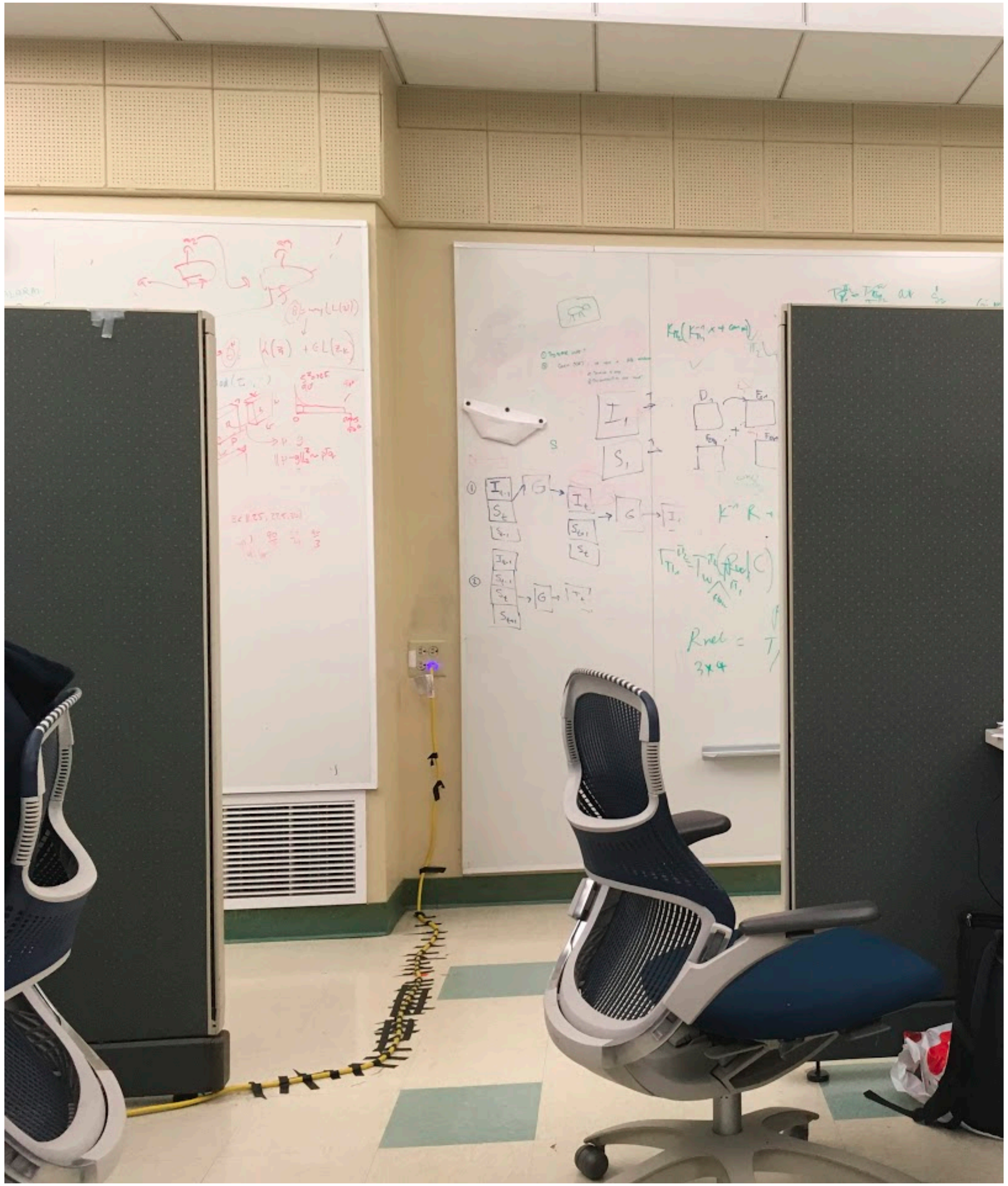
But there is a serious distortion in the way the Seine is represented. In reality the path of the Seine resembles a wave that enters Paris at the Quai Bercy, rises sharply northward, tapers slightly as it flows into separate streams around the islands, initiates its flat northernmost segment at the Place de la Concorde, then turns sharply in a great 60° bend at the Place d'Alma to flow out of the southwestern tip of the city. But in their drawings, 91.6 percent of the subjects understated the river's degree of curvature. Several subjects pulled it through the city as a straight line, and the typical subject represented the Seine as a gentle arc of slight but uniform curvature.

Perhaps, accurate full 3D is unnecessary?

Stanley Milgram. Psychological Maps of Paris. In Environmental Psychology 1976.

Lacks Semantics

Speculating about space not directly observed.



Lacks Semantics

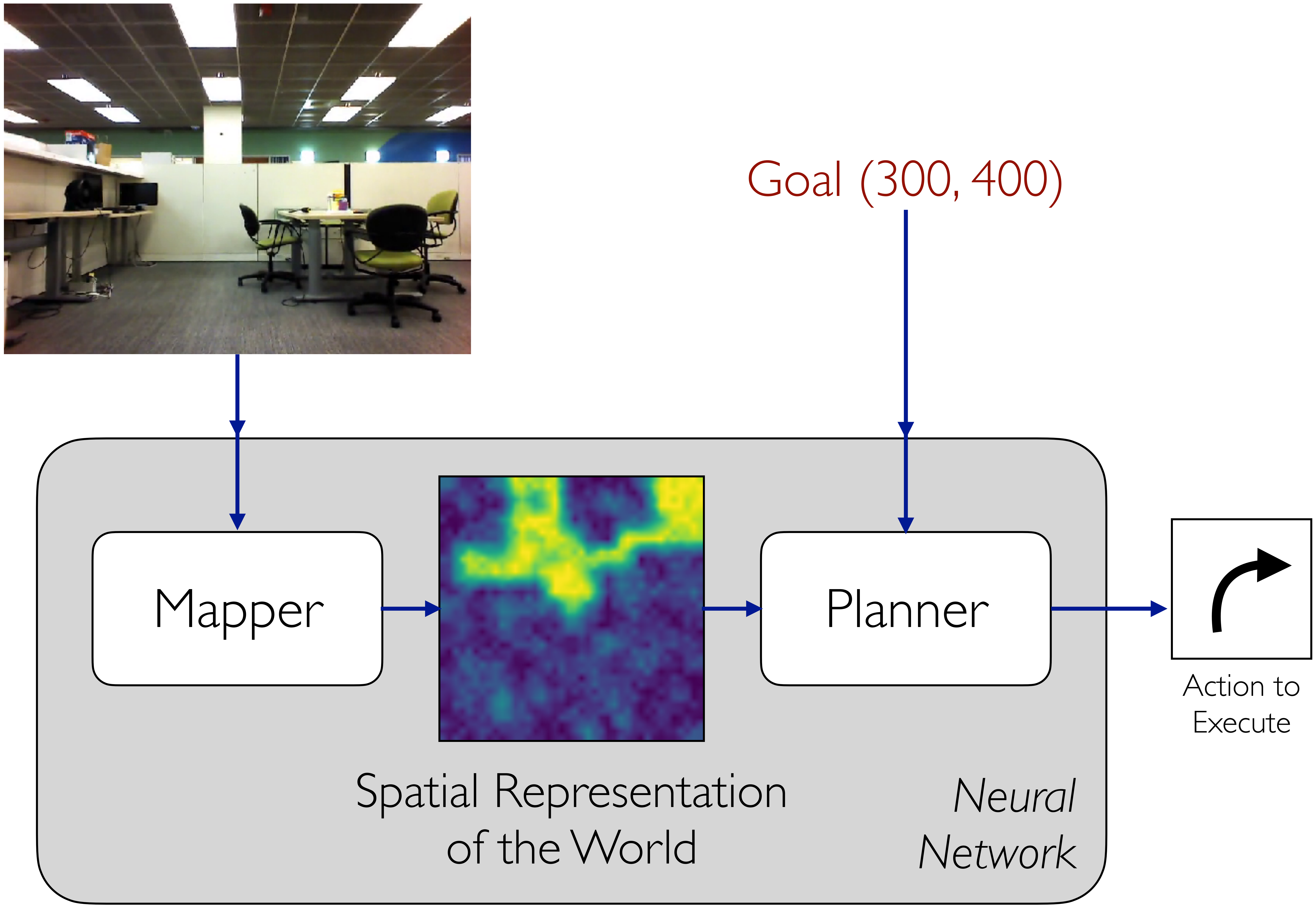
Eg: Finding a bathroom in a new restaurant

In this talk,

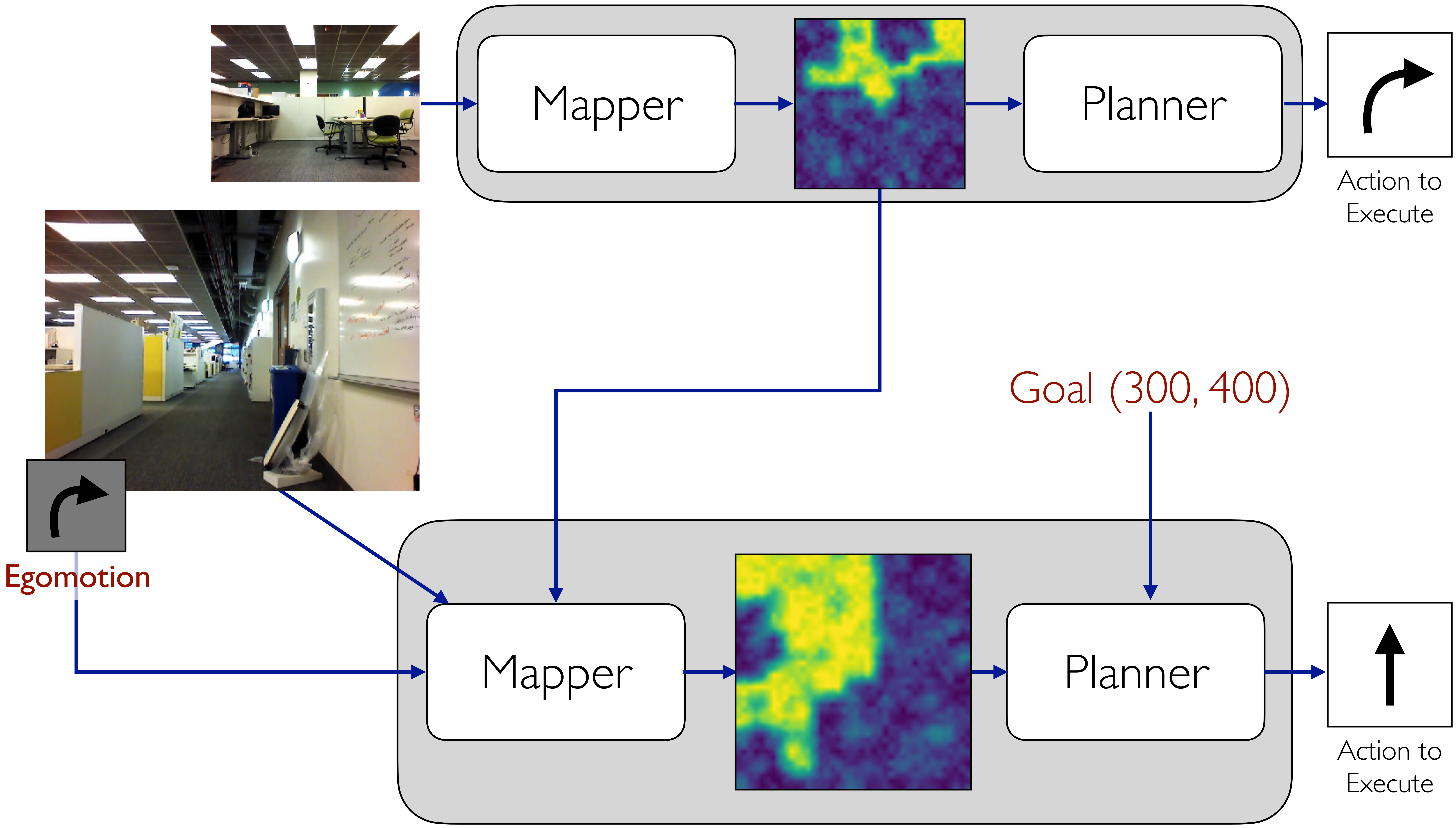
Representations for Places that Afford Navigation in Novel Environments

- *Augmenting metric representations with semantic reasoning*
- *Relaxing the need for metric representations*
- *Scaling-up training of such representations*

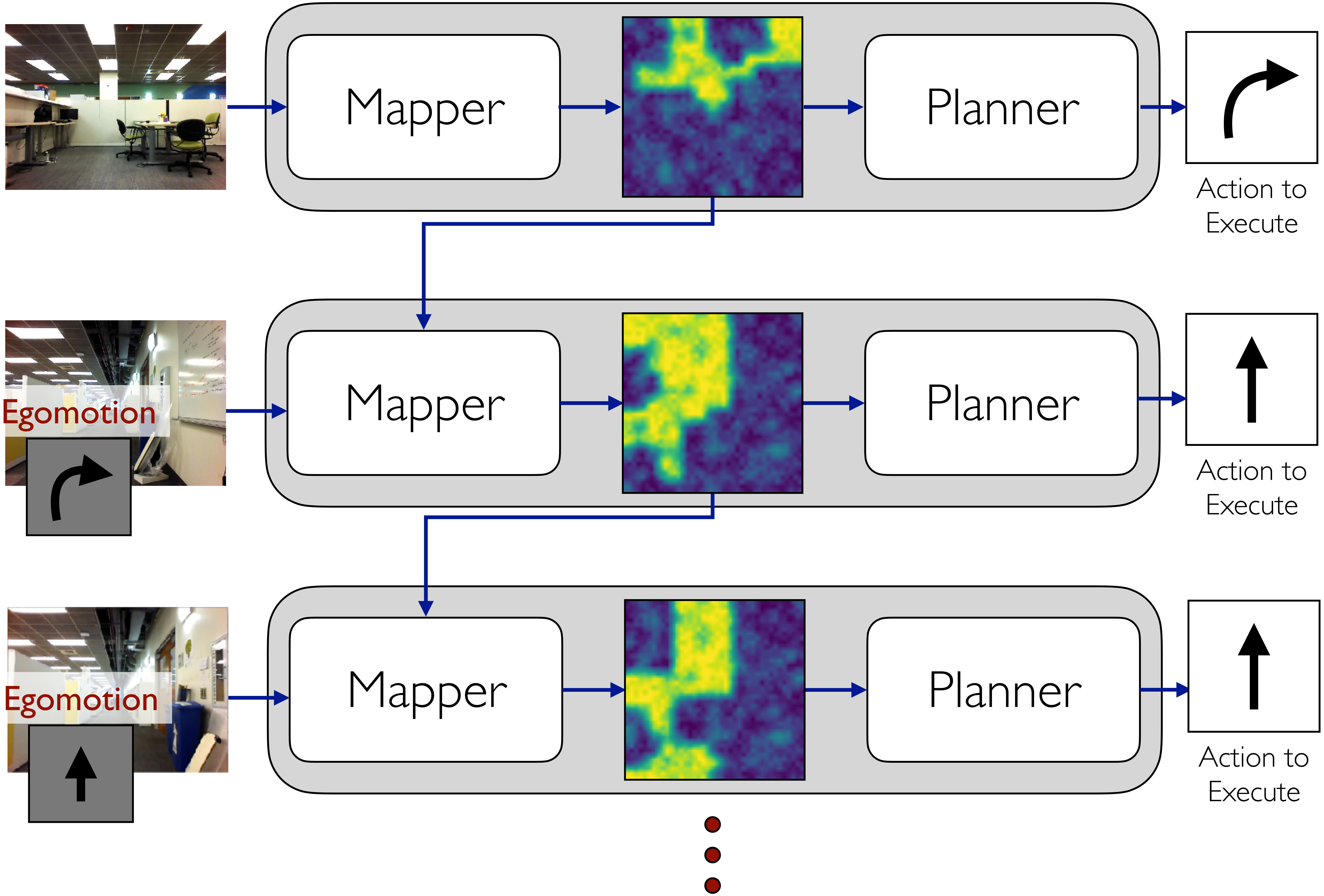
Operationalize insights from classical robotics into learning paradigms



S. Gupta et al., *CVPR 2017, IJCV 2020.* Cognitive Mapping and Planning for Visual Navigation

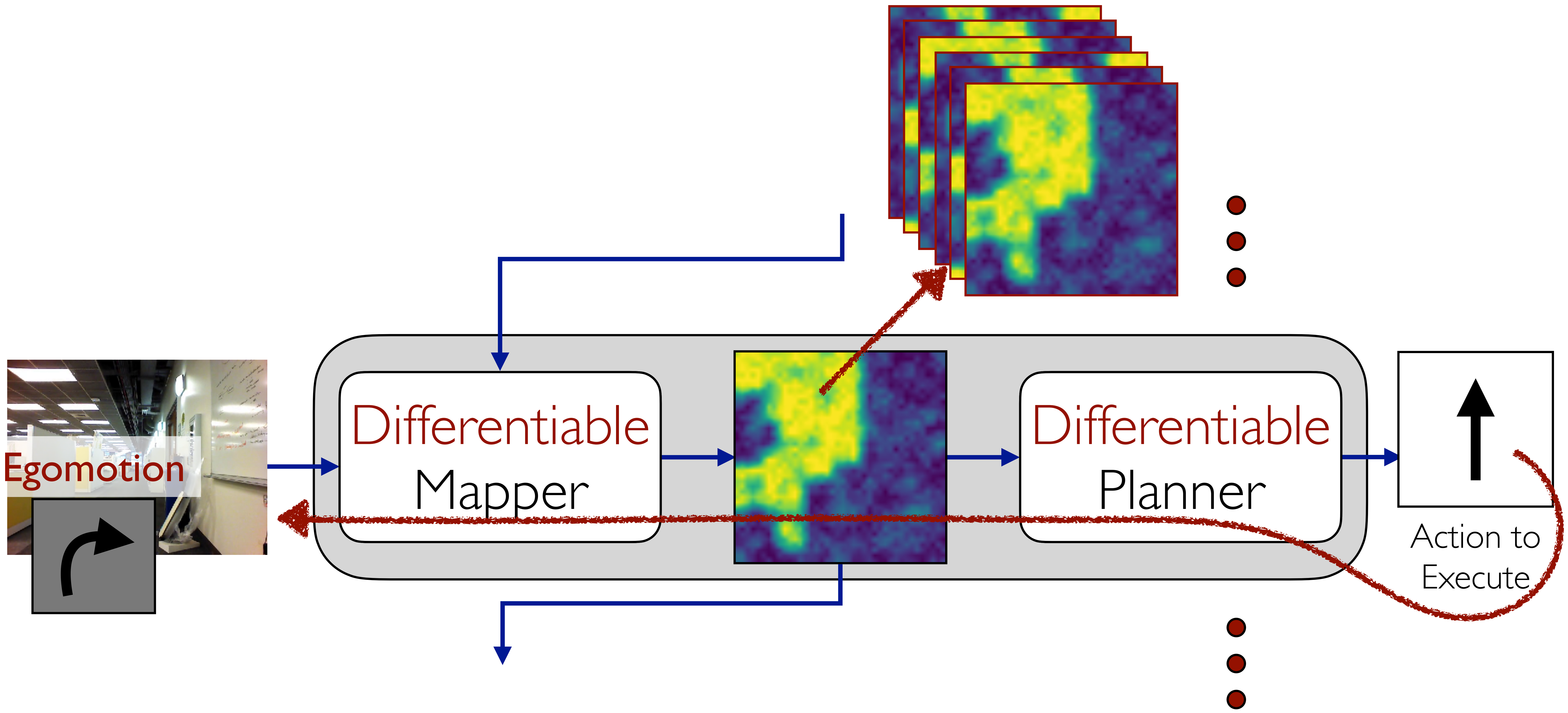


Mapper
Planner
Action to Execute
Goal (300, 400)
Egomotion
Mapper
Planner
Action to Execute

