Last-Mile Embodied Visual Navigation

In submission to CoRL 2022

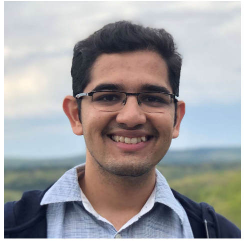
EAI seminar | Aug 19th 2022

Justin Wasserman*

Karmesh Yadav

Girish Chowdhary

Abhinav Gupta



Unnat Jain*

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Carnegie Mellon University

Meta AI

Image-goal navigation

- Observes RGBD and pose at each step

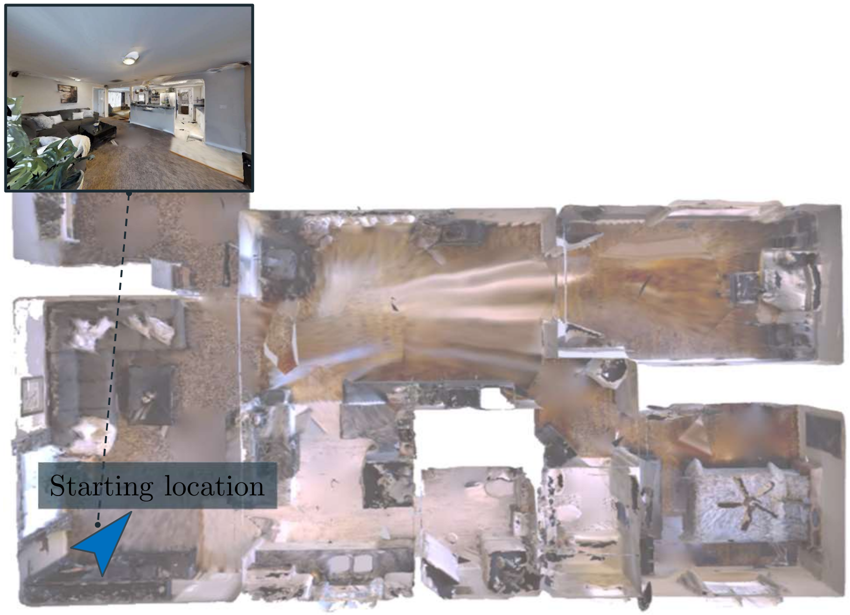


Image-goal navigation

❖Observes RGBD and pose at each step

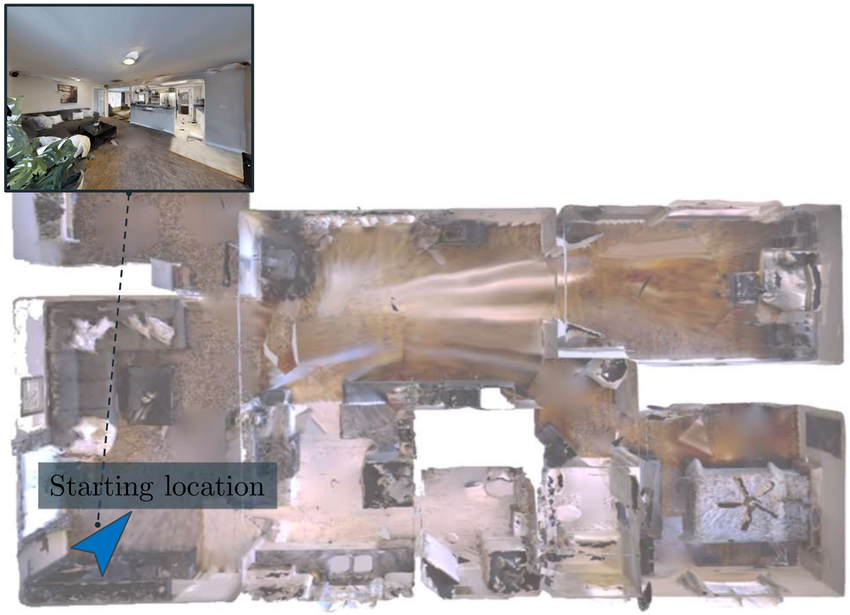


Image-goal navigation

- ❖Observes RGBD and pose at each step
- ❖Must navigate to goal RGB image

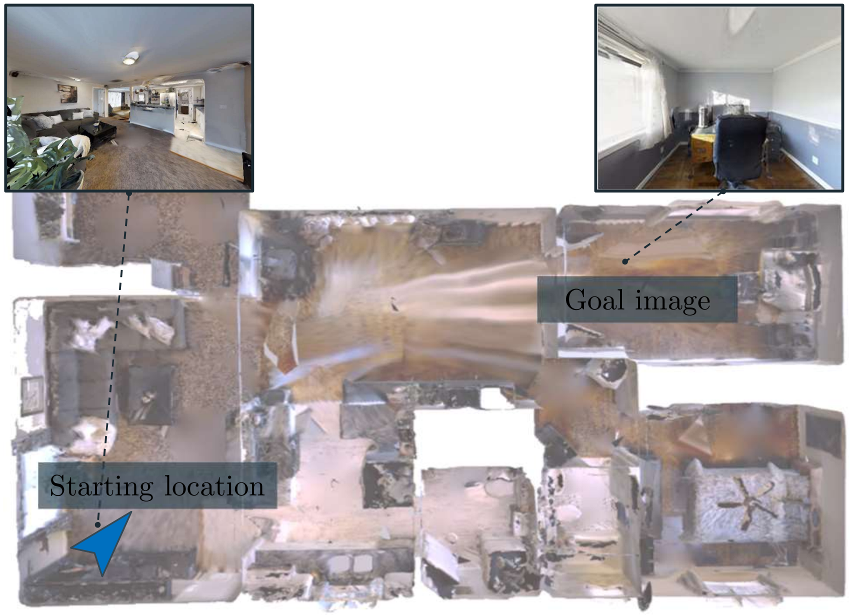


Image-goal navigation

- ❖Observes RGBD and pose at each step
- ❖Must navigate to goal RGB image
- ❖Action space: {forward, turn right, turn left, stop}