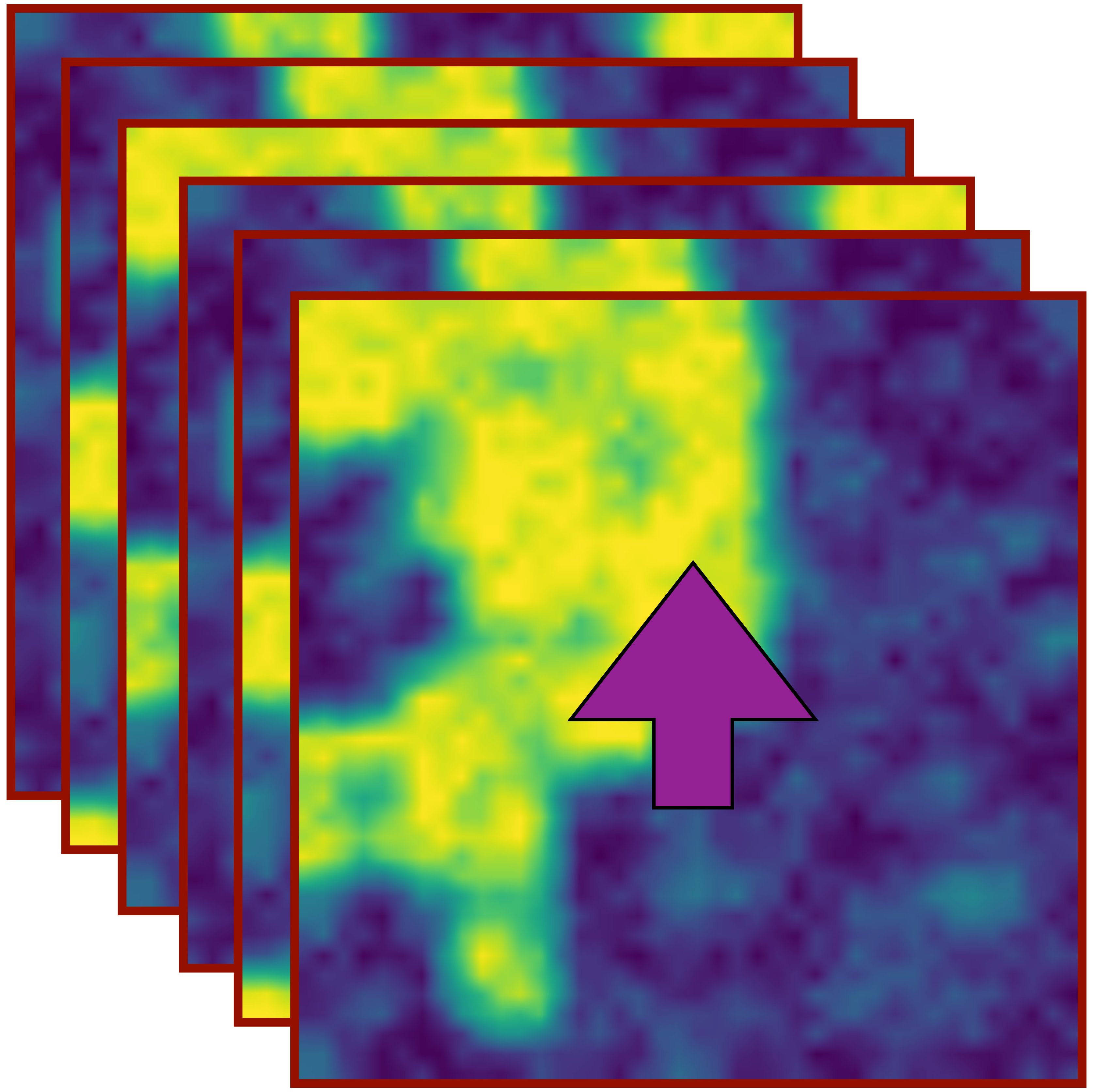


Mapper
Planner
Action to
Execute
Egomotion
Mapper
Planner
Action to
Execute
Egomotion
Mapper
Planner
Action to
Execute

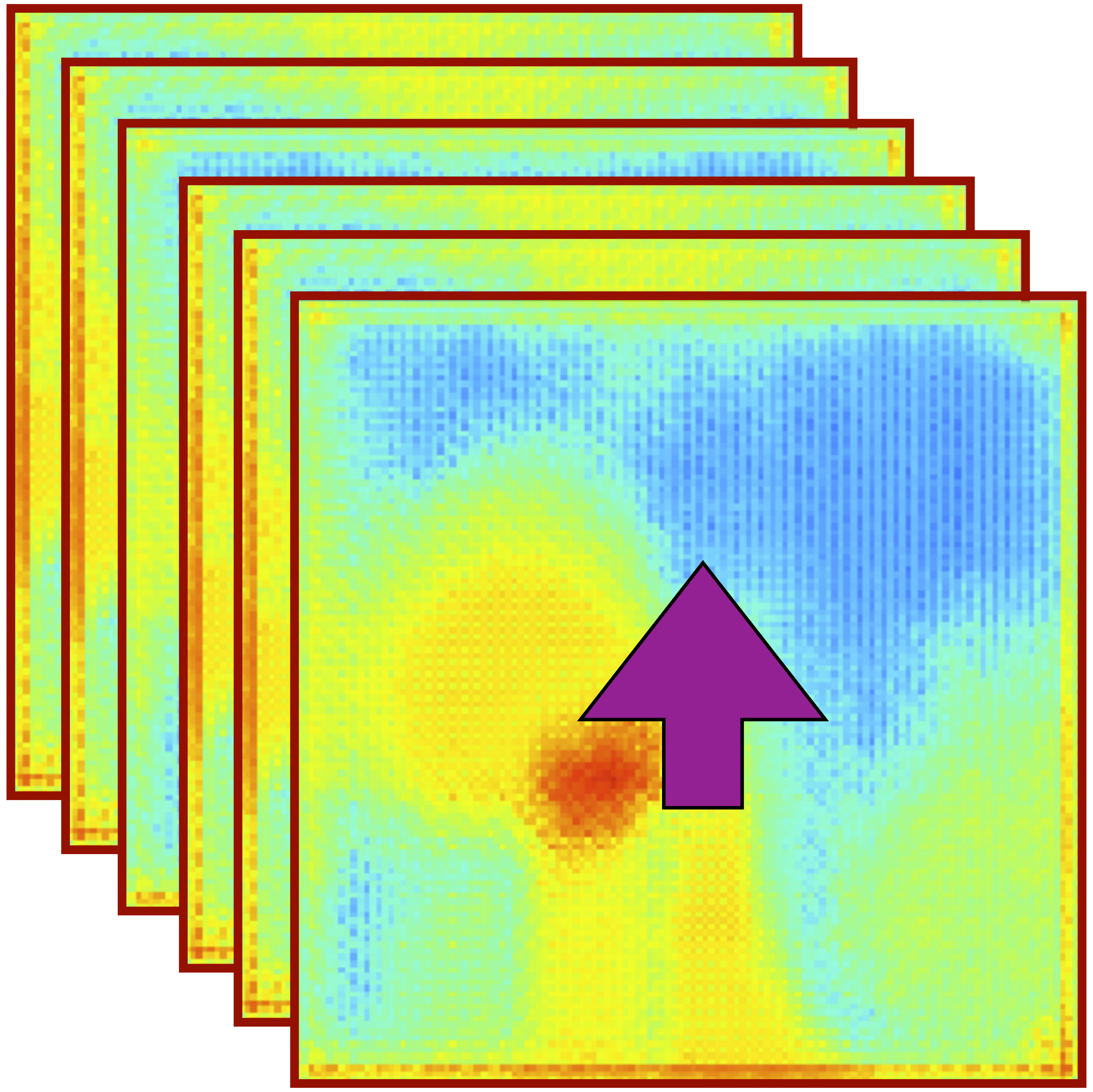
- Mapper and planner are *differentiable functions*
- Mapper and planner are *learned for end task*
- Hand-crafted obstacle maps to *task-driven semantic maps*



Spatial Representations



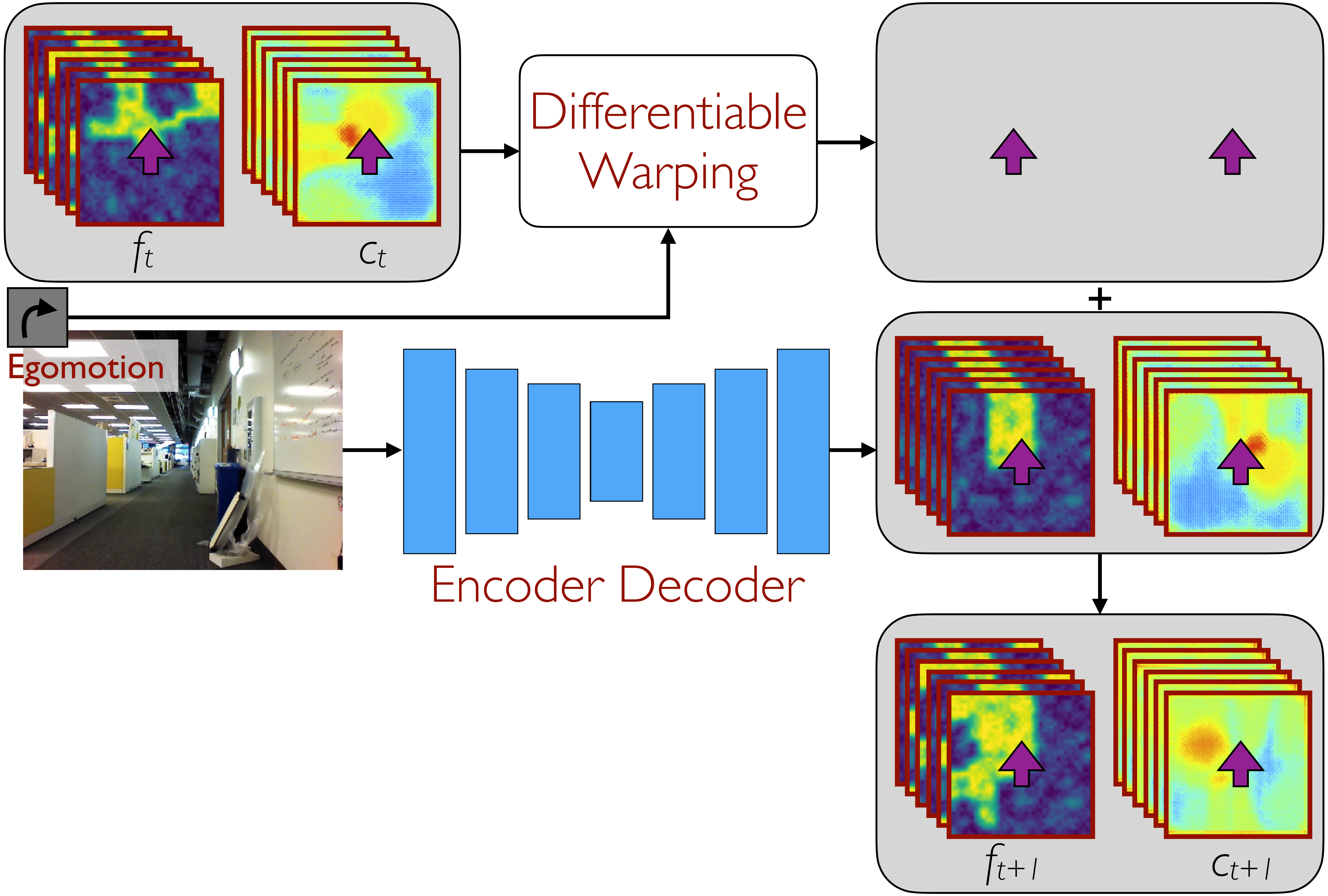
Feature f_t

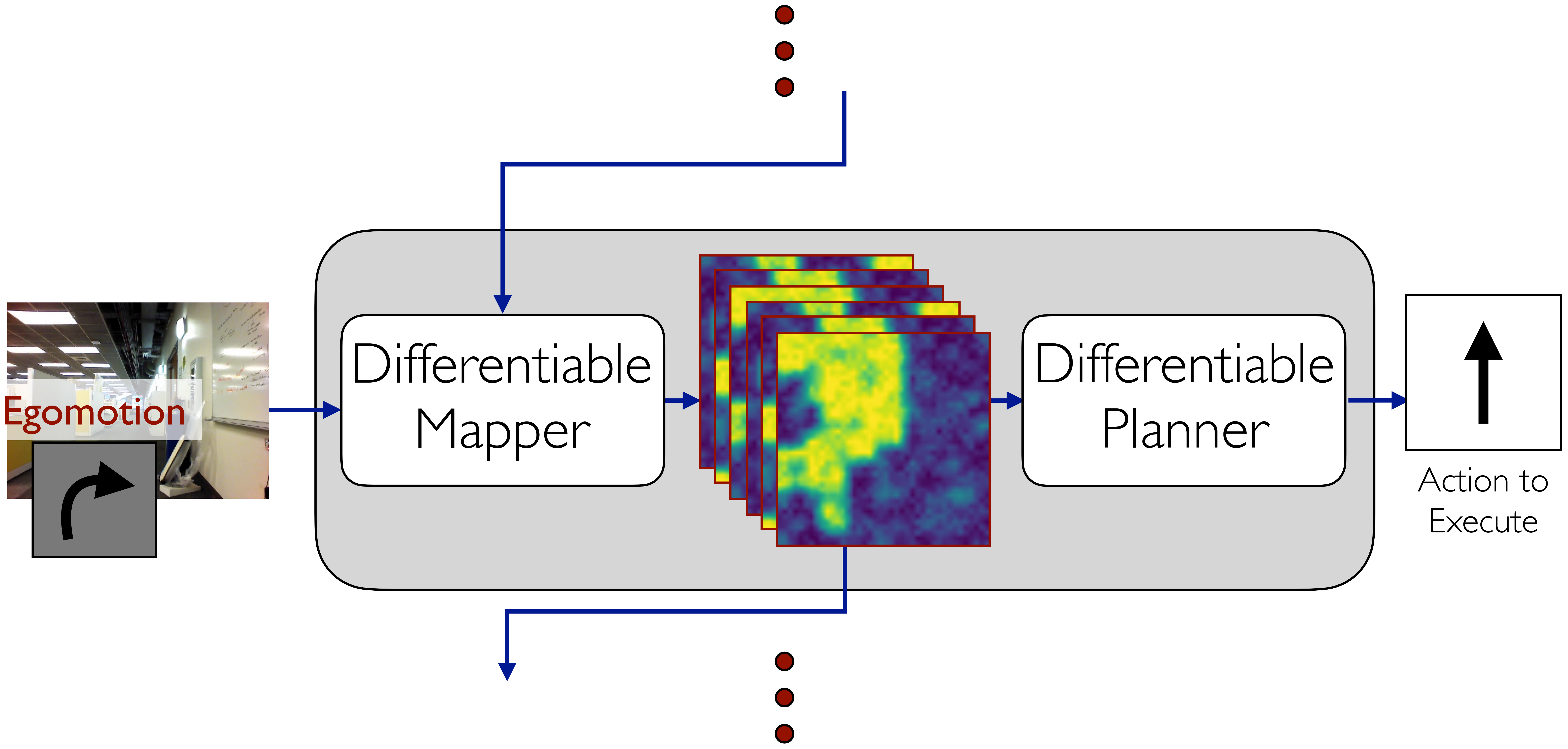


Confidence c_t

Egocentric Bird's Eye Coordinate Frame

Differentiable Mapper



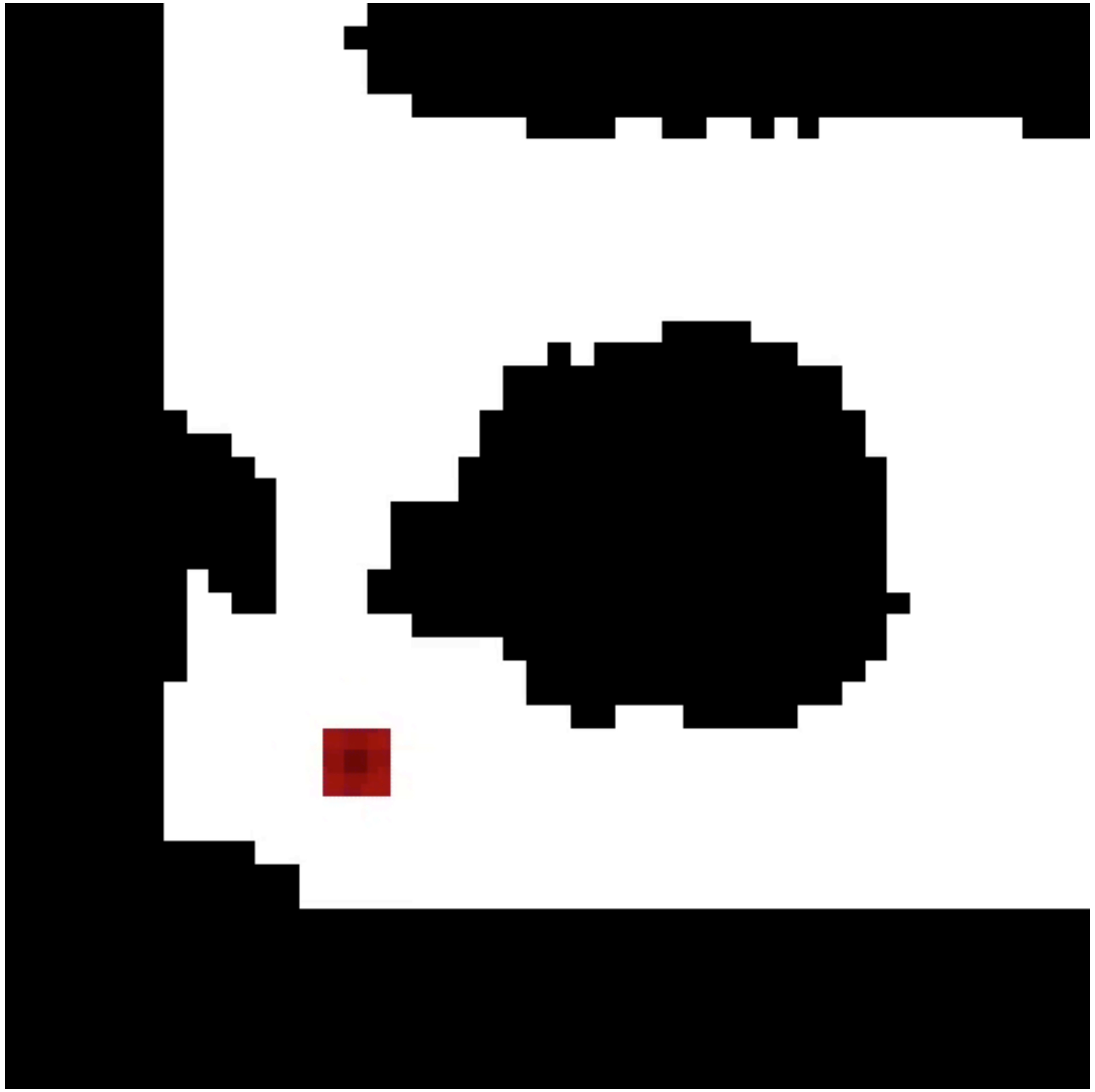


Egomotion
Differentiable Mapper
Differentiable Planner
Action to Execute

Differentiable Planner

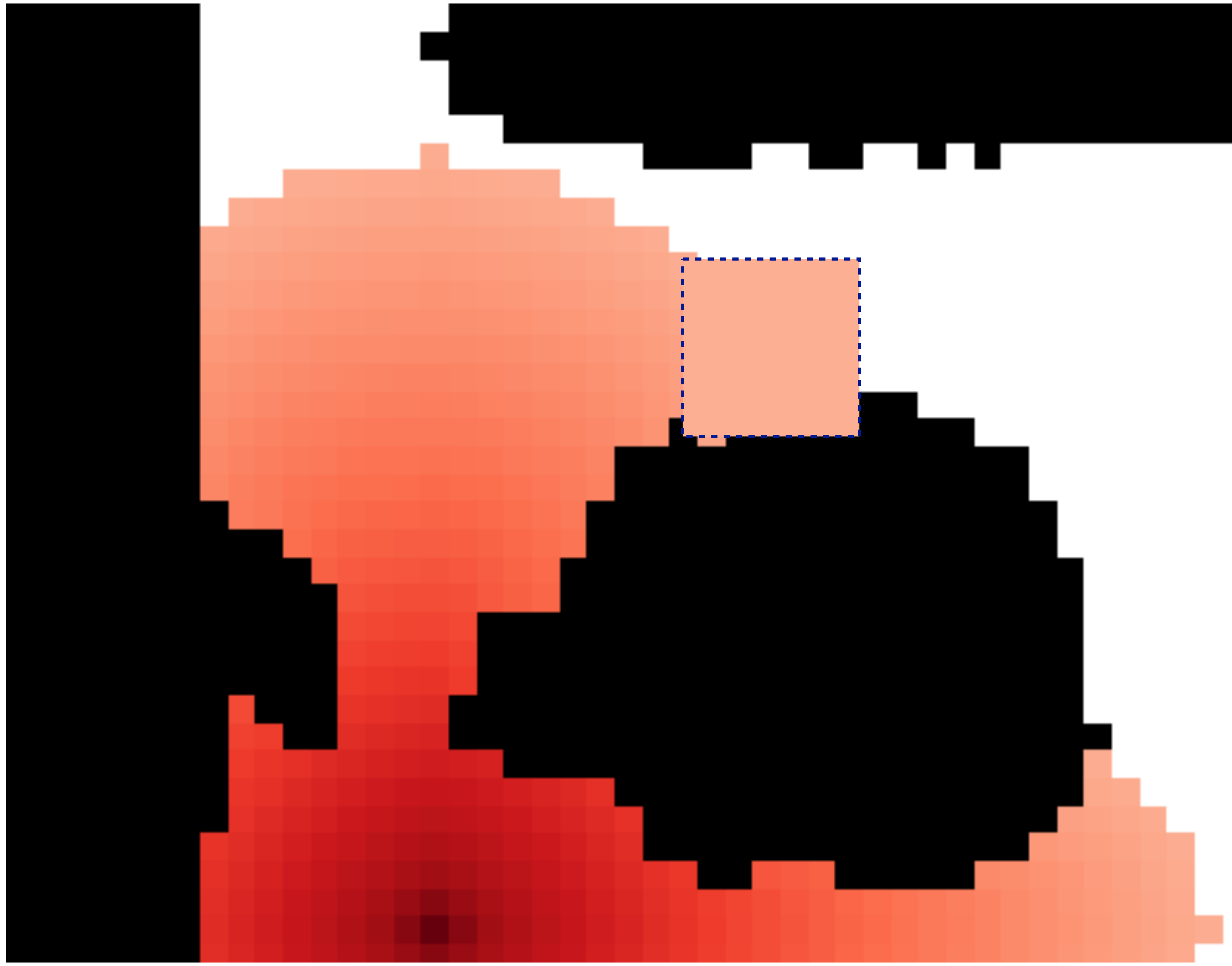
R. Bellman. 1957. A Markovian Decision Process.

Differentiable Planner



R. Bellman. 1957. A Markovian Decision Process.

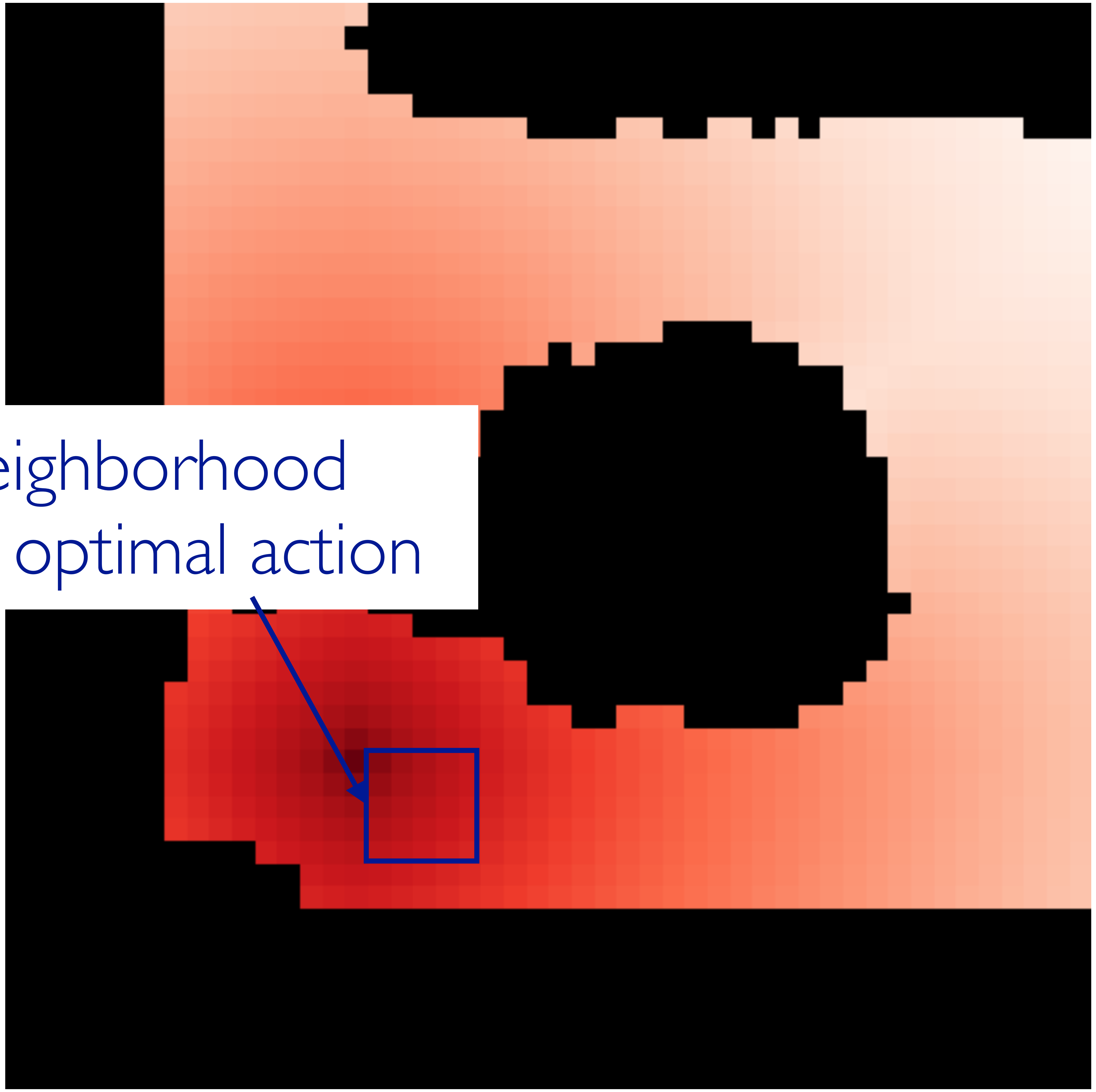
Differentiable Planner



Local computation that can be done using *Convolutions* and *Channel-wise Max-Pooling.*

R. Bellman. 1957. A Markovian Decision Process.
Tamar et al. NIPS 2016. Value Iteration Networks.

Differentiable Planner



R. Bellman. 1957. A Markovian Decision Process.

Policy Training

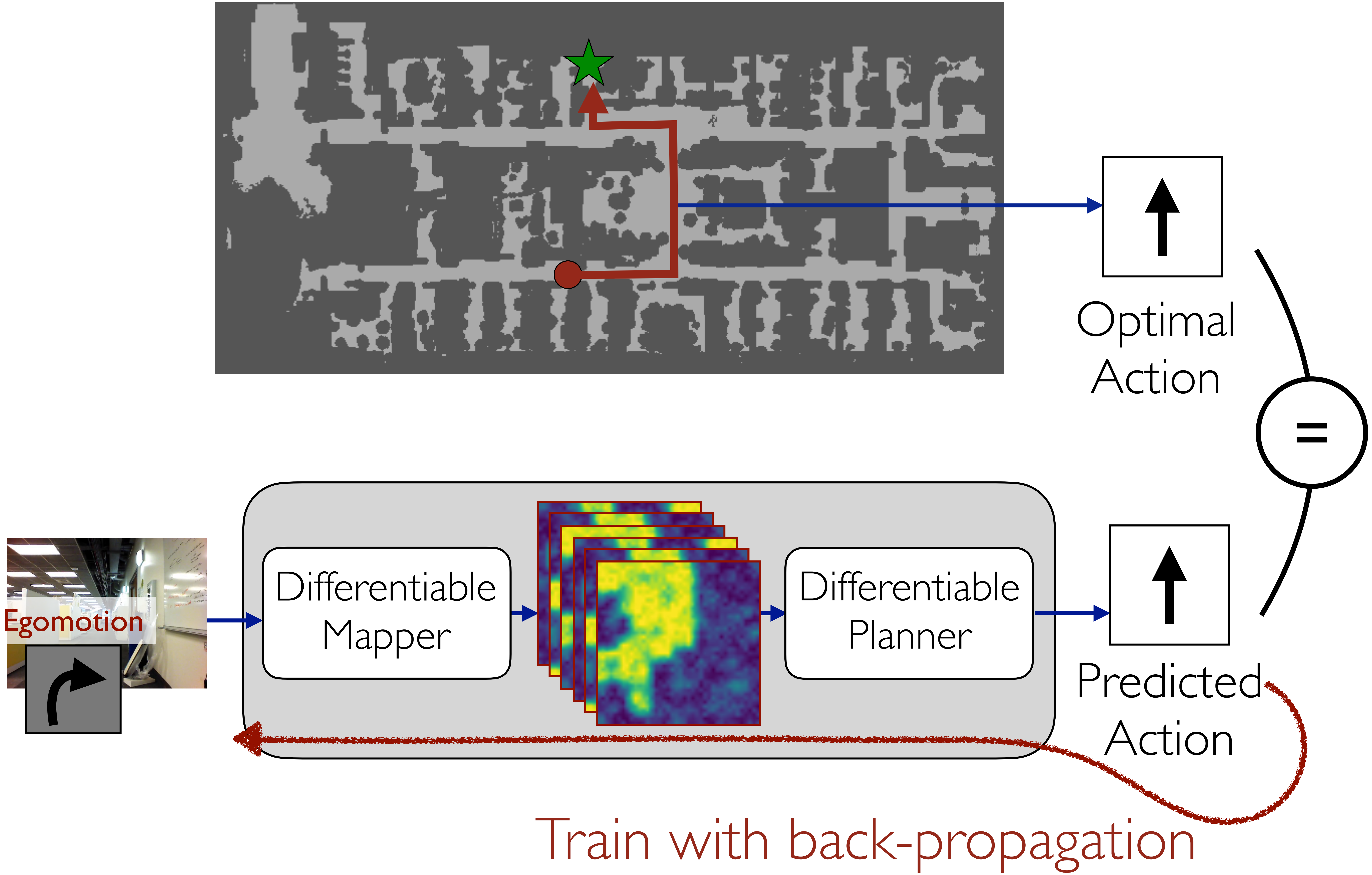
Simulator based on scans of *Real World Environments*

Simulate robot views and motion

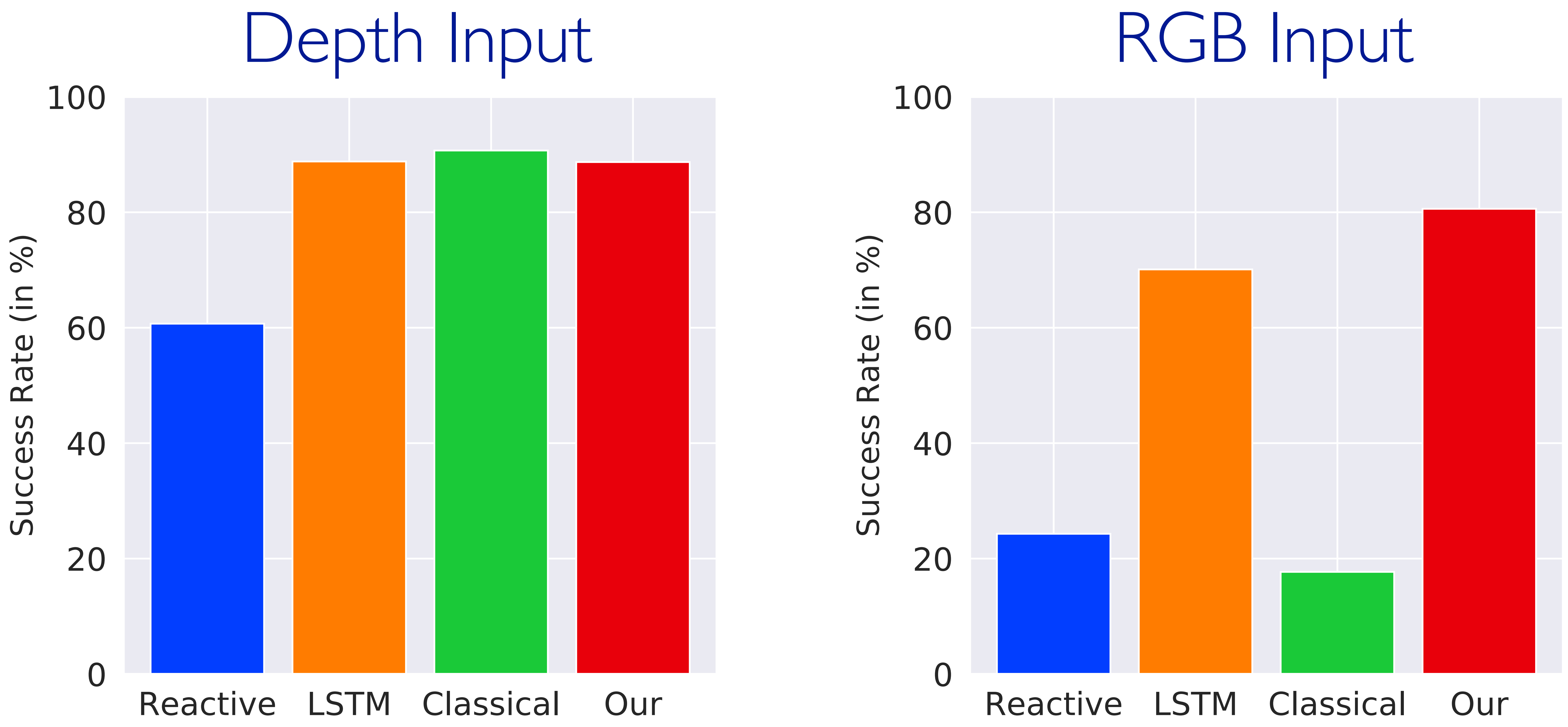
Compute ground truth traversability

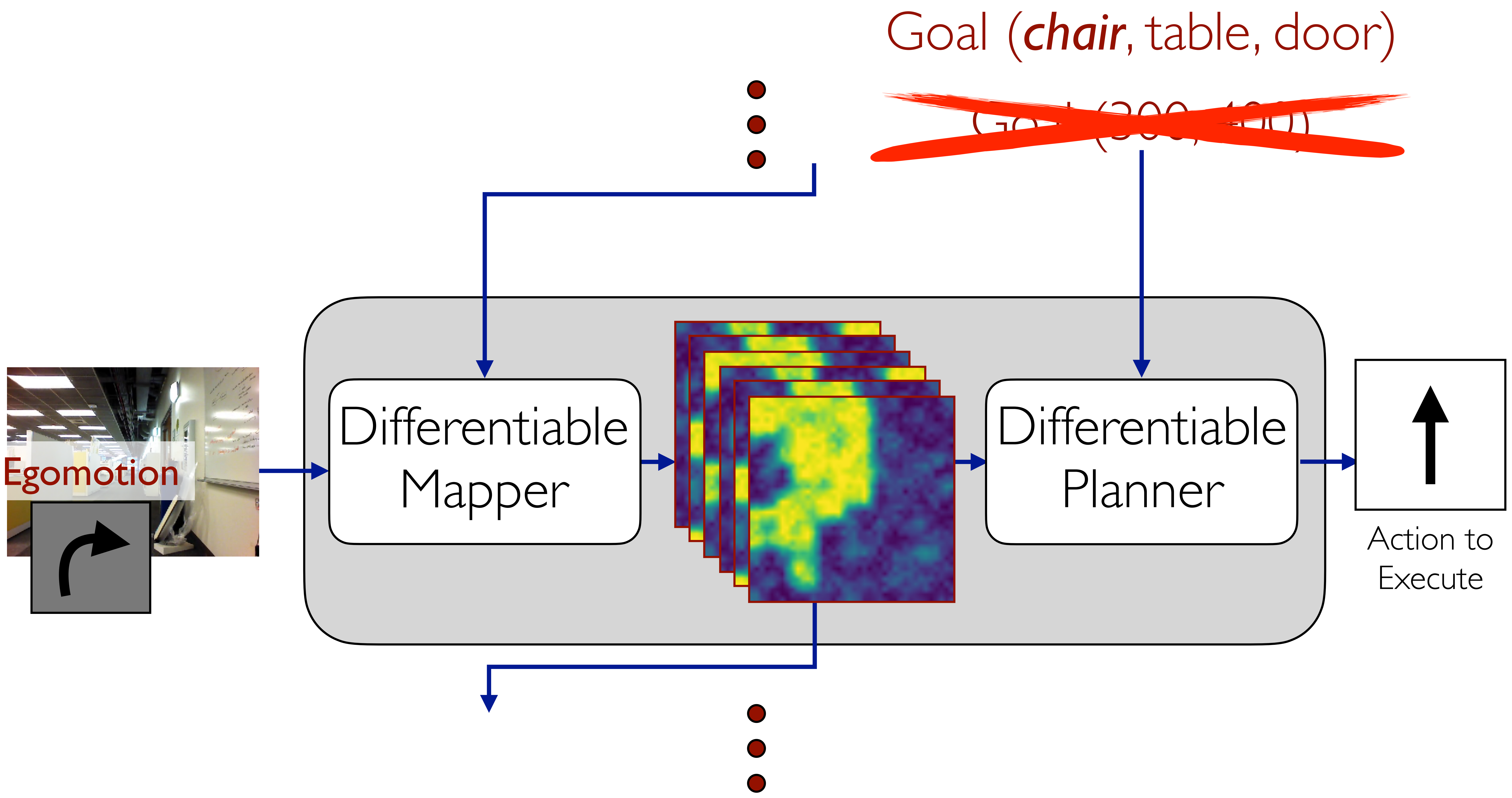
Gupta et al., CVPR 2017. Cognitive Mapping and Planning for Visual Navigation
Armeni et al. CVPR 2016. 3D Semantic Parsing of Large-Scale Indoor Spaces