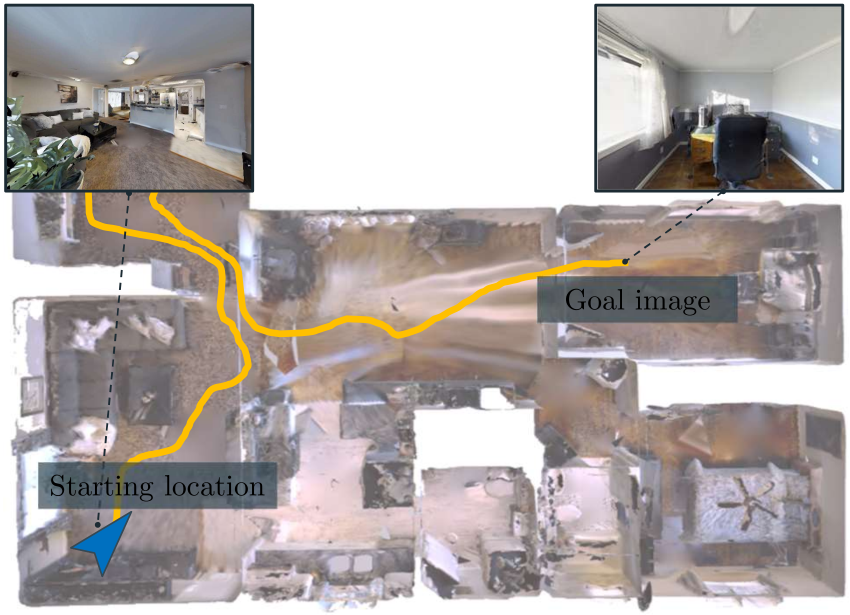
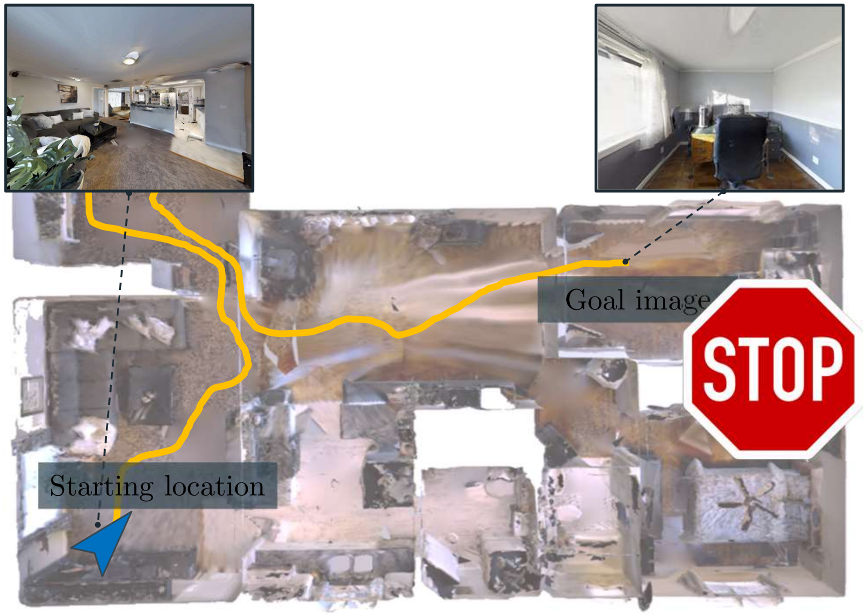
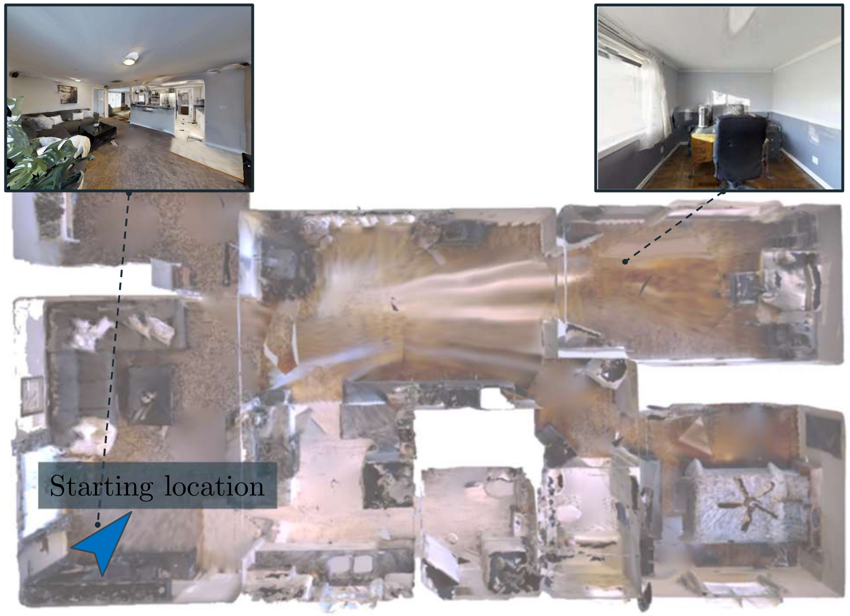


Image-goal navigation

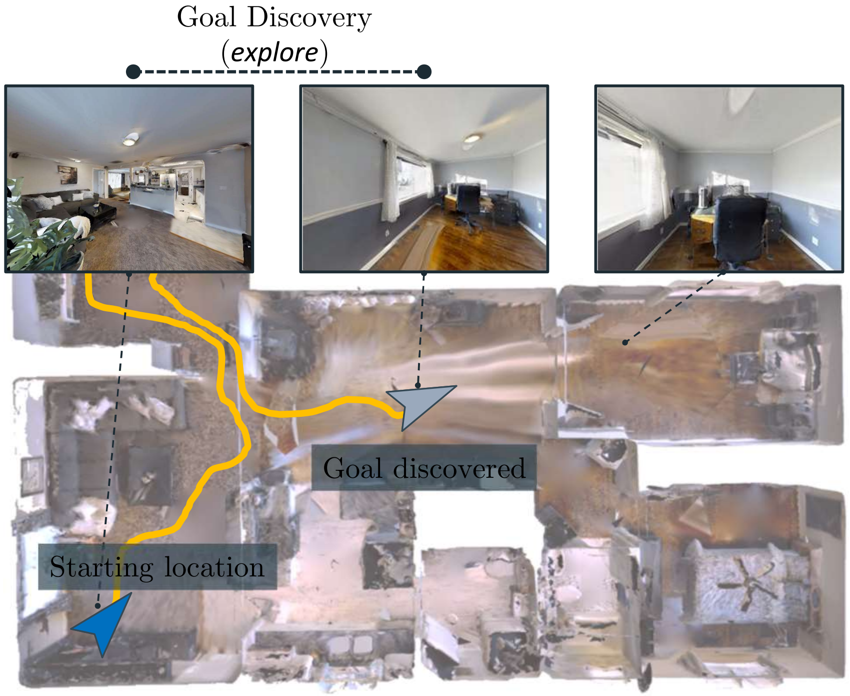
- Observes RGBD and pose at each step
- Must navigate to goal RGB image
- Action space: {forward, turn right, turn left, stop}
- Success: “stop” within 1m of the goal within 500 steps



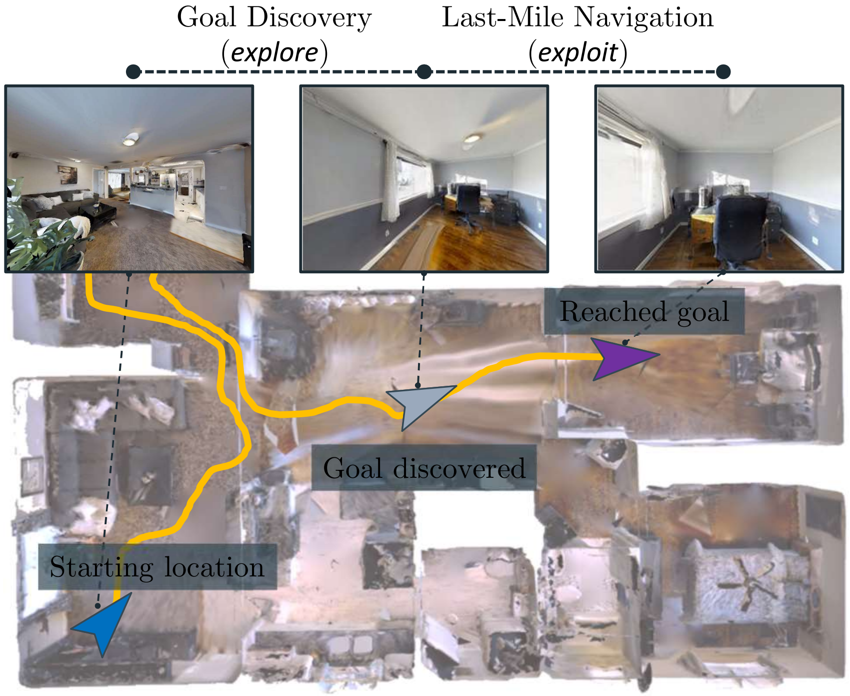
Motivation



Motivation

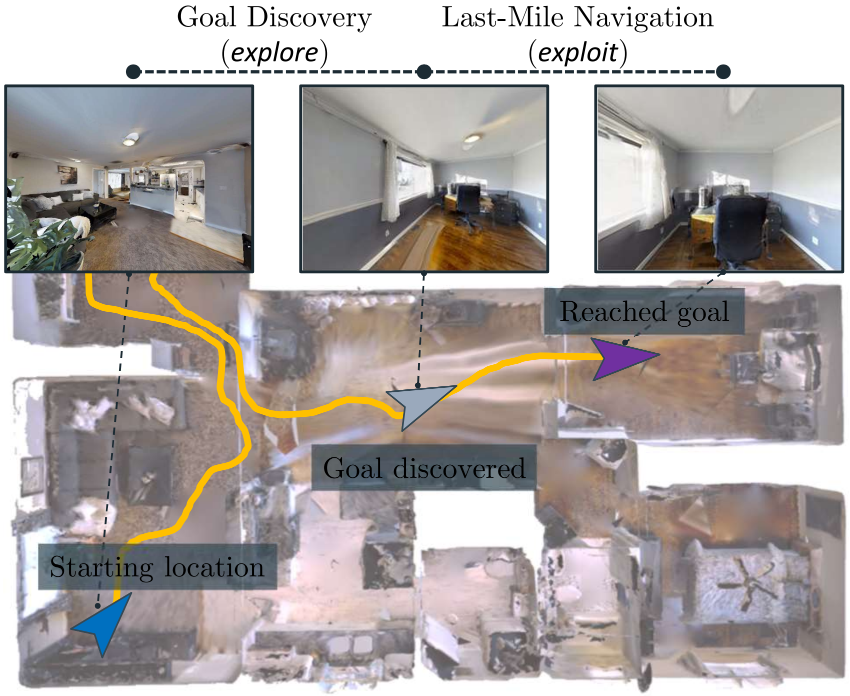

Goal Discovery
(*explore*)
Goal discovered
Starting location

Motivation


Goal Discovery
(*explore*)
Last-Mile Navigation
(*exploit*)
Reached goal
Goal discovered
Starting location