Policy Training by Expert Imitation



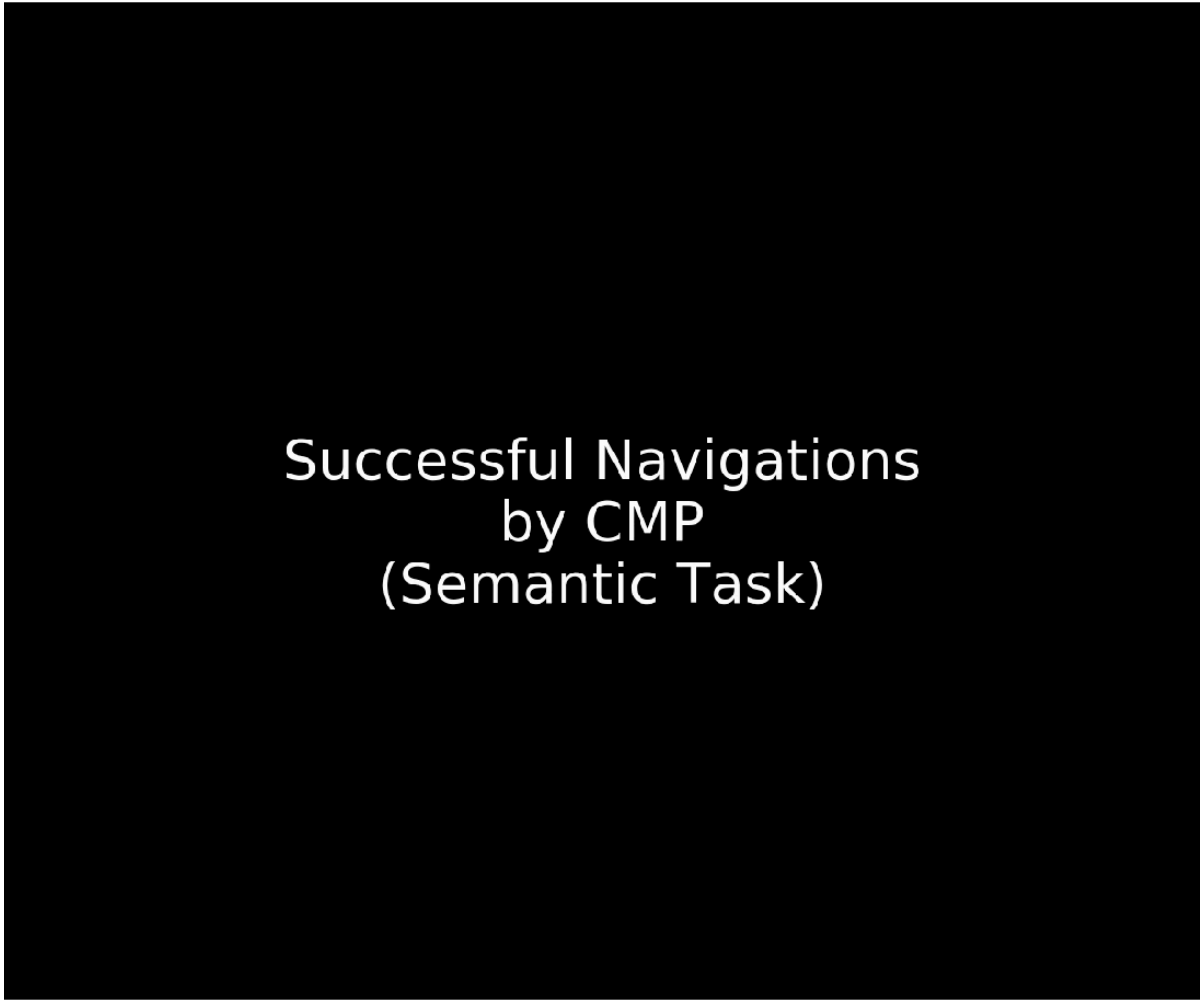
Results (Novel Env., Go To Relative Offset)



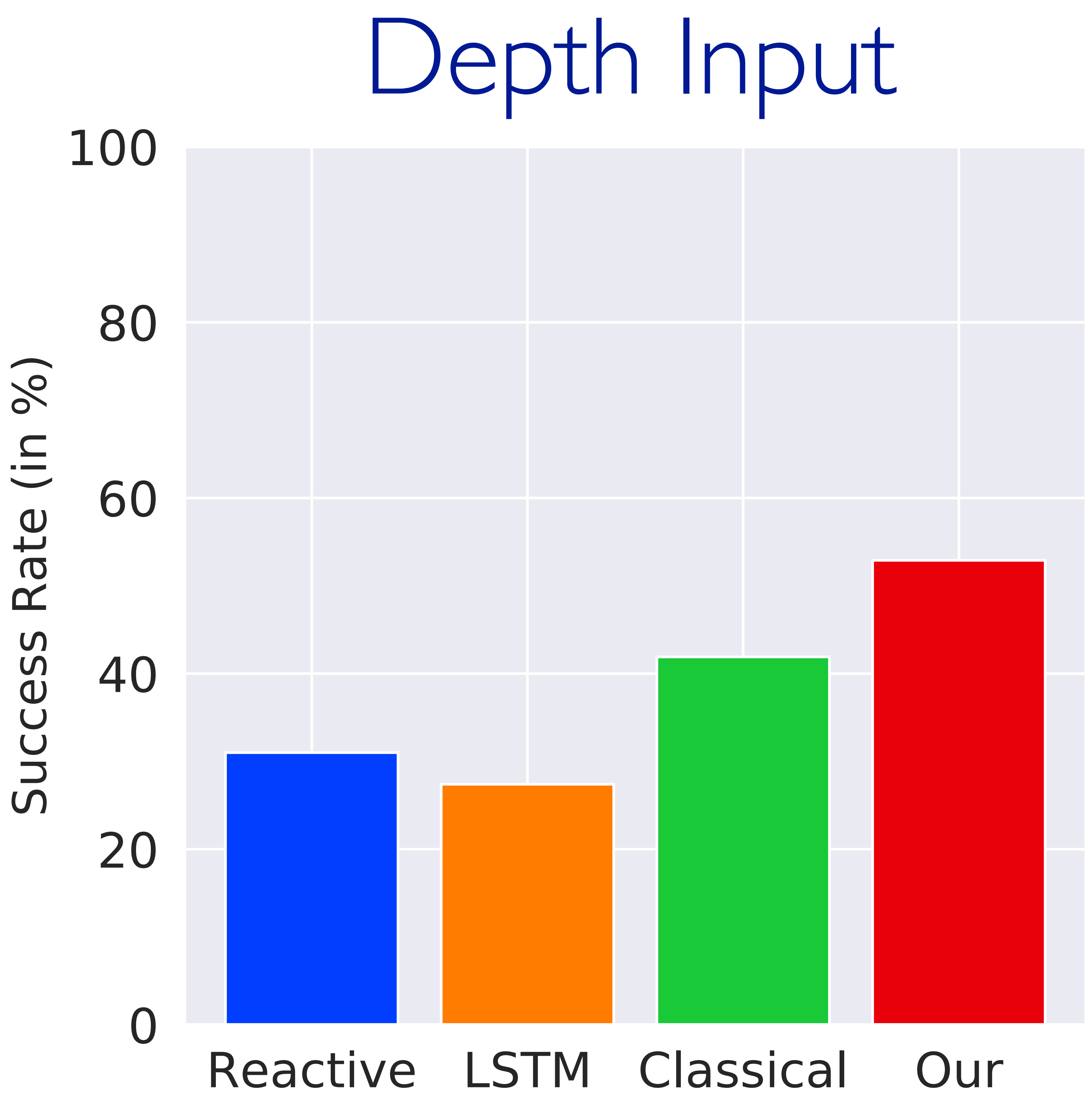


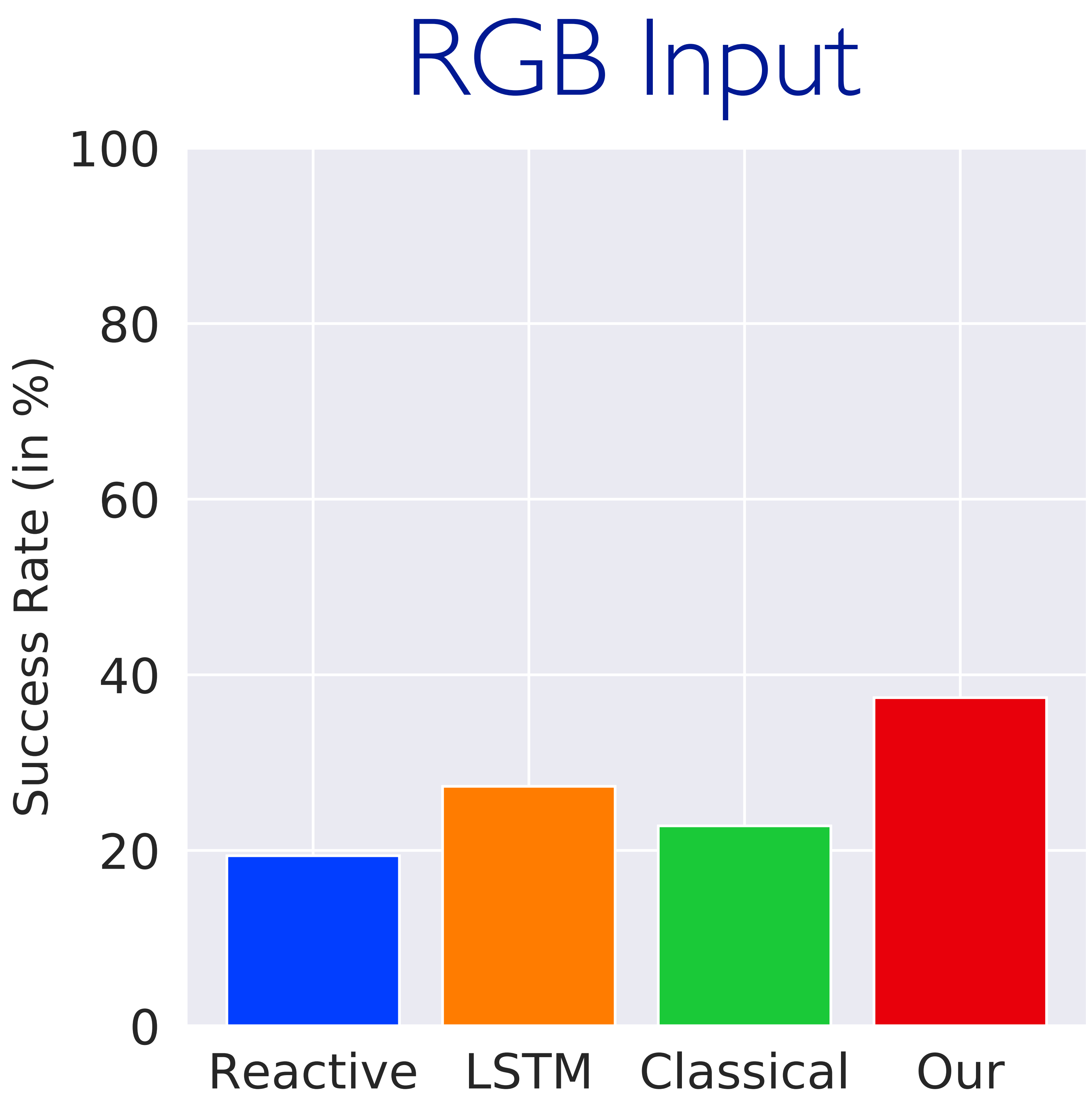
Goal (*chair*, table, door)
Egomotion
Differentiable Mapper
Differentiable Planner
Action to Execute

Semantic Tasks (Go to a chair)

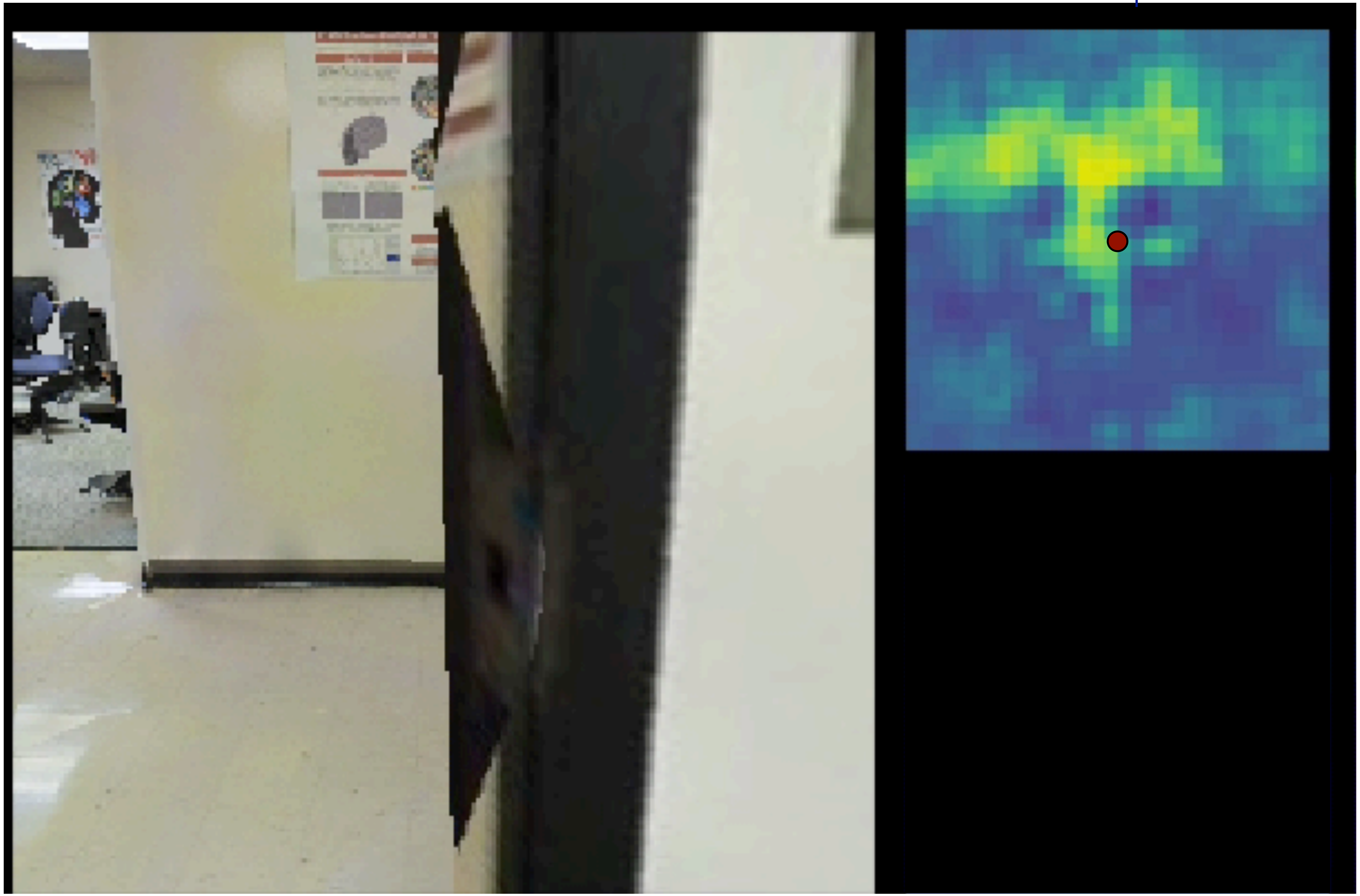


Results (Novel Env, Go To Object)



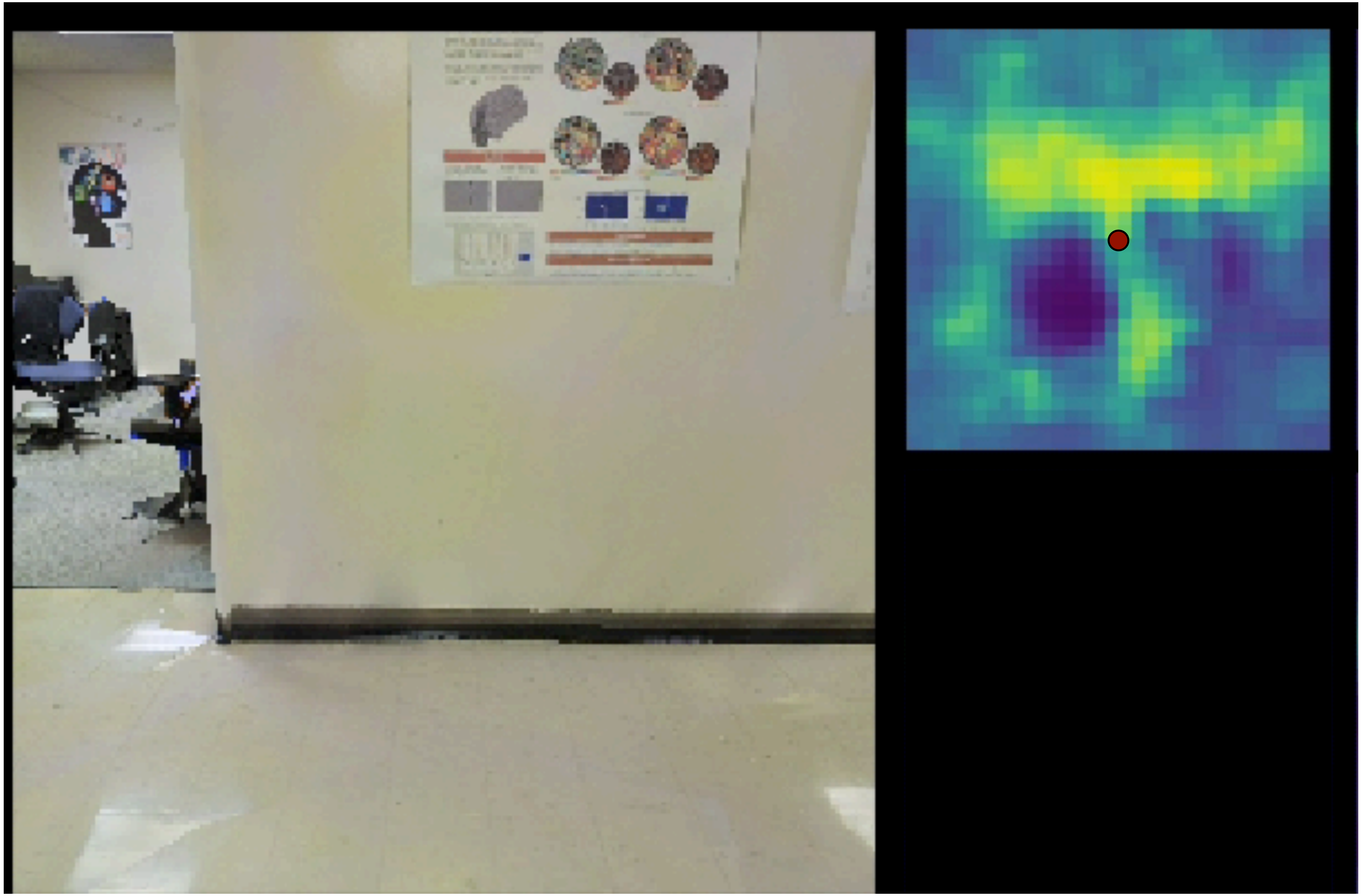


Agent can make predictions about its surroundings



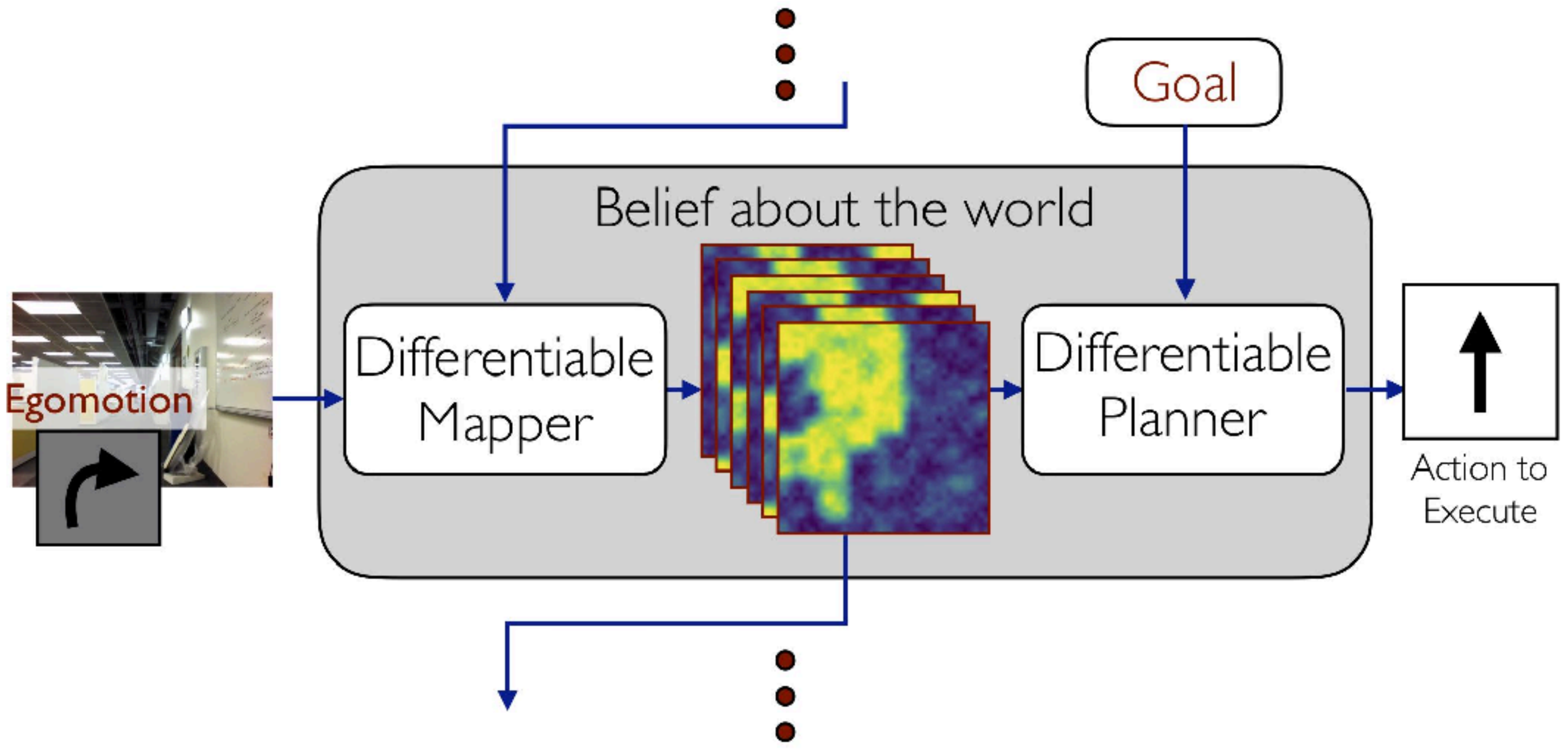
Agent can make predictions about its surroundings

Free Space



Agent can make predictions about its surroundings

Representation for Places



- Spatial reasoning
- Semantic reasoning

- Sensitive to pose error
- Interactive training (DAgger easier than RL, but still)
- Long training horizons

Can we relax the need for spatially consistent global maps?