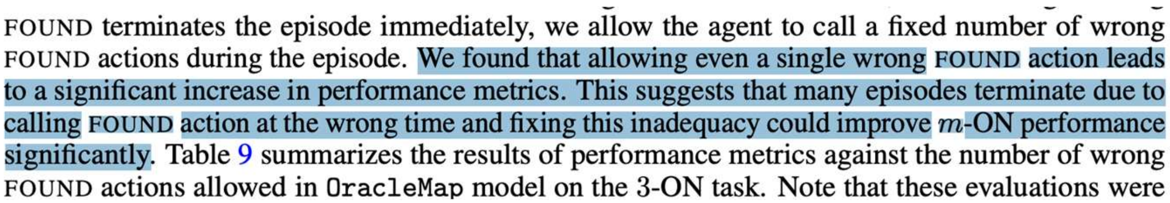
Motivation

Agents across embodied AI have difficulties getting close to the goal even when they see the goal.



Motivation

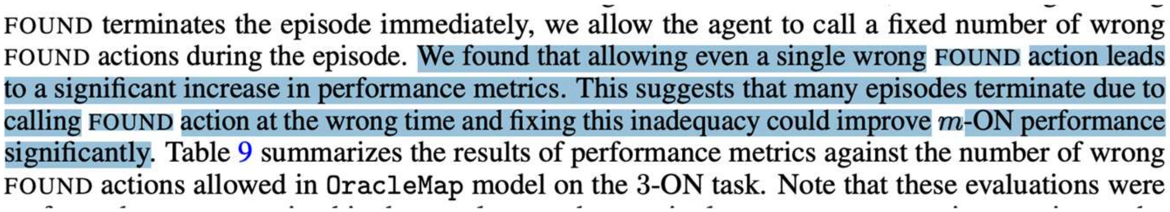
S. Wani et al. **MultiON** [NeurIPS 2020]



FOUND terminates the episode immediately, we allow the agent to call a fixed number of wrong FOUND actions during the episode. We found that allowing even a single wrong FOUND action leads to a significant increase in performance metrics. This suggests that many episodes terminate due to calling FOUND action at the wrong time and fixing this inadequacy could improve m-ON performance significantly. Table 9 summarizes the results of performance metrics against the number of wrong FOUND actions allowed in OracleMap model on the 3-ON task. Note that these evaluations were

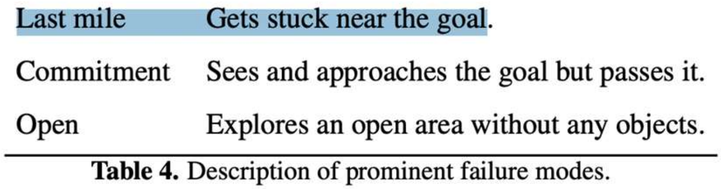
Motivation

S. Wani et al. **MultiON** [NeurIPS 2020]



FOUND terminates the episode immediately, we allow the agent to call a fixed number of wrong FOUND actions during the episode. We found that allowing even a single wrong FOUND action leads to a significant increase in performance metrics. This suggests that many episodes terminate due to calling FOUND action at the wrong time and fixing this inadequacy could improve m-ON performance significantly. Table 9 summarizes the results of performance metrics against the number of wrong FOUND actions allowed in `OracleMap` model on the 3-ON task. Note that these evaluations were

P. Chattopadhyay et al. **RobustNav** [ICCV 2021]

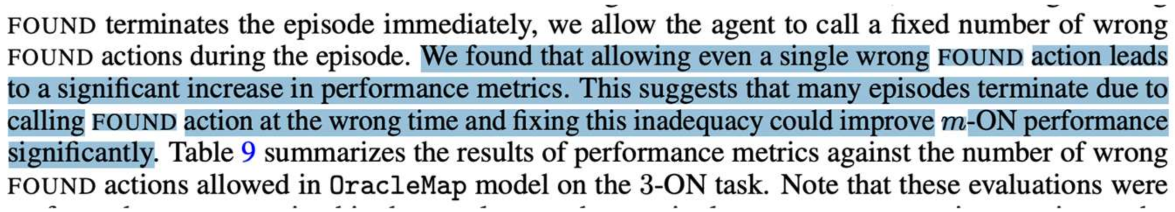


Last mile	Gets stuck near the goal.
Commitment	Sees and approaches the goal but passes it.
Open	Explores an open area without any objects.

Table 4. Description of prominent failure modes.

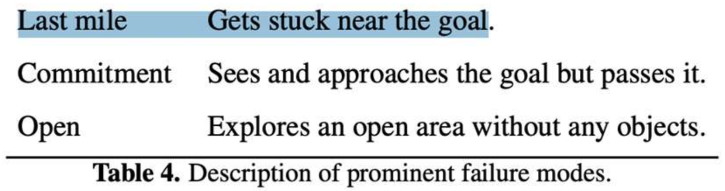
Motivation

S. Wani et al. **MultiON**
[NeurIPS 2020]



FOUND terminates the episode immediately, we allow the agent to call a fixed number of wrong FOUND actions during the episode. We found that allowing even a single wrong FOUND action leads to a significant increase in performance metrics. This suggests that many episodes terminate due to calling FOUND action at the wrong time and fixing this inadequacy could improve m-ON performance significantly. Table 9 summarizes the results of performance metrics against the number of wrong FOUND actions allowed in `OracleMap` model on the 3-ON task. Note that these evaluations were

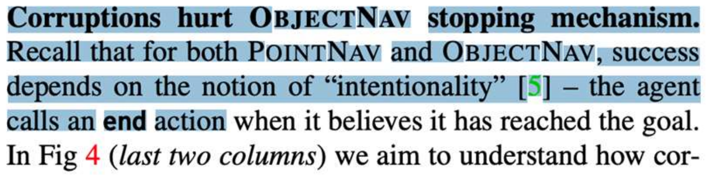
P. Chattopadhyay et al. **RobustNav**
[ICCV 2021]



Last mile	Gets stuck near the goal.
Commitment	Sees and approaches the goal but passes it.
Open	Explores an open area without any objects.

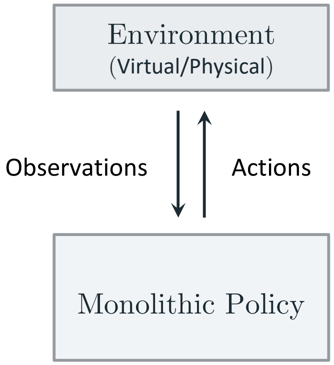
Table 4. Description of prominent failure modes.

J. Ye et al. **ObjectAux**
[ICCV 2021]

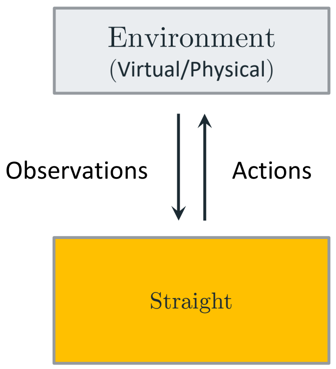


Corruptions hurt ObjectNav stopping mechanism. Recall that for both PointNav and ObjectNav, success depends on the notion of "intentionality" [5] – the agent calls an **end** action when it believes it has reached the goal. In Fig 4 (*last two columns*) we aim to understand how cor-

Solving image-goal navigation

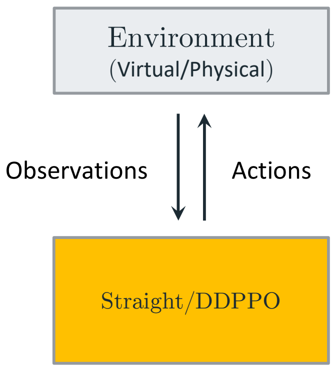


Solving image-goal navigation



[Heuristic]
T. Chen et al. Learning exploration policies for navigation.
ICLR 2019