Reaching Image Goals

D. Chaplot, R. Salakhutdinov, A. Gupta, S. Gupta. Neural topological slam for visual navigation. In *CVPR*, 2020.

Image Goal Task

- Agent observations are panoramic images
- Take actions to navigate to the goal location
- Take the 'stop' action at the goal location
- *Actuation noise, robot does not precisely know how much it has moved*

Representation

4 Functions

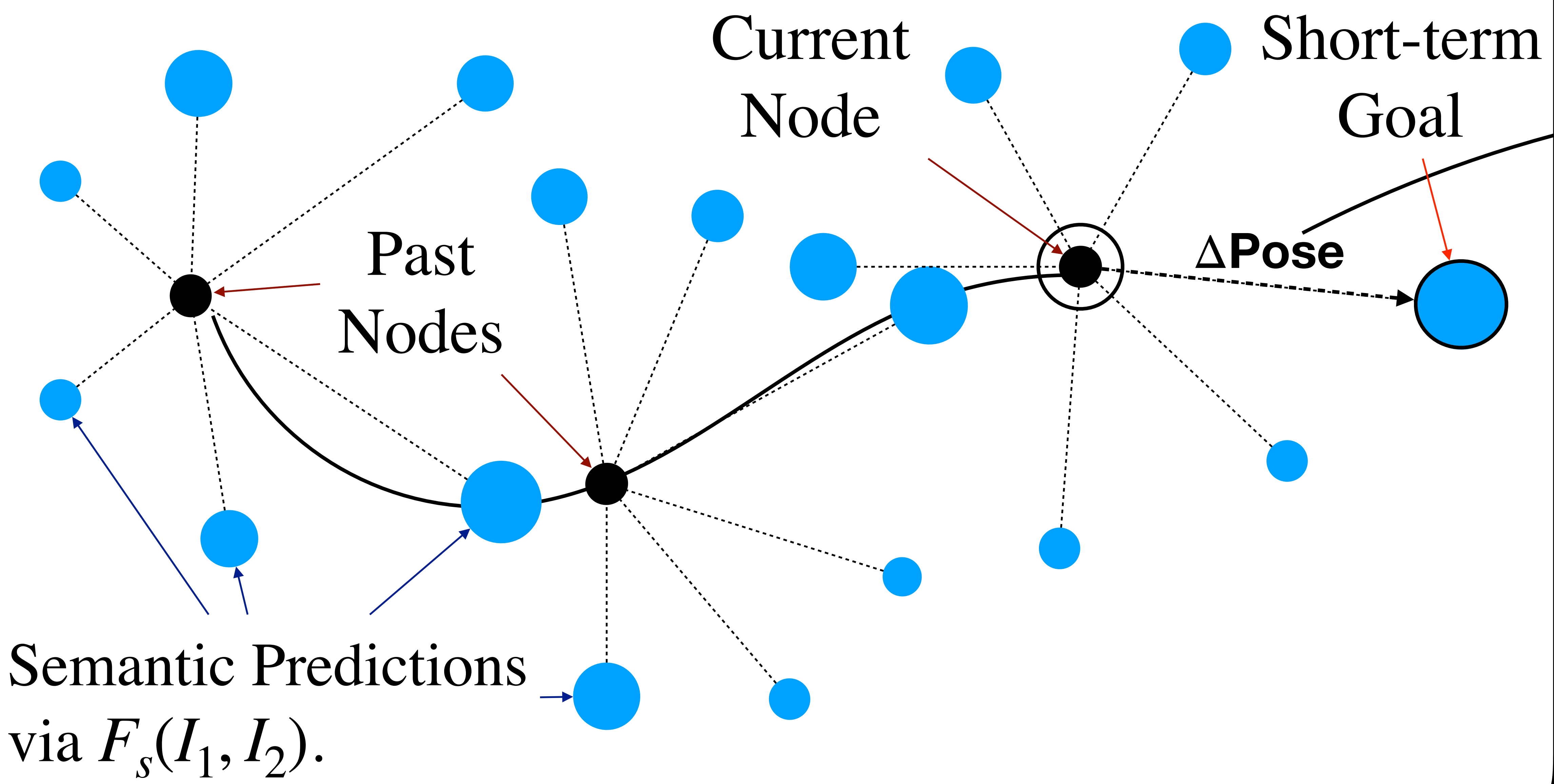
- $F_G(I_1)$ = Geometric Prediction: Free directions
- $F_S(I_1, I_2)$ = Semantic Prediction: Closeness to target
- $F_R(I_1, I_2)$ = Relative Pose
- $F_L(I_1, I_2)$ = Localization

Using the Representation

Hierarchical Policy

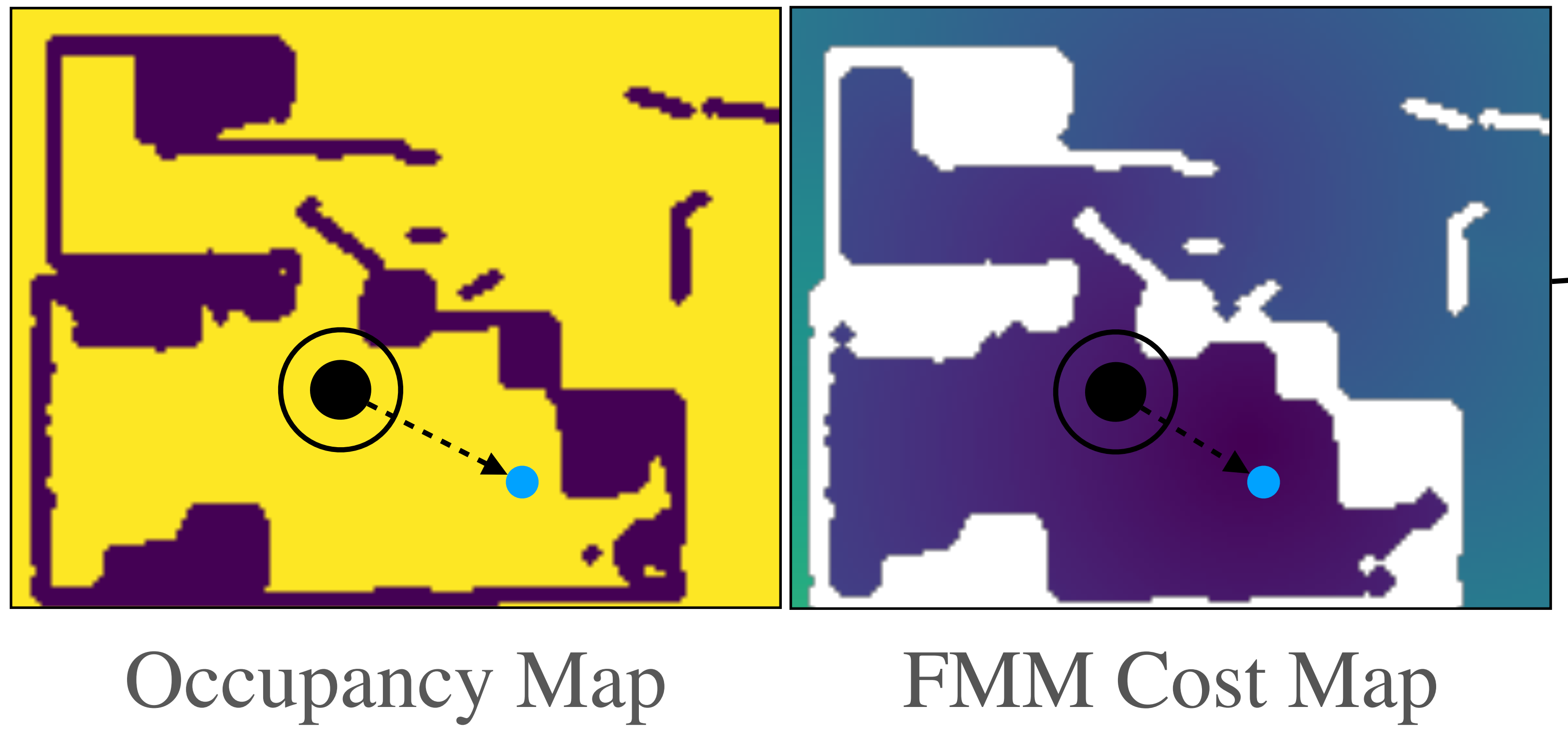
High-Level Policy

- Decides where to go next and emits short-term goal
- Builds a topological map that keeps track of visited nodes, ghost nodes and values predicted by $F_s(I_1, I_2)$ for different directions

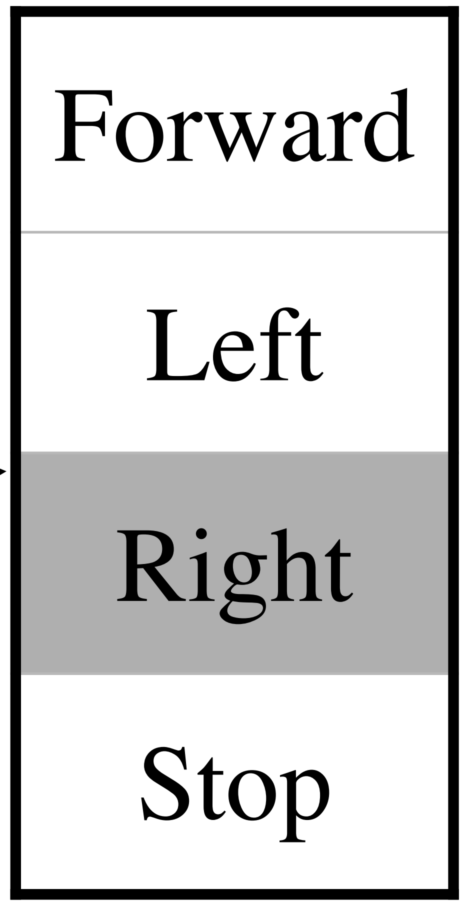


Low-Level Policy

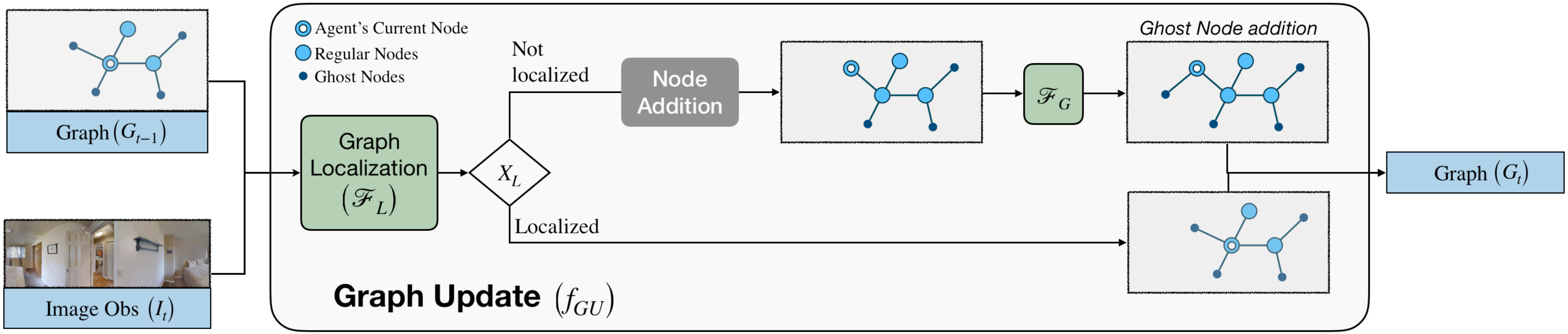
- Executes actions to achieve short-term goal
- Option 1: predict local occupancy, plan paths



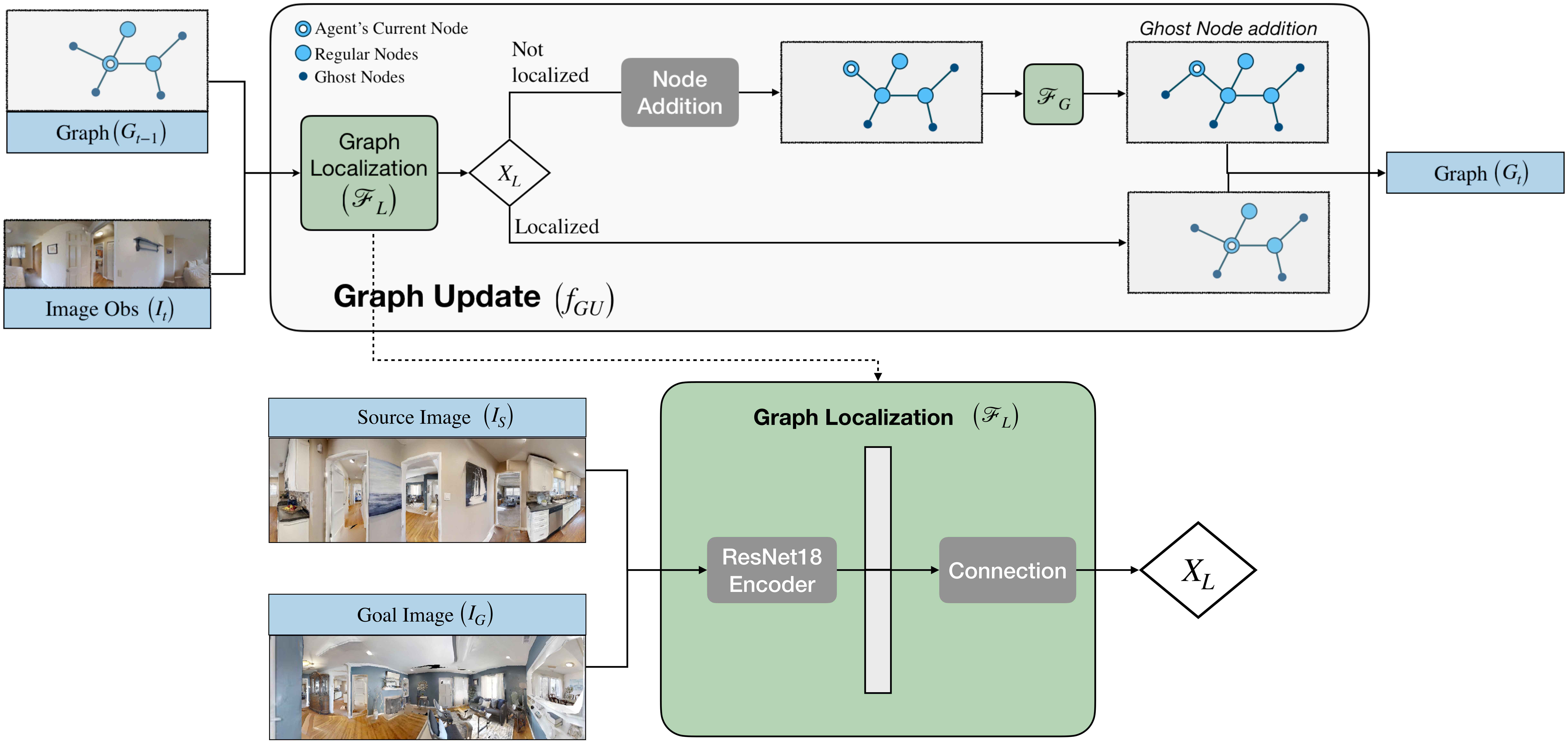
- Option 2: learn a low-level controller