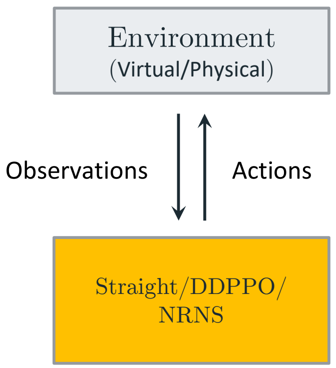
Solving image-goal navigation



[Heuristic]
T. Chen et al. Learning exploration policies for navigation.
ICLR 2019

[End-to-end deep RL]
E. Wijmans et al. DDPPO
ICLR 2019

Solving image-goal navigation

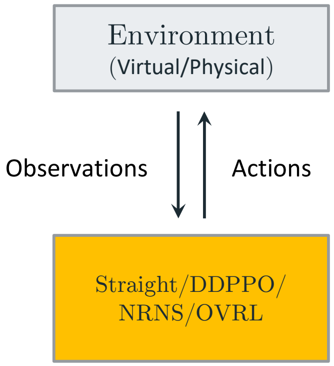


[Heuristic]
T. Chen et al. Learning exploration policies for navigation.
ICLR 2019

[End-to-end deep RL]
E. Wijmans et al. DDPPO
ICLR 2019

[Modular]
M. Hahn et al. No rl, no simulation
NeurIPS 2021

Solving image-goal navigation



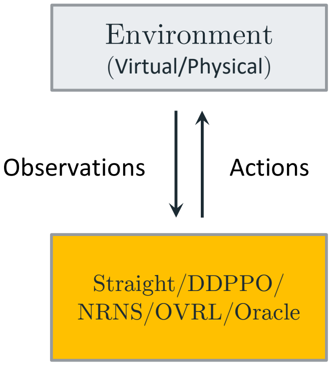
[Heuristic]
T. Chen et al. Learning exploration policies for navigation.
ICLR 2019

[End-to-end deep RL]
E. Wijmans et al. DDPPO
ICLR 2019

[Modular]
M. Hahn et al. No rl, no simulation
NeurIPS 2021

[SSL]
K. Yadav et al. Offline visual repr. Learning for embodied navigation.
arXiv 2022

Solving image-goal navigation



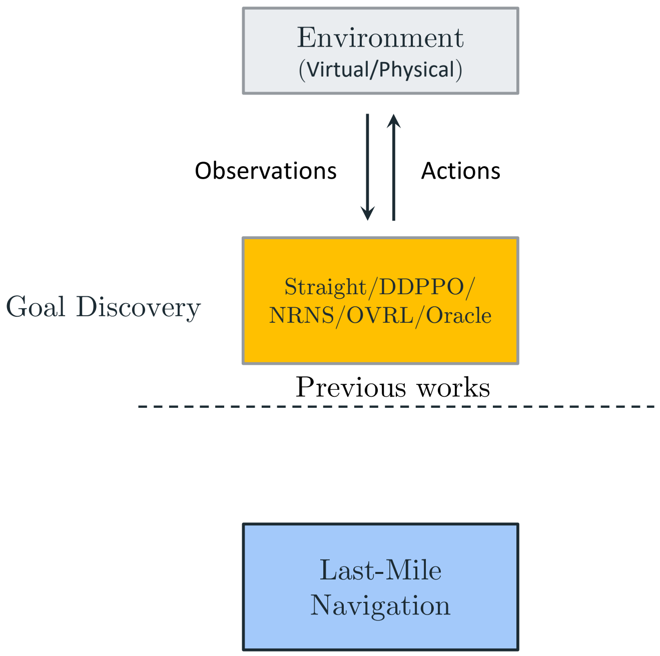
[Heuristic]
T. Chen et al. Learning exploration policies for navigation.
ICLR 2019

[End-to-end deep RL]
E. Wijmans et al. DDPPO
ICLR 2019

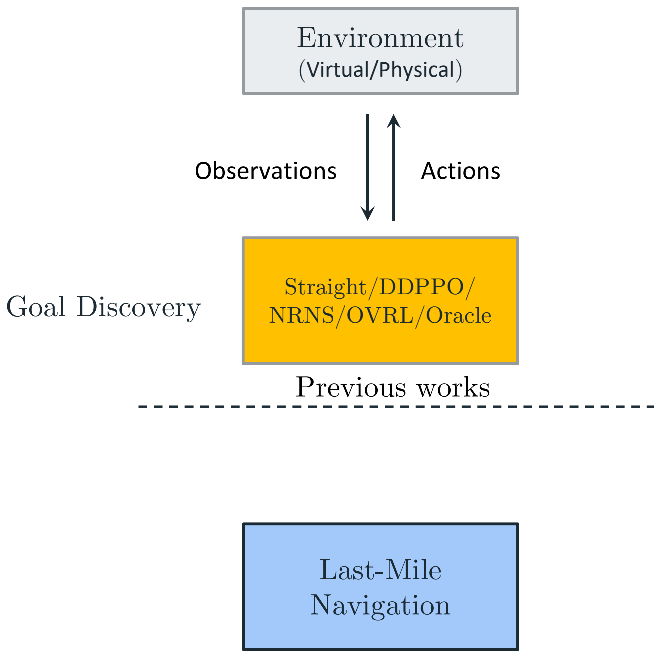
[Modular]
M. Hahn et al. No rl, no simulation
NeurIPS 2021

[SSL]
K. Yadav et al. Offline visual repr. Learning for embodied navigation.
arXiv 2022