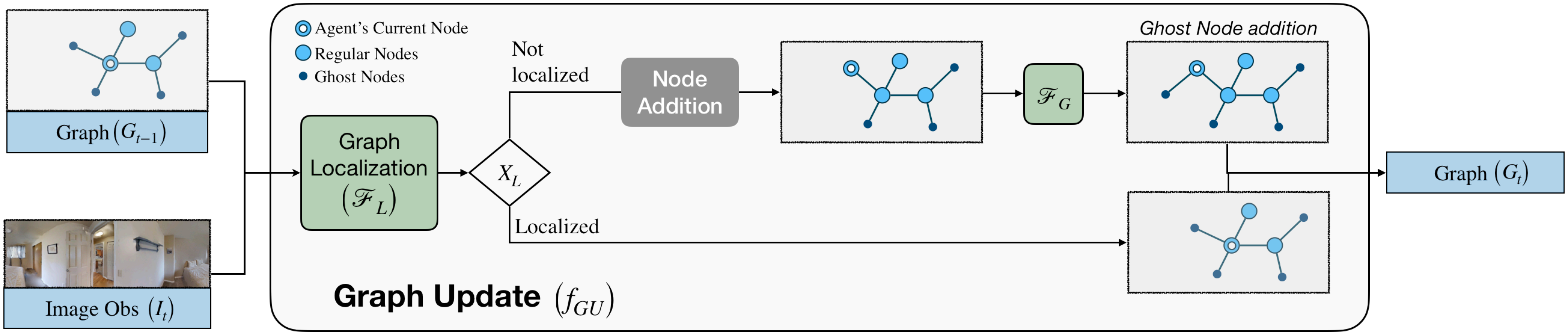
Building the Representation



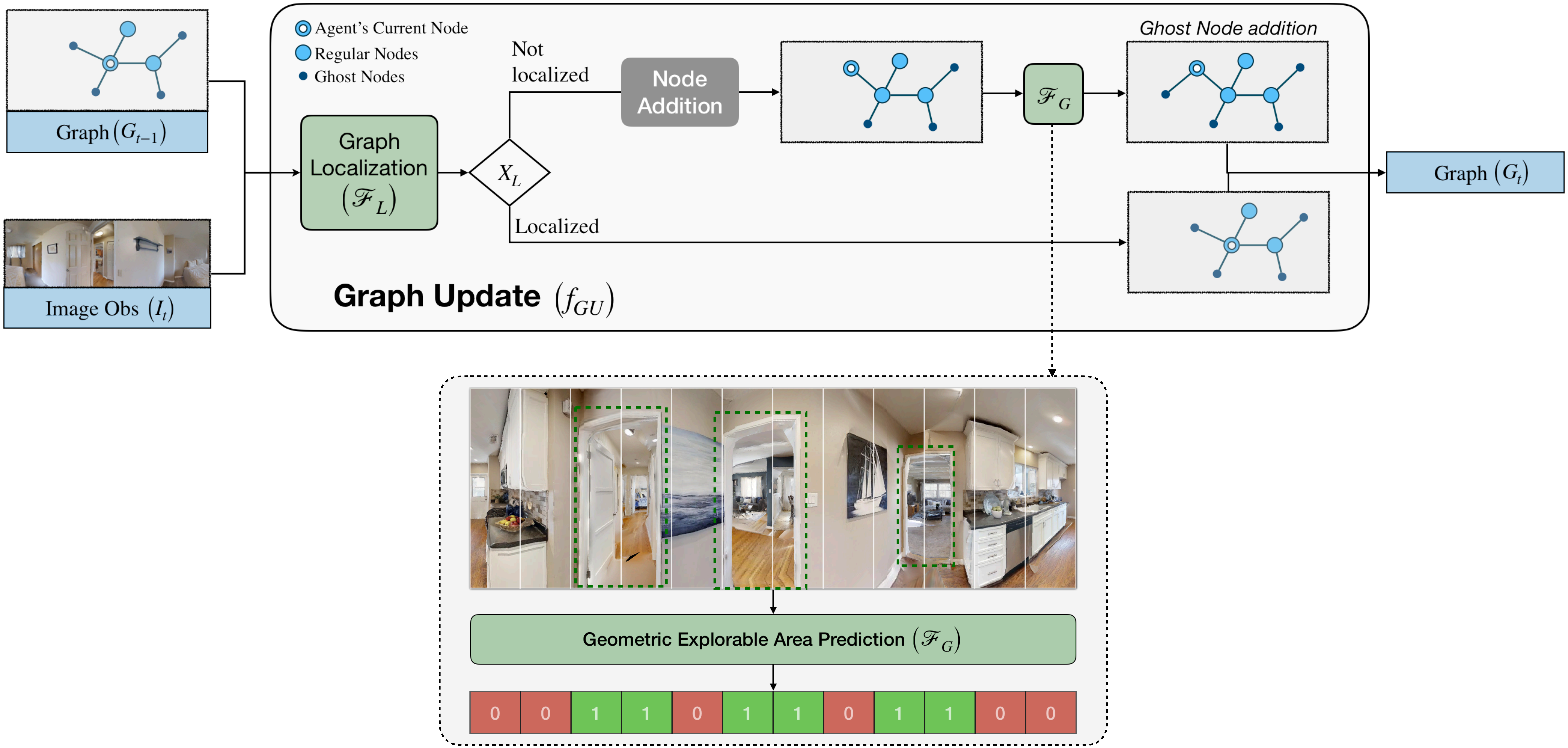
Building the Representation



Building the Representation

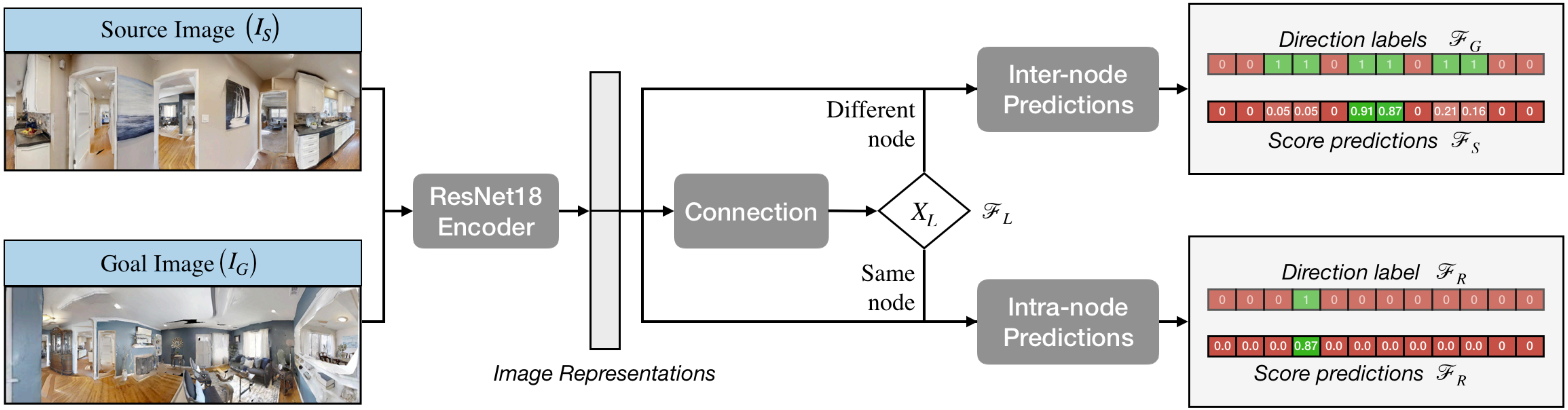


Building the Representation

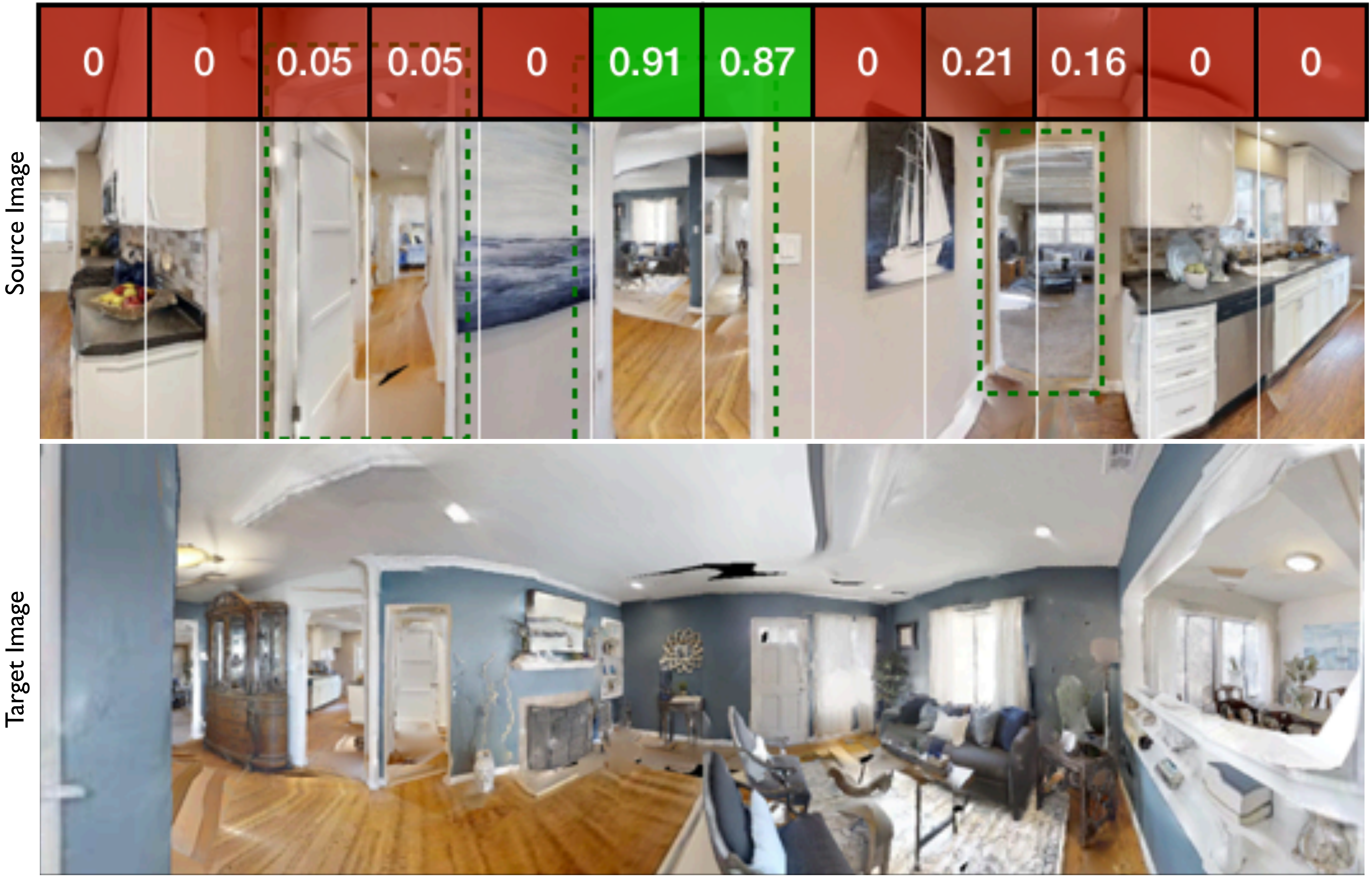


Single Supervised Learning Model

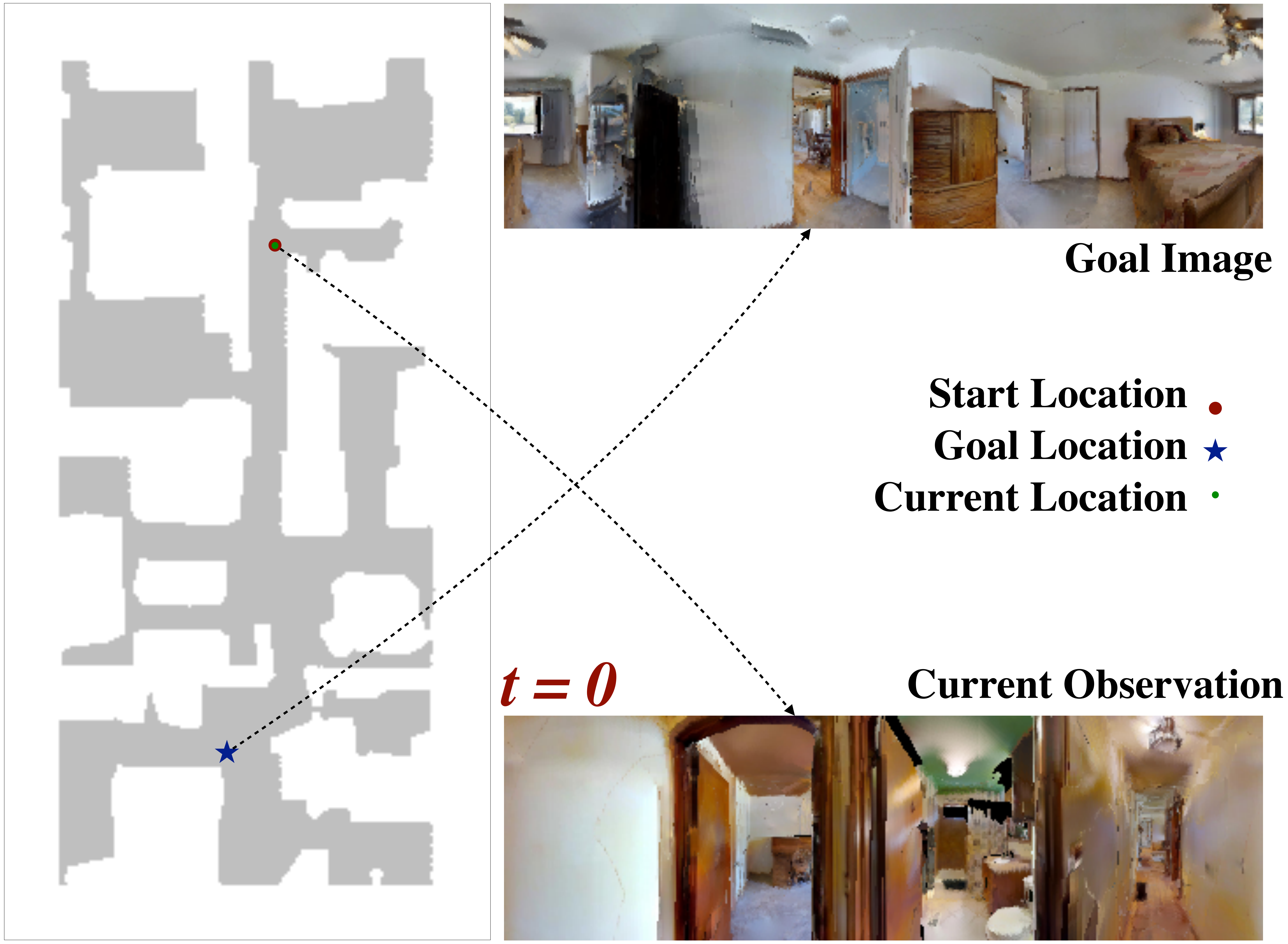
- $F_G(I_1)$ = Geometric Prediction: Free directions
- $F_S(I_1, I_2)$ = Semantic Prediction: Closeness to target
- $F_R(I_1, I_2)$ = Relative Pose
- $F_L(I_1, I_2)$ = Localization



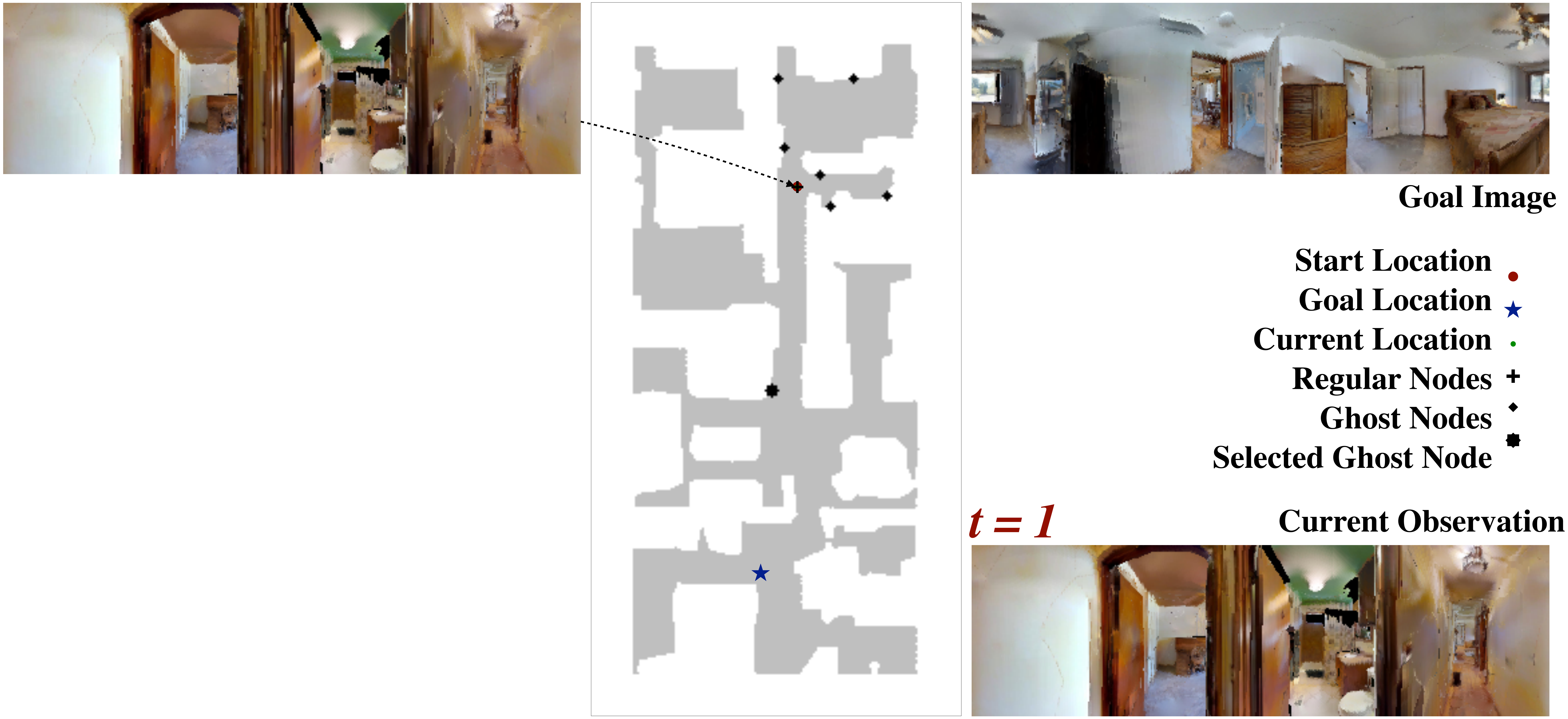
- No reinforcement learning, no interaction needed
- Can be trained completely with static data


0
0
0.05
0.05
0
0.91
0.87
0
0.21
0.16
0
0
Source Image
Target Image

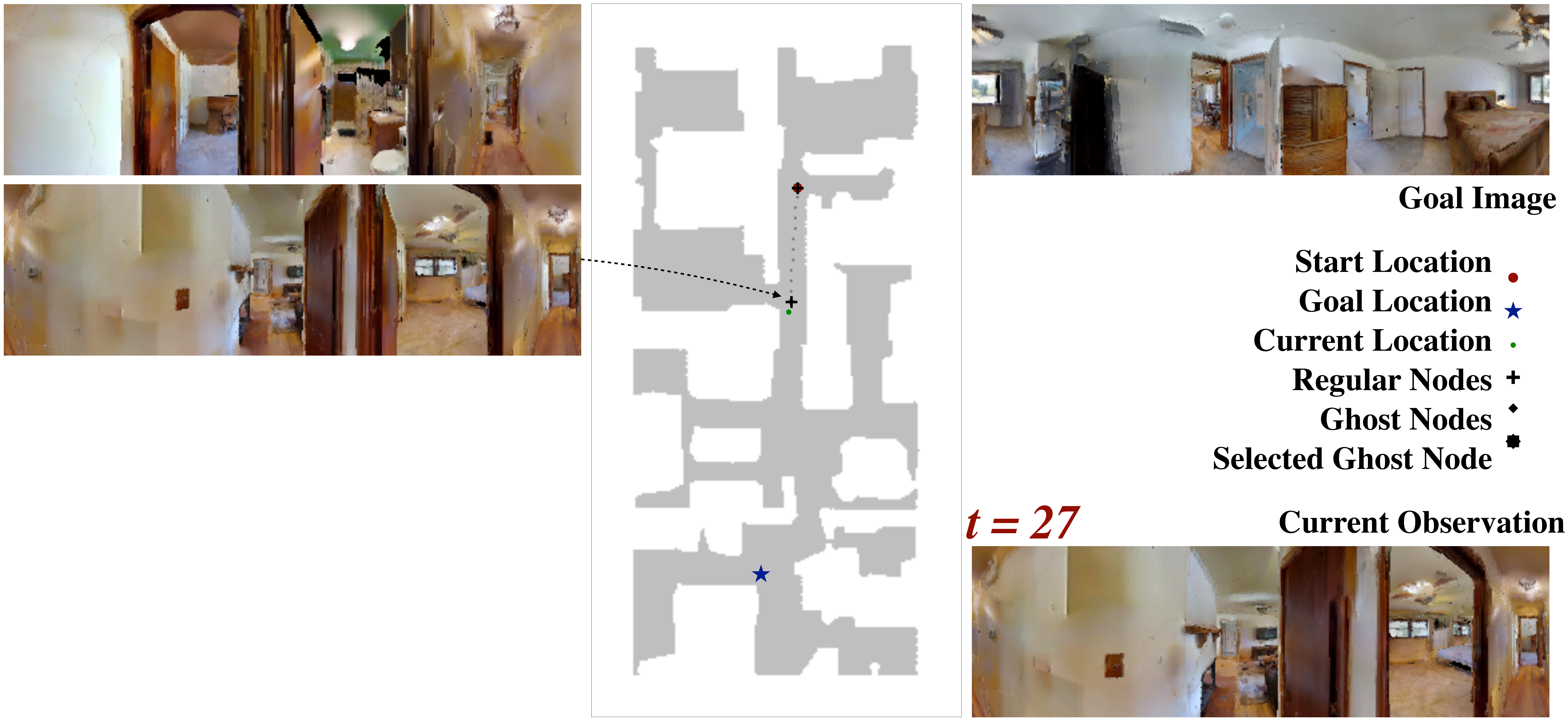

0
0
0.68
0.69
0
0.21
0.23
0
0.16
0.16
0
0
Source Image
Target Image



Goal Image
Start Location
Goal Location
Current Location
t = 0
Current Observation