Focus on Last-Mile navigation

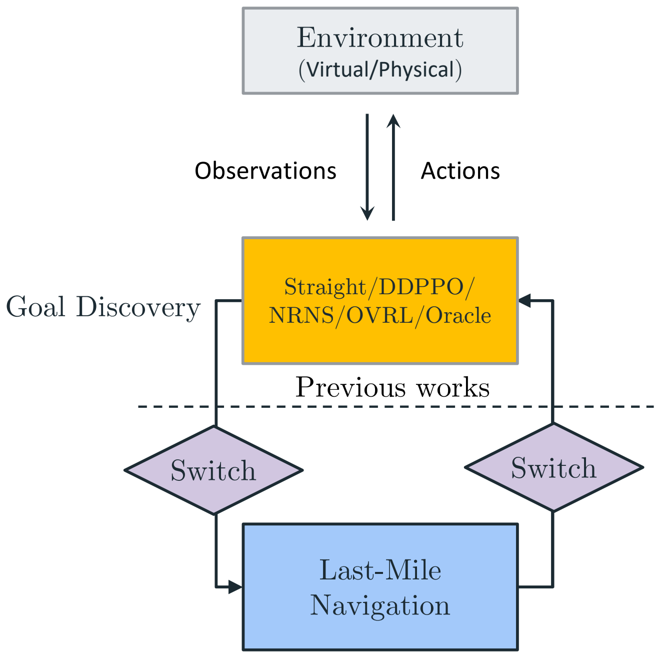


Focus on Last-Mile navigation



Principled 3D vision module for the last mile

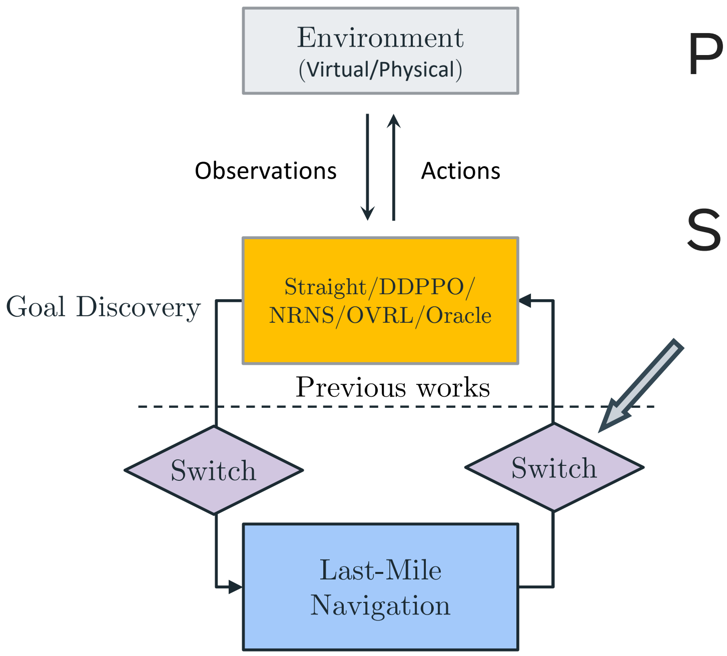
Focus on Last-Mile navigation



Principled 3D vision module for the last mile

Simple switches to plug-and-play

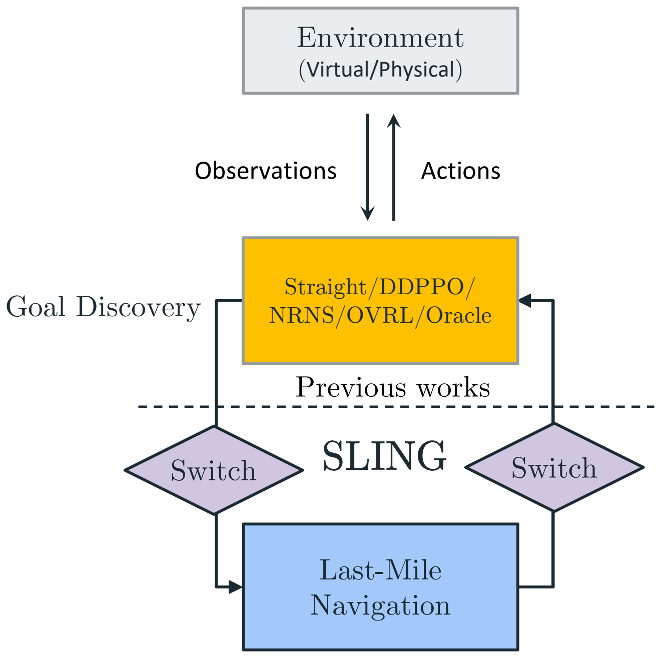
Focus on Last-Mile navigation



Principled 3D vision module for the last mile

Simple switches to plug-and-play

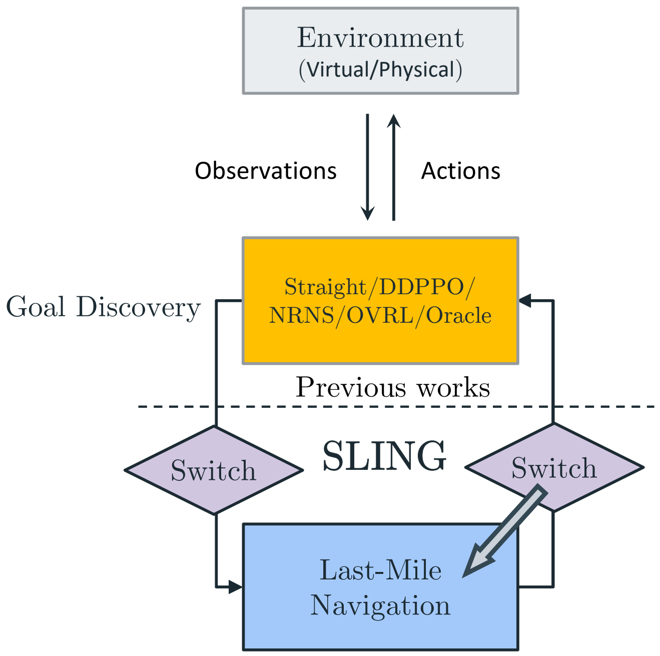
Switchable Last-Mile Image-Goal Navigation (SLING)



Principled 3D vision module for the last mile

Simple switches to plug-and-play

Switchable Last-Mile Image-Goal Navigation (SLING)



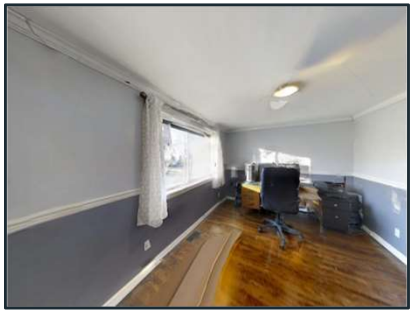
Principled 3D vision module for the last mile

Simple switches to plug-and-play

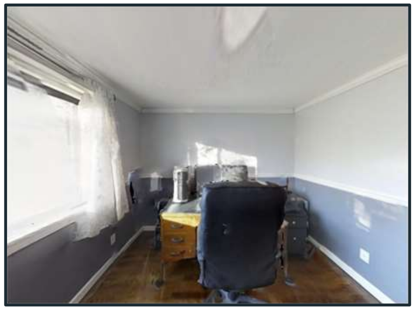
Our simple solution to Last-Mile Navigation