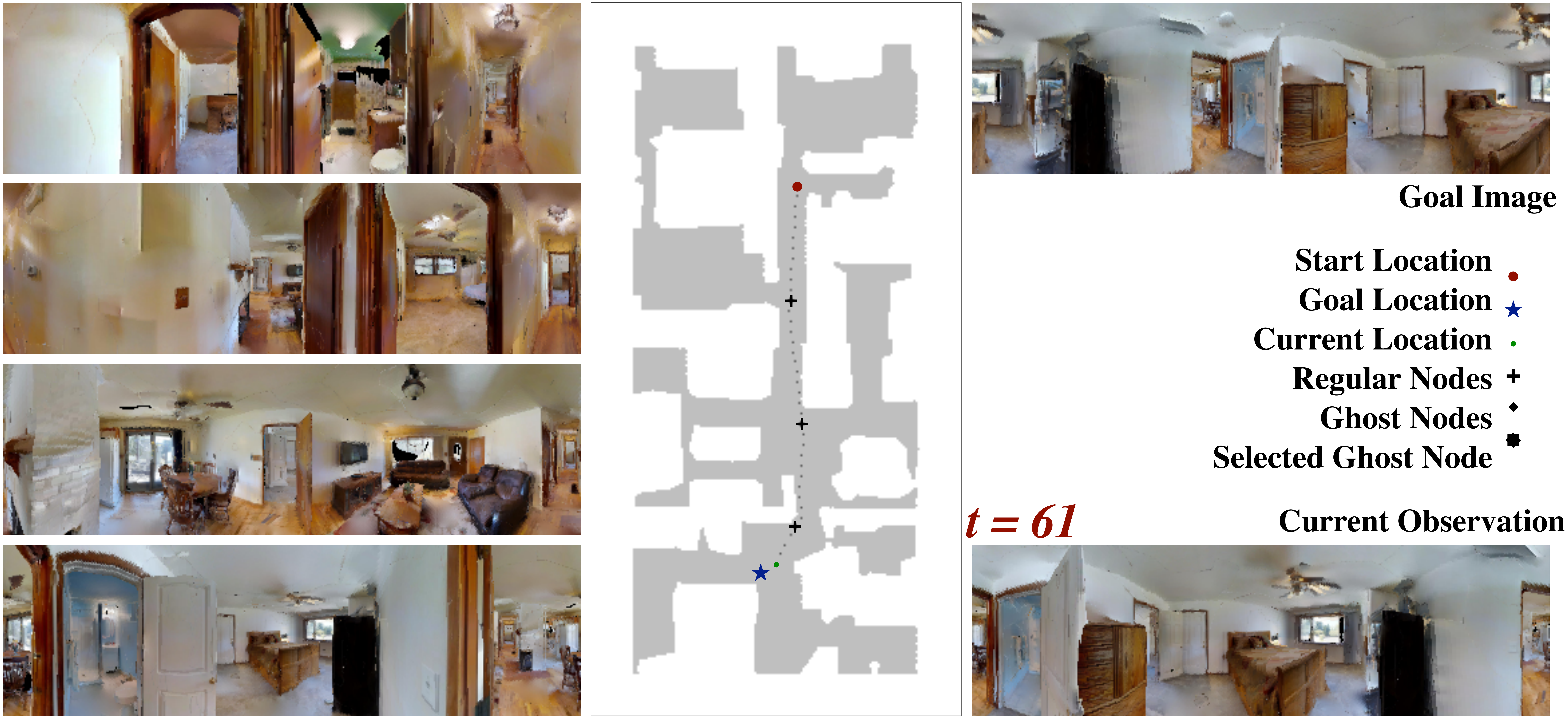

Goal Image
Start Location
Goal Location
Current Location
Regular Nodes
Ghost Nodes
Selected Ghost Node
t = 1
Current Observation


Goal Image
Start Location
Goal Location
Current Location
Regular Nodes
Ghost Nodes
Selected Ghost Node
t = 20
Current Observation


Goal Image
Start Location
Goal Location
Current Location
Regular Nodes
Ghost Nodes
Selected Ghost Node
t = 27
Current Observation


Goal Image
Start Location
Goal Location
Current Location
Regular Nodes
Ghost Nodes
Selected Ghost Node
t = 27
Current Observation

Goal Image
Start Location
Goal Location
Current Location
Regular Nodes
Ghost Nodes
Selected Ghost Node
t = 56
Current Observation



Goal Image
Start Location
Goal Location
Current Location
Regular Nodes
Ghost Nodes
Selected Ghost Node
t = 61
Current Observation

Results (SPL)

		RGB	RGBD	RGBD (No Noise)	RGBD (No Stop)
Vanilla LSTM Memory	LSTM + Imitation	0.10	0.14	0.15	0.18
	LSTM + RL	0.10	0.13	0.14	0.17
Metric Maps	Occupancy Maps + FBE + RL		0.26	0.31	0.24
	ANS	0.23	0.29	0.35	0.39
Topological Maps	**NTS (Our)**	**0.38**	**0.43**	**0.45**	**0.60**

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

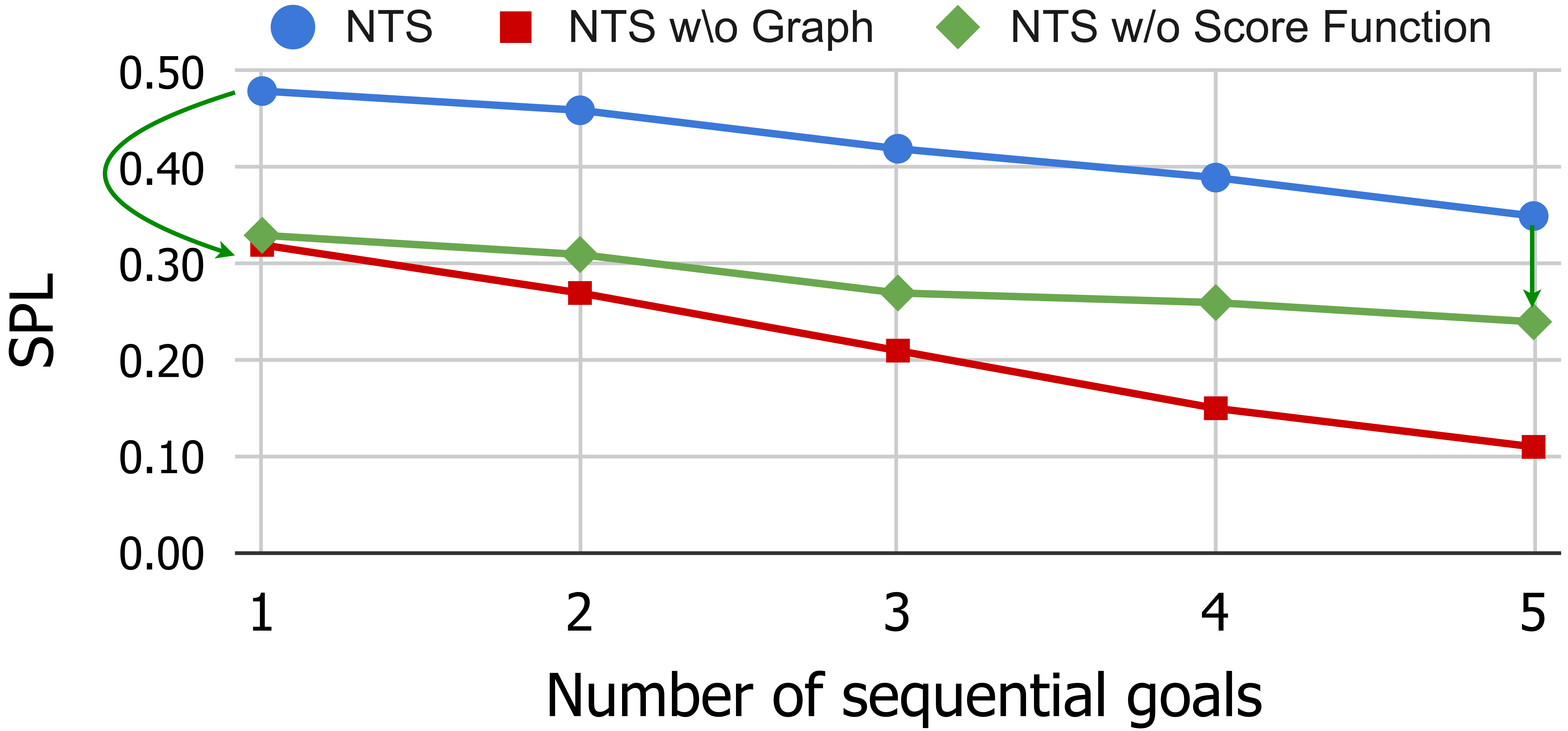
NTS is better than occupancy map models, captures and uses semantic priors.

Robustness to Noise

Results

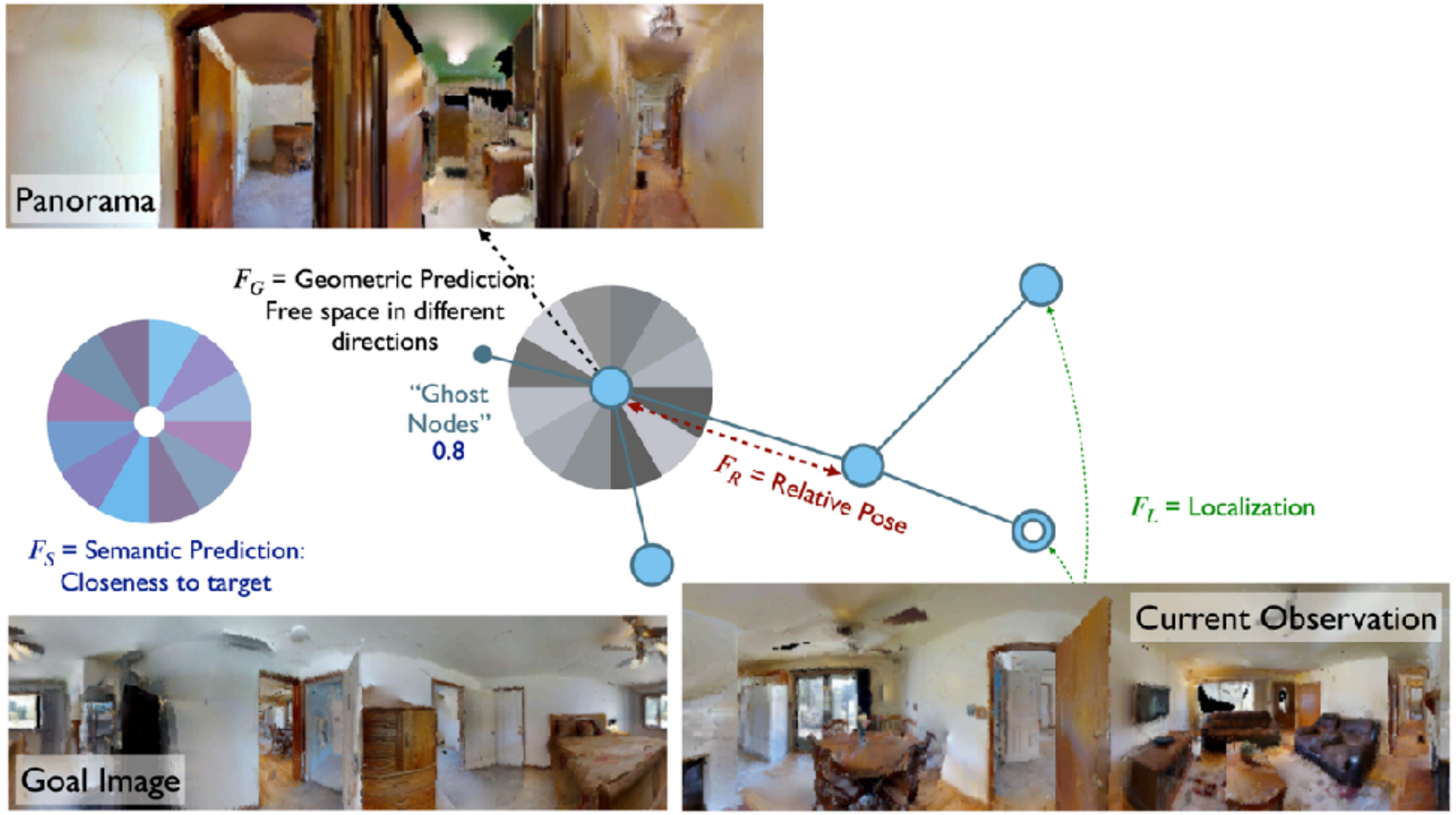
Semantic score function improves efficiency when no prior experience with environment is available.

As experience in environment increases, utility of semantic function decreases



But, at the same time, importance of the topological representation increases

Representation for Places



- Spatial reasoning
- Semantic reasoning

- ~~Sensitive to pose error~~ Robust to pose error
- ~~Interactive training~~ Offline supervised training - Still requires a simulator
- ~~Long training horizons~~ Modularized policy

$F_G(I_1)$: Geometry prediction

$F_R(I_1, I_2)$: Relative Pose

$F_L(I_1, I_2)$: Localization

$\boxed{F_S(I_1, I_2)\text{: Semantic Prediction}}$

Can we simplify and scale-up training further?

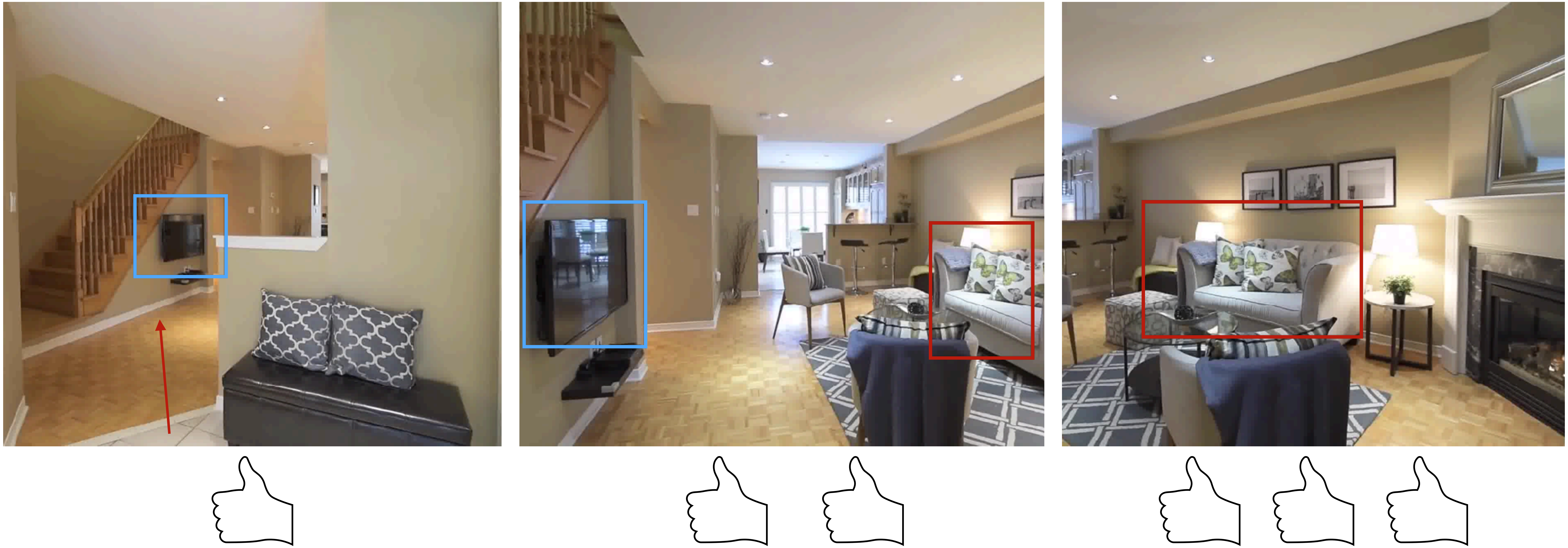
Learning F_s by Watching YouTube Videos

M. Chang, A. Gupta, S. Gupta. Semantic Visual Navigation by Watching YouTube Videos. In *NeurIPS*, 2020.

Basic Intuition

Mine for spatial co-occurrences

Video *time*



e.g. cues for finding a couch

Challenges in Using Such Videos

- Videos don't come with action labels

  $\Longrightarrow$ Action Grounding via an Inverse Model [1]

- Goals and intents are not known

  $\Longrightarrow$ Use off-the-shelf object detectors to label frames with desired objects

- Depicted trajectories may not be optimal

  $\Longrightarrow$ Use Q-learning to learn optimal behavior from sub-optimal data [2]

[1] A. Kumar, S. Gupta, J. Malik. Learning navigation subroutines by watching videos. In *CoRL*, 2019.

[2] Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Value Learning from Videos (VLV)