Our simple solution to Last-Mile Navigation

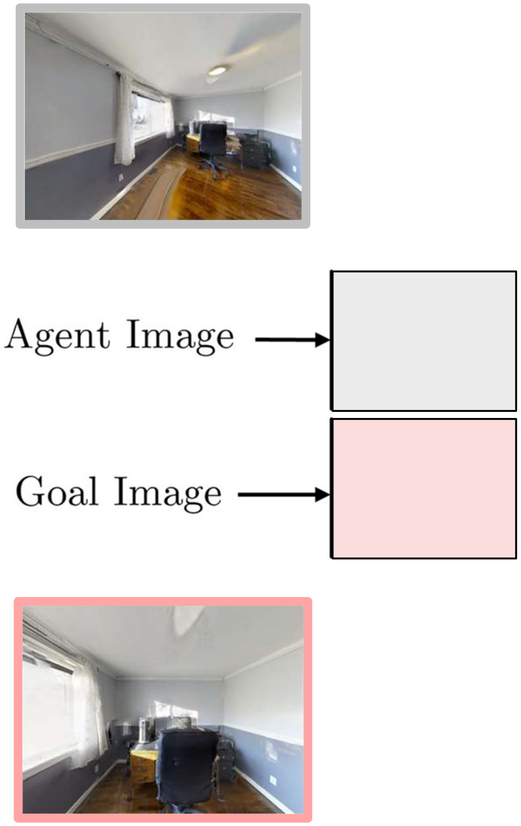
Agent Image



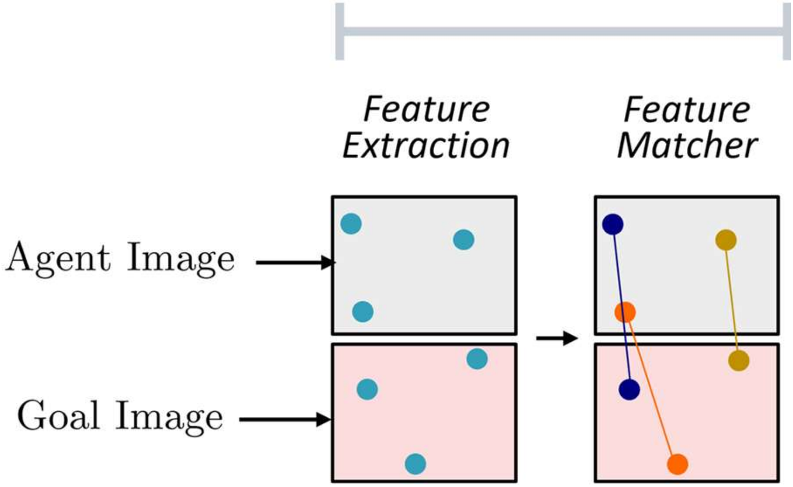
Goal Image



Our simple solution to Last-Mile Navigation



Our simple solution to Last-Mile Navigation



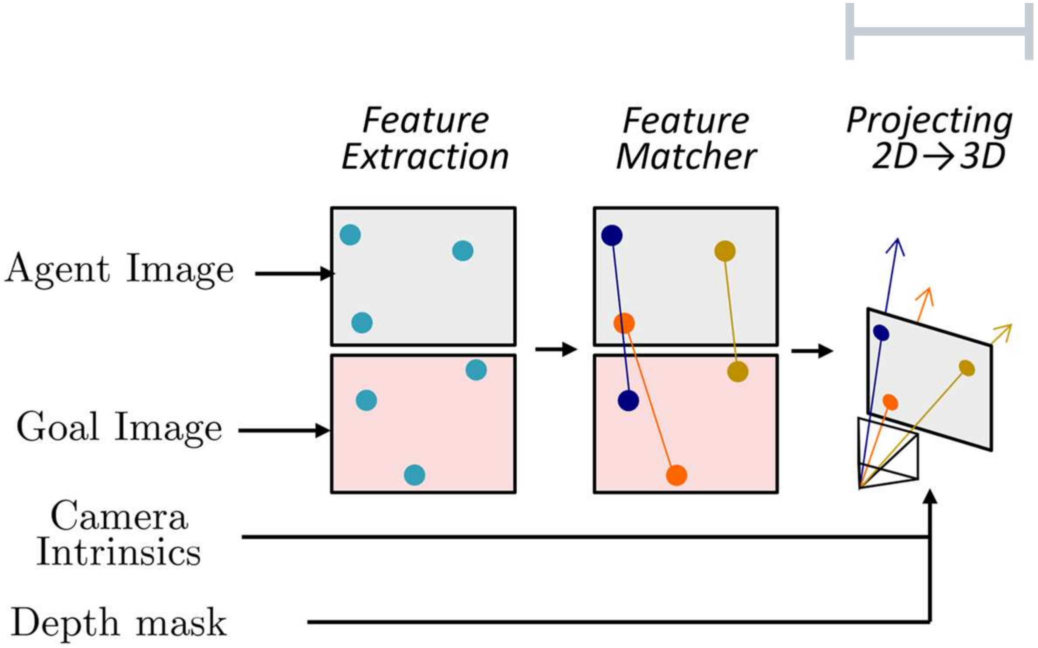
Feature Extraction

Feature Matcher

Agent Image

Goal Image

Our simple solution to Last-Mile Navigation



Feature Extraction

Feature Matcher

Projecting 2D→3D

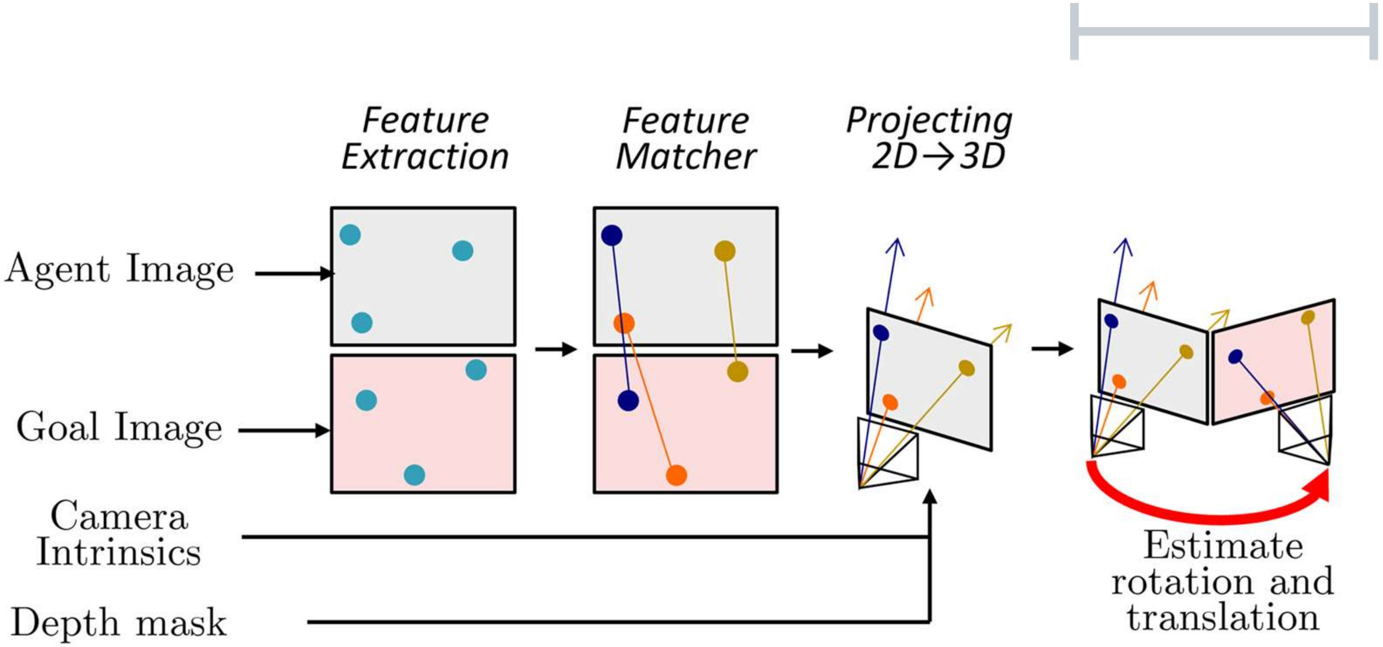
Agent Image

Goal Image

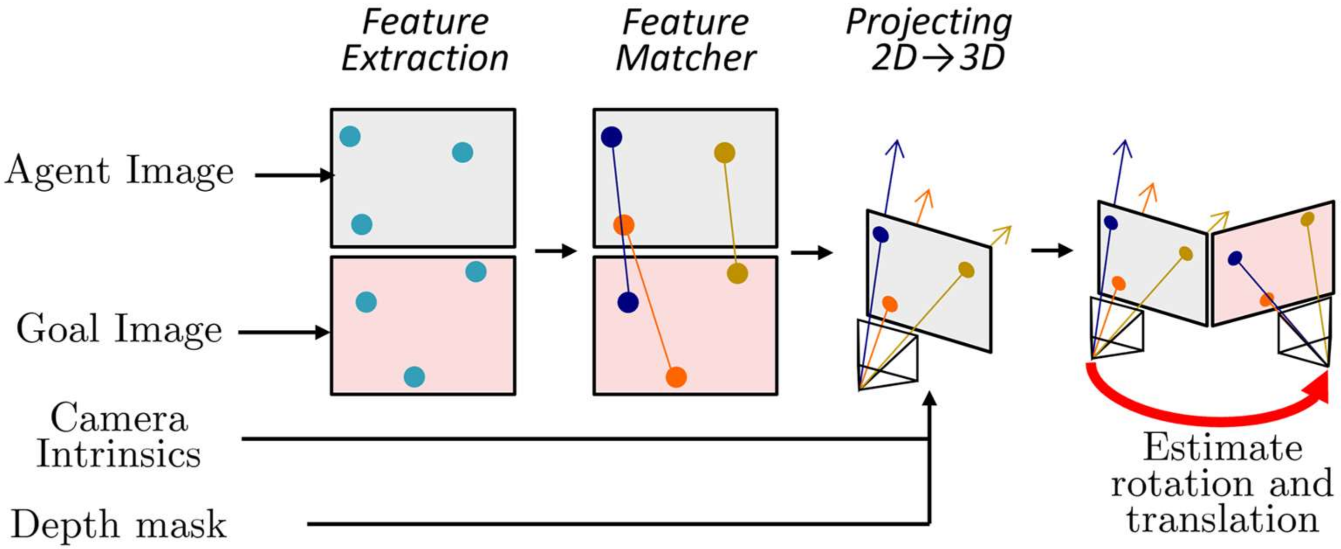
Camera Intrinsics

Depth mask

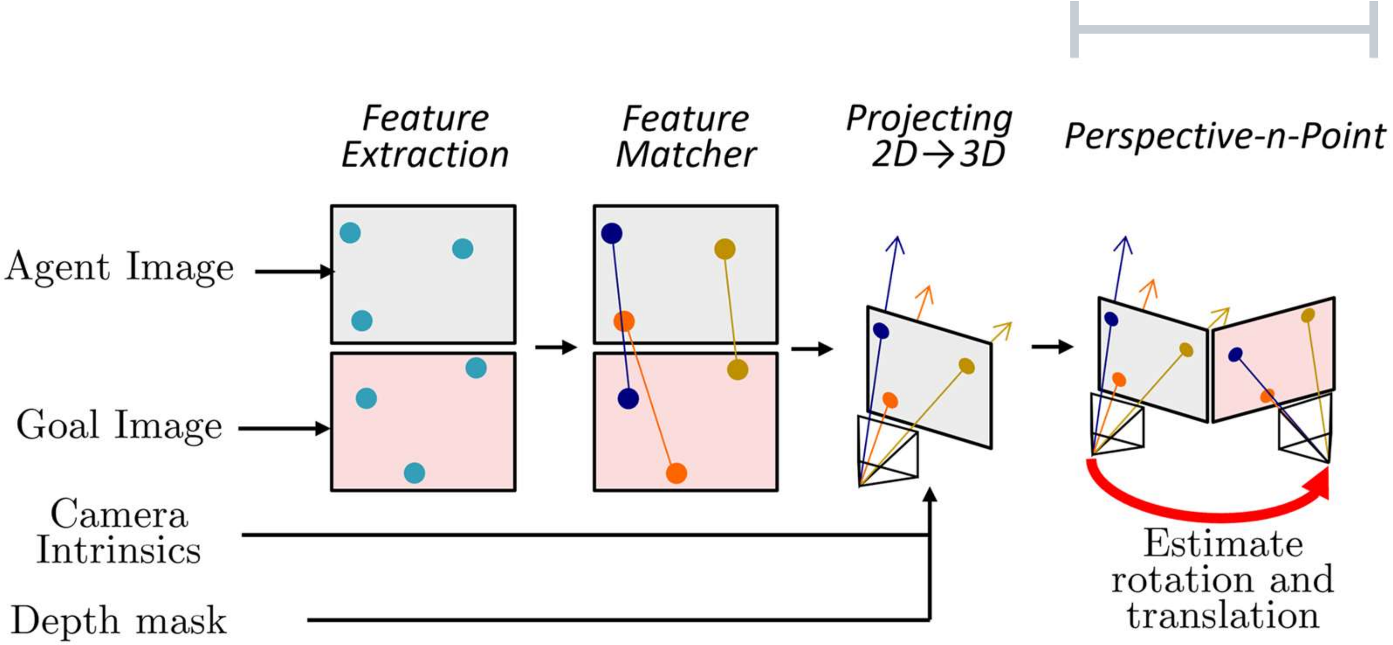
Our simple solution to Last-Mile Navigation



Our simple solution to Last-Mile Navigation



Our simple solution to Last-Mile Navigation

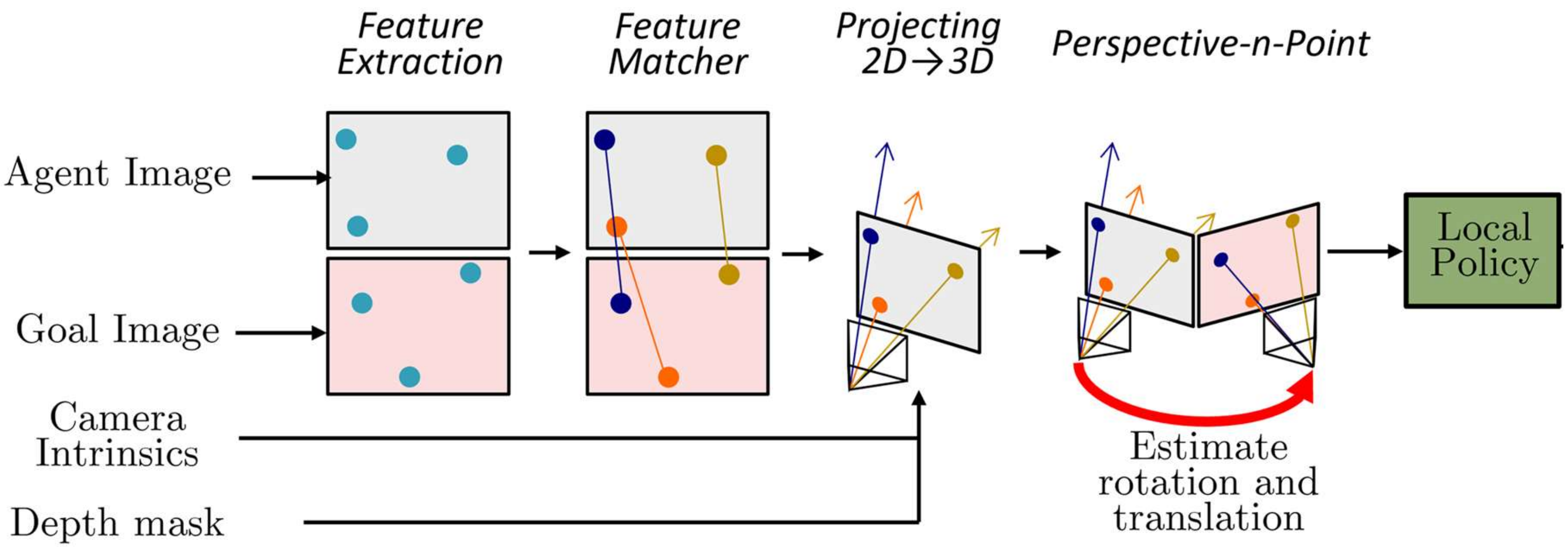


M. A. Fischler et al. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography.
ACM 1981

C.-P. Lu et al. Fast and globally convergent pose estimation from video images.
TPAMI 2000.

V. Lepetit et al. An accurate o(n) solution to the pnp problem.
IJCV 2009.

Our simple solution to Last-Mile Navigation



Feature Extraction

Feature Matcher

Projecting 2D→3D

Perspective-n-Point

Agent Image

Goal Image

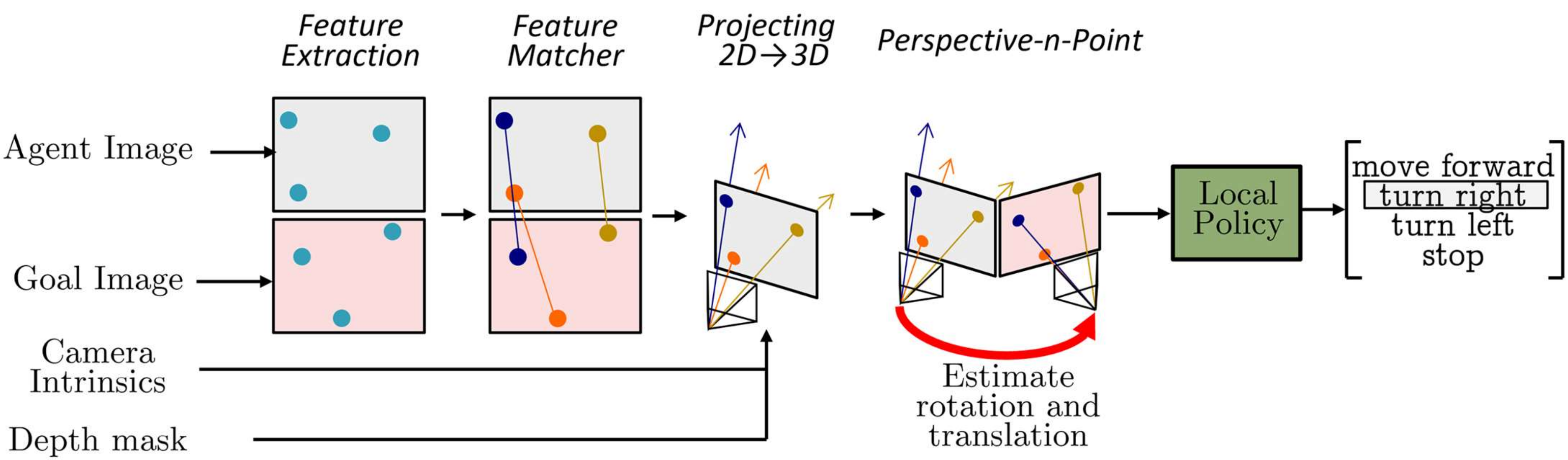
Camera Intrinsics

Depth mask

Estimate rotation and translation

Local Policy

Our simple solution to Last-Mile Navigation



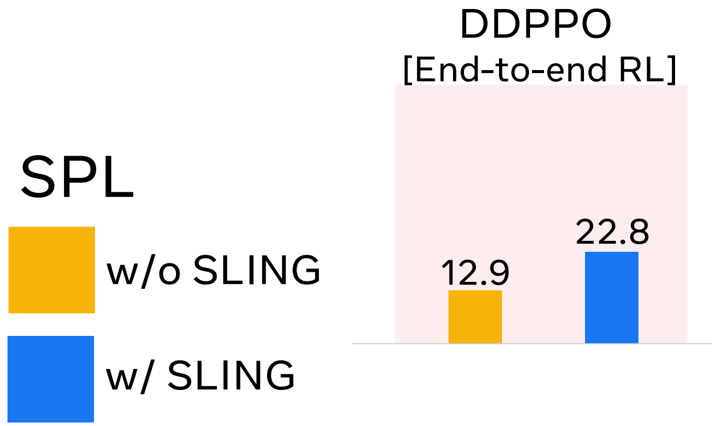
SOTA on Image-goal navigation in AIHabitat