Learned Value Function

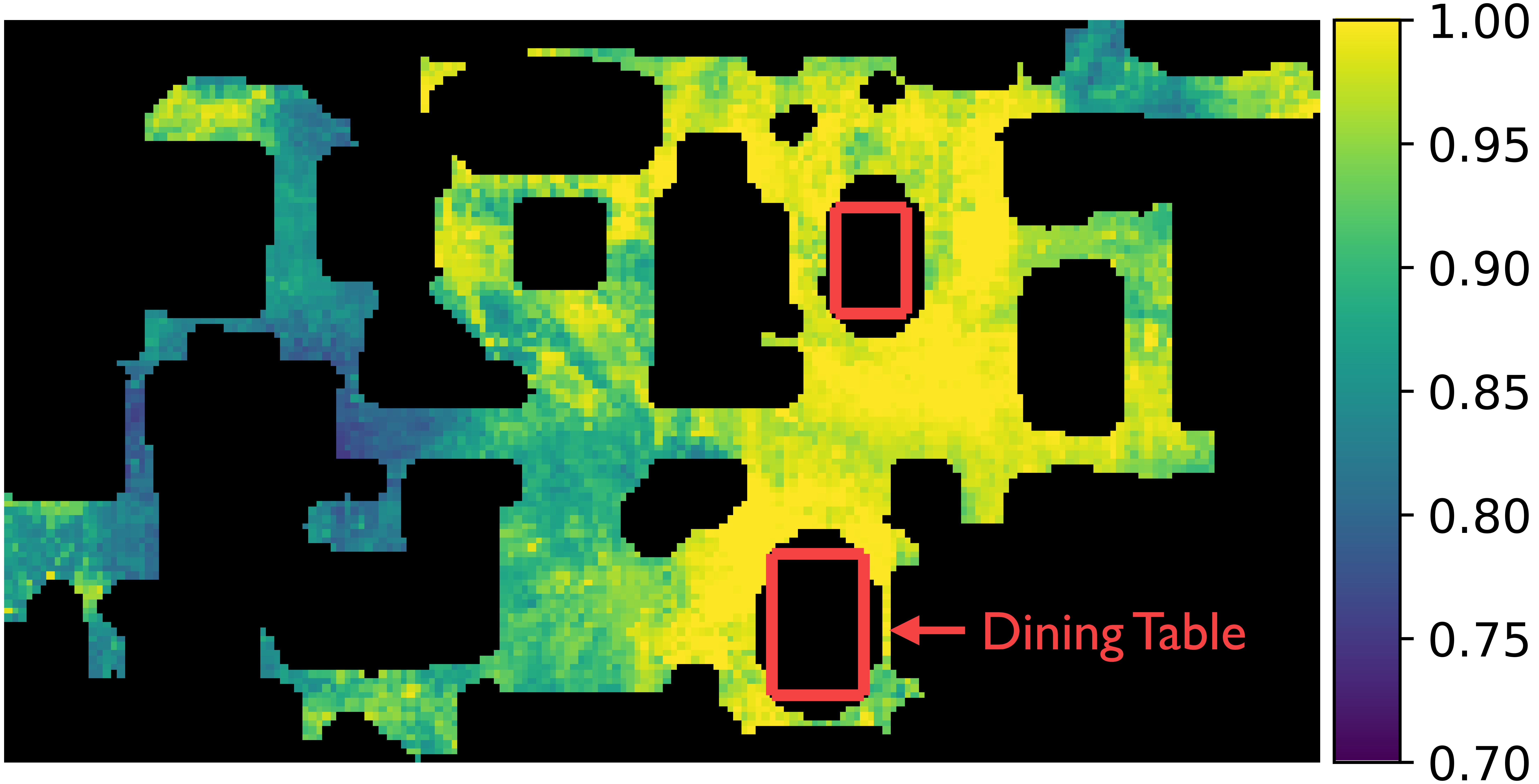
$f(I, c) \approx$ nearness to goal

Value function predicts a proxy for nearness to a goal object for a given image

Learned Value Function

$f(I, c) \approx$ nearness to goal

Value function predicts a proxy for nearness to a goal object for a given image

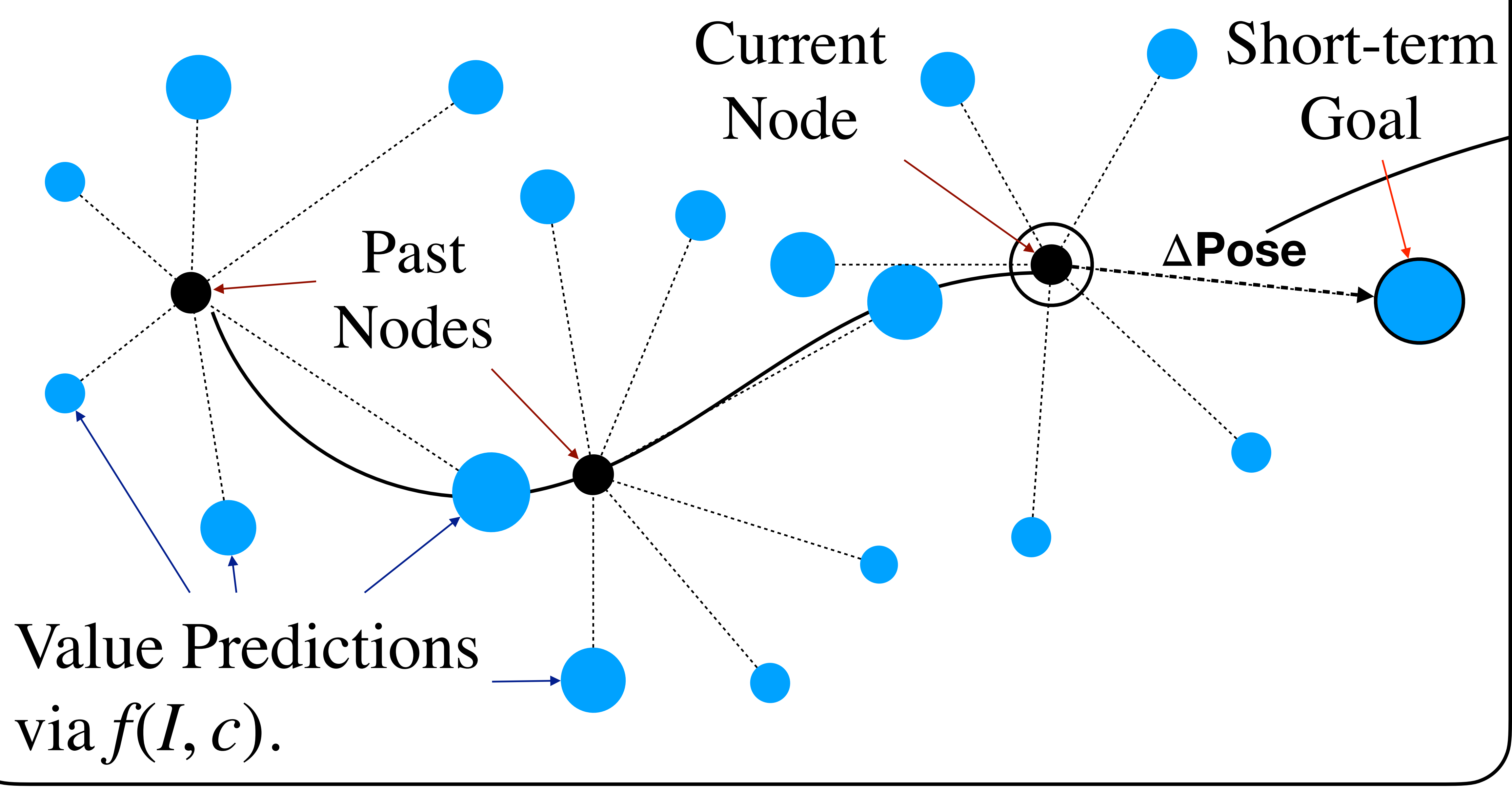


Using Learned Values for Semantic Navigation

Hierarchical Policy

High-Level Policy

- Decides where to go next and emits short-term goal
- Builds a topological map [1] that stores values predicted by $f(I, c)$ at different locations in different directions



Low-Level Policy

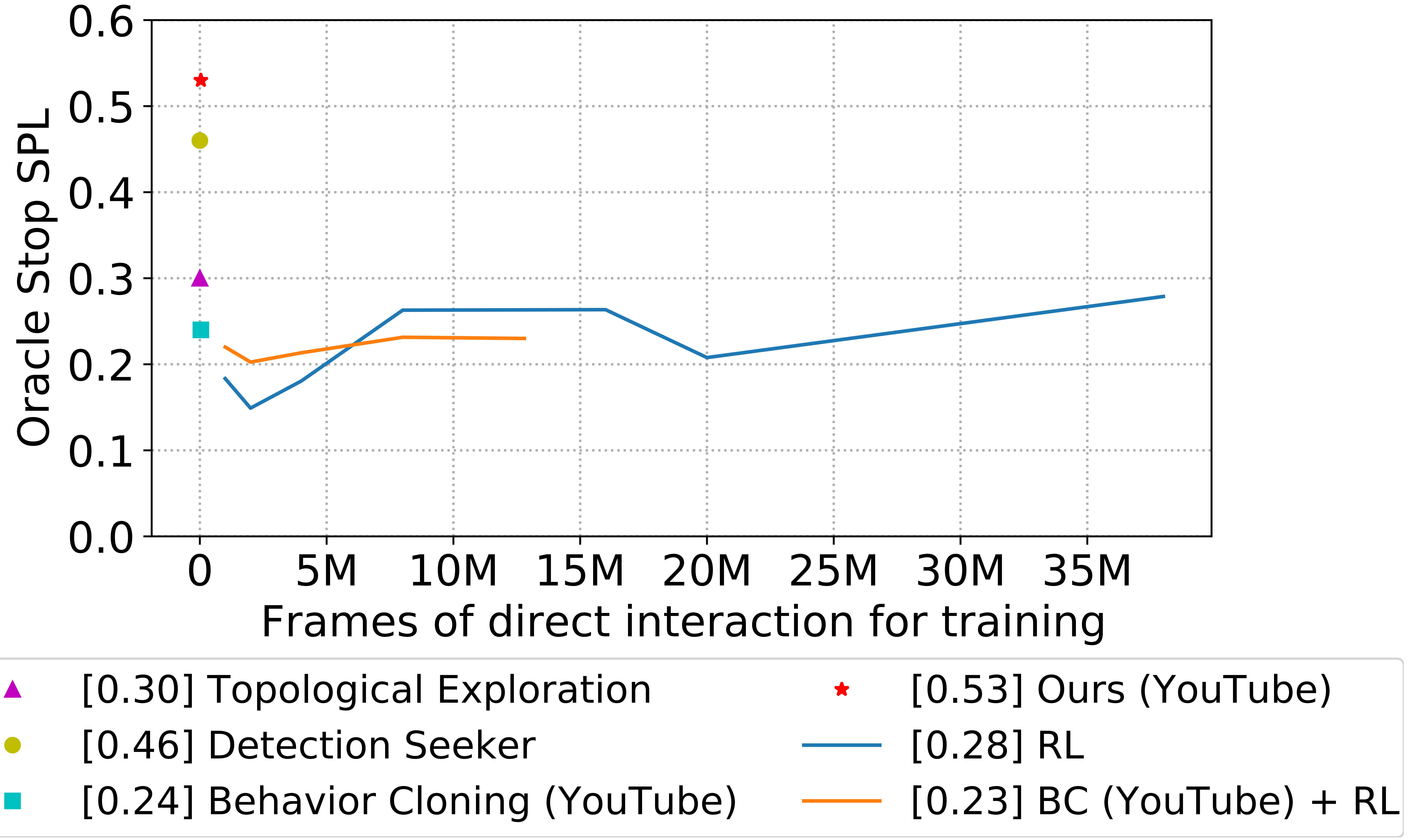
- Executes actions to achieve short-term goal
- Incrementally builds occupancy map from depth camera, plans paths

Occupancy Map | FMM Cost Map

Forward
Left
Right
Stop

[1] D. Chaplot, R. Salakhutdinov, A. Gupta, S. Gupta. Neural topological slam for visual navigation. In *CVPR*, 2020.

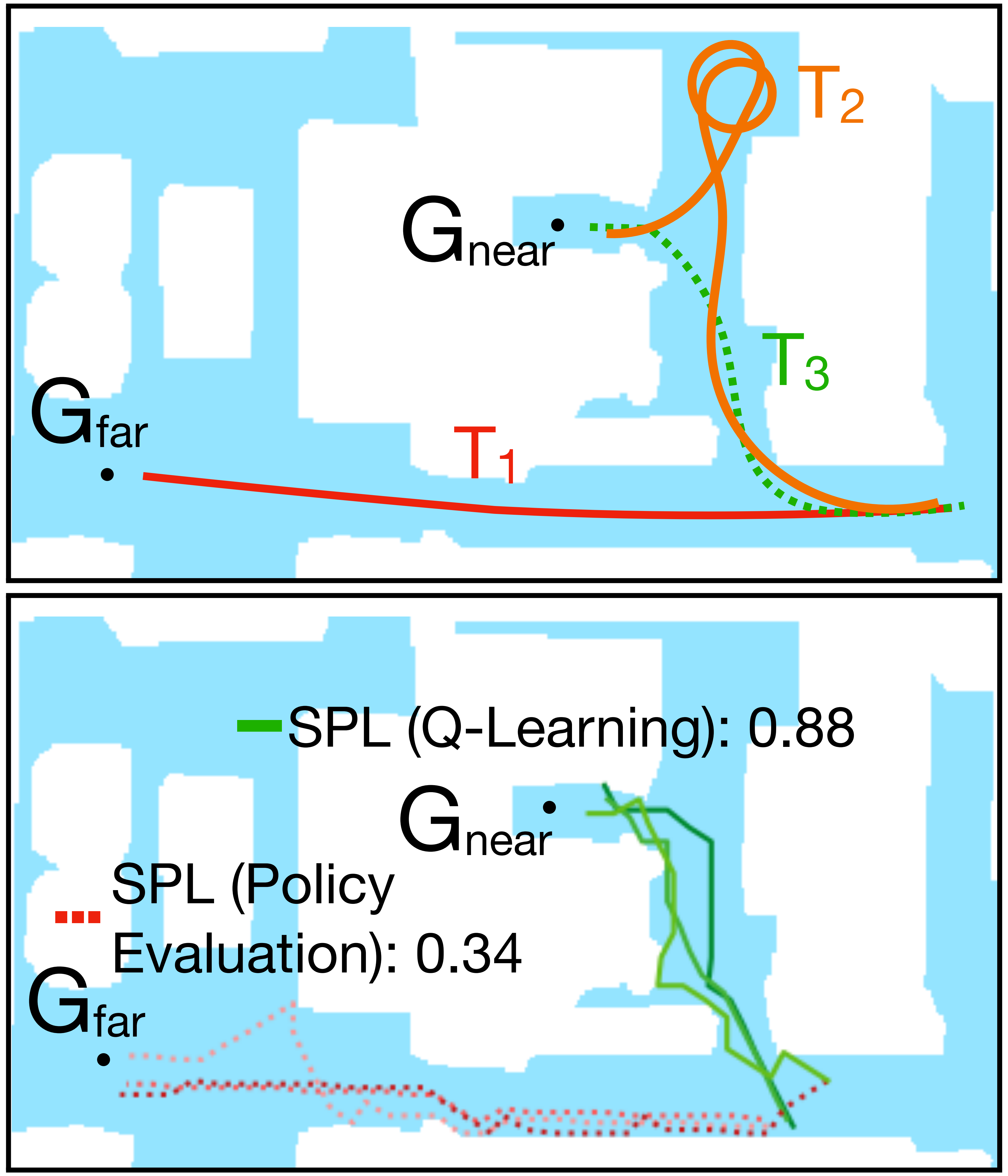
Results


Oracle Stop SPL
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5M
10M
15M
20M
25M
30M
35M
Frames of direct interaction for training
[0.30] Topological Exploration
[0.46] Detection Seeker
[0.24] Behavior Cloning (YouTube)
[0.53] Ours (YouTube)
[0.28] RL
[0.23] BC (YouTube) + RL

Ablations

Oracle Stop SPL

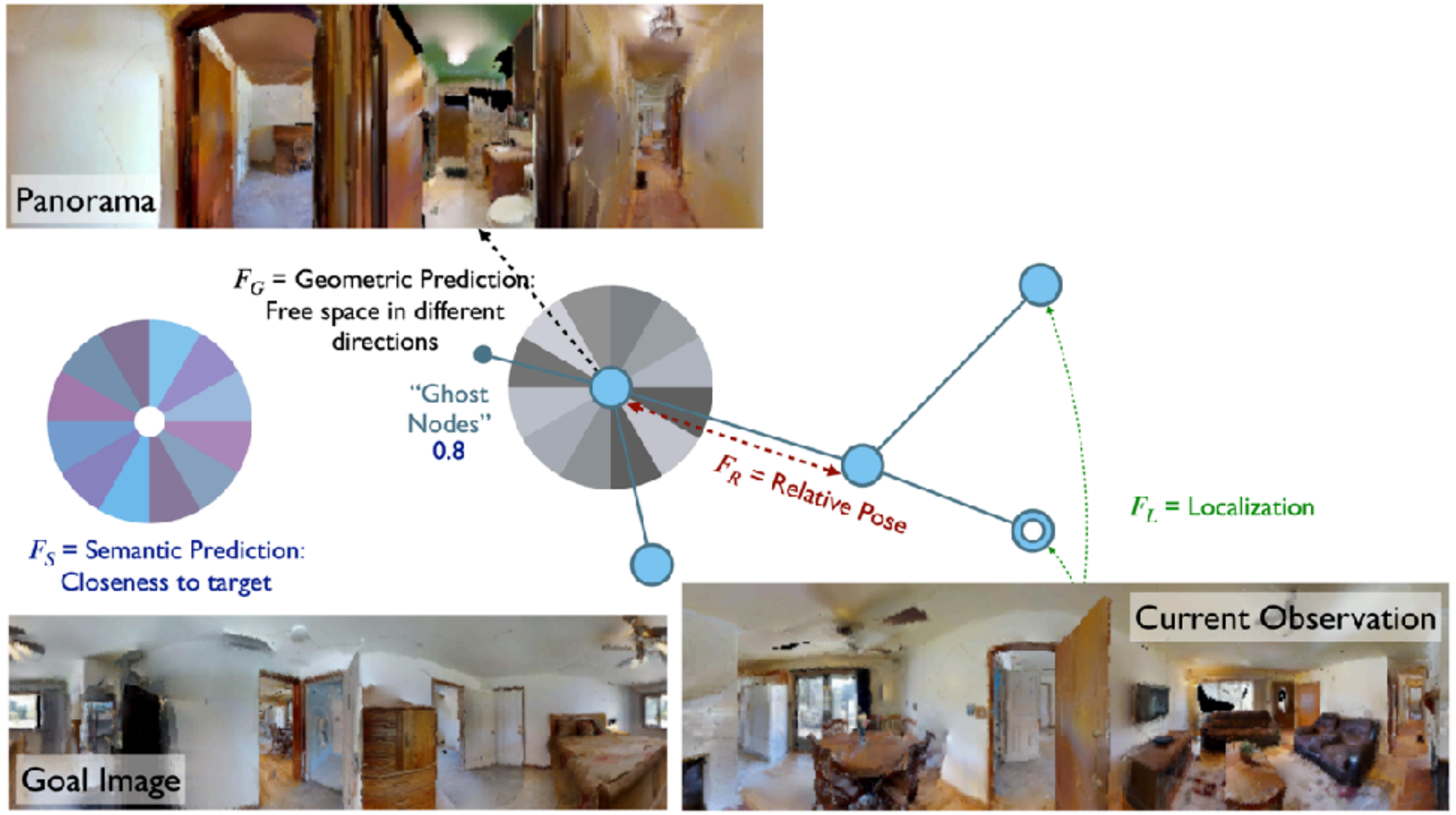
Method	Easy	Medium	Hard	Overall
Base Setting	0.62	0.42	0.23	0.40
True Actions	0.61	0.45	0.25	0.41
True Detections	0.62	0.45	0.22	0.40
True Rewards	0.64	0.46	0.21	0.41
Optimal Trajectories	0.65	0.46	0.25	0.43
Detector Score	0.73	0.48	0.26	0.46
Train on 360° Videos	0.66	0.51	0.32	0.47
No Hierarchy	0.38	0.10	0.02	0.15



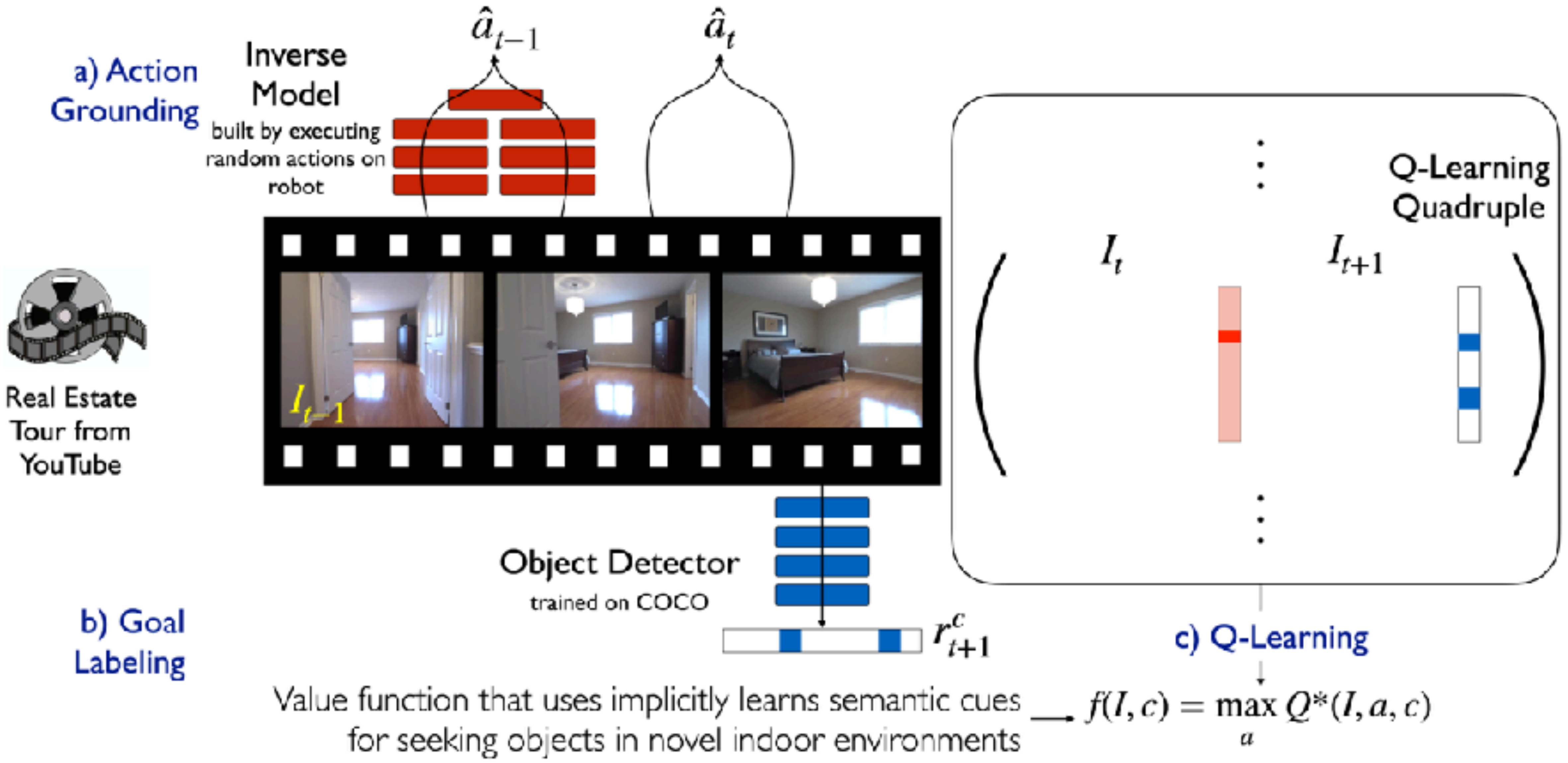
- Inverse model and detector do not hurt performance significantly
- Detector at test time helps for close objects, panorama helps for far objects
- Q-Learning outperforms simple policy evaluation for challenging environments
- Hierarchical policy is a major factor in strong performance

Navigation to couches in novel environments

Representation for Places



- Spatial reasoning
- Semantic reasoning
- Robust to pose error
- Modularized policy
- Training from in-the-wild videos



In this talk,

Representations for Places that Afford Navigation in Novel Environments

- *Augmenting metric representations with semantic reasoning*
- *Relaxing the need for metric representations*
- *Scaling-up training of such representations*

Operationalize insights from classical robotics into learning paradigms