SPL

🟨 w/o SLING

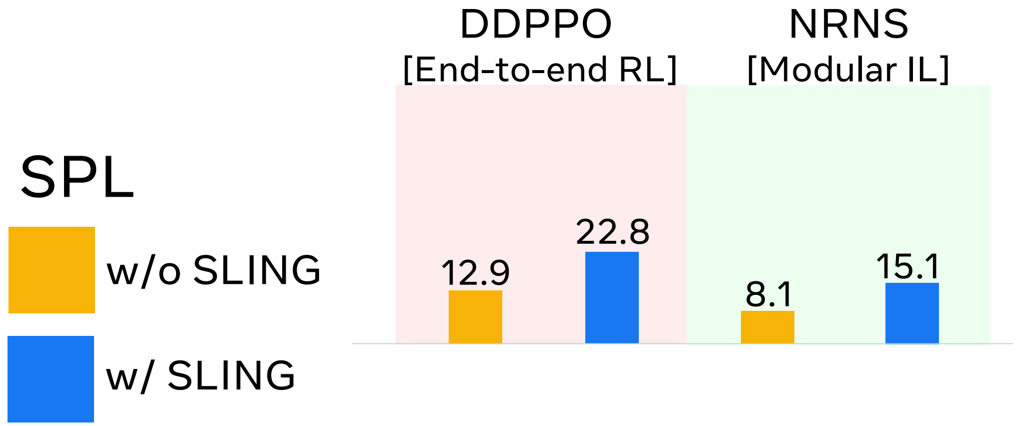
🟦 w/ SLING

SOTA on Image-goal navigation in AIHabitat



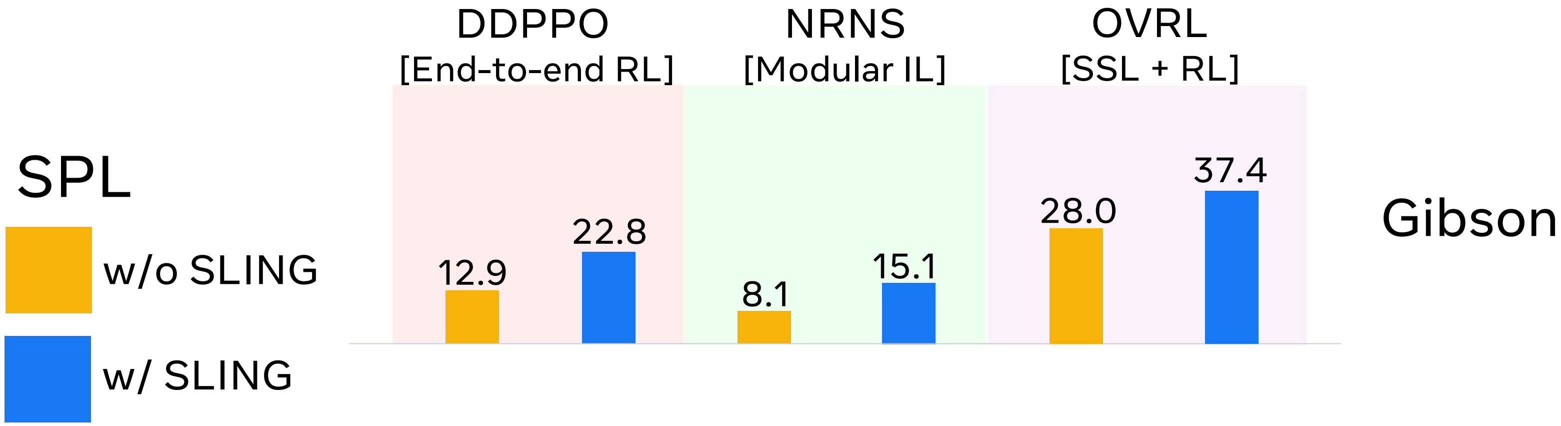
Gibson

SOTA on Image-goal navigation in AIHabitat

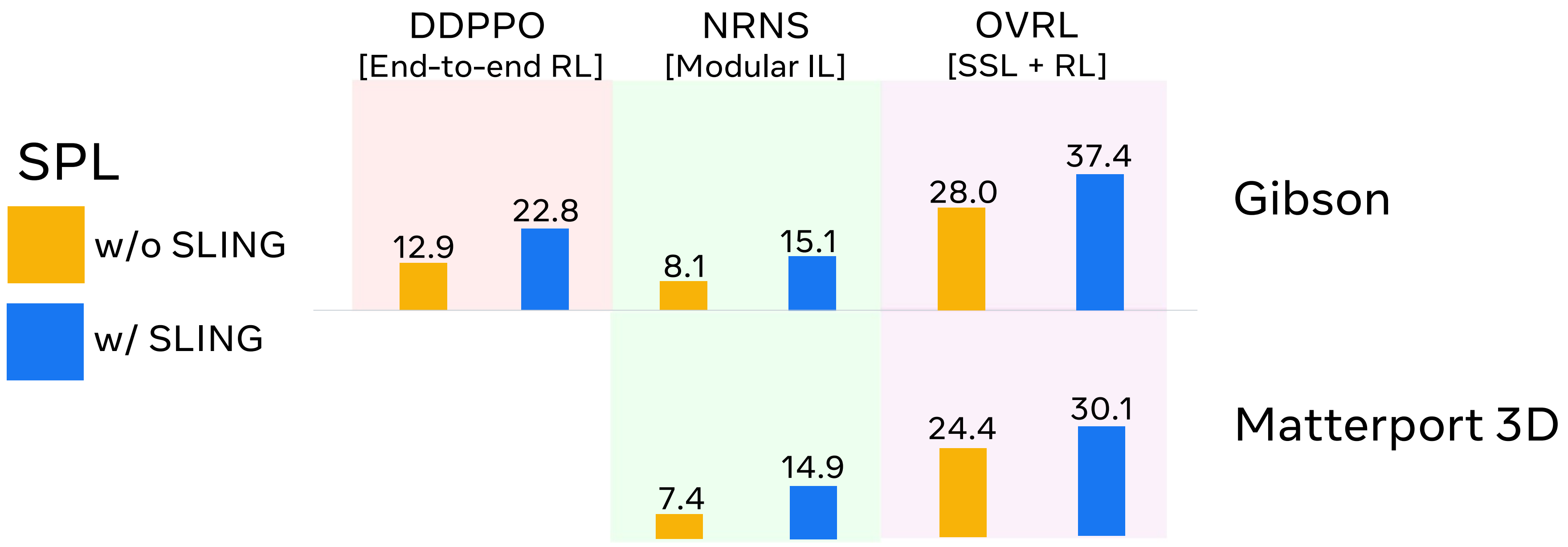


Gibson

SOTA on Image-goal navigation in AIHabitat

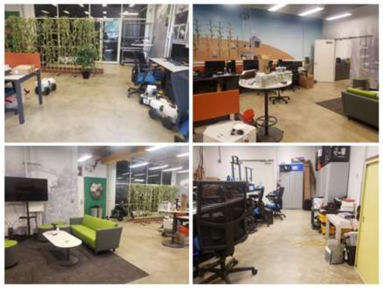


SOTA on Image-goal navigation in AIHabitat

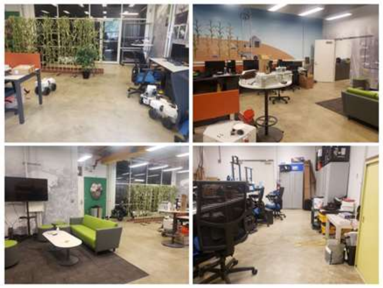


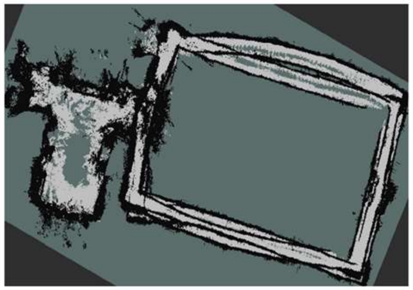
SLING transferred to a robot

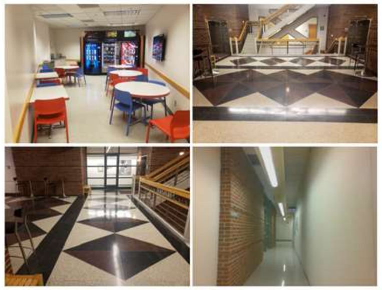
SLING transferred to a robot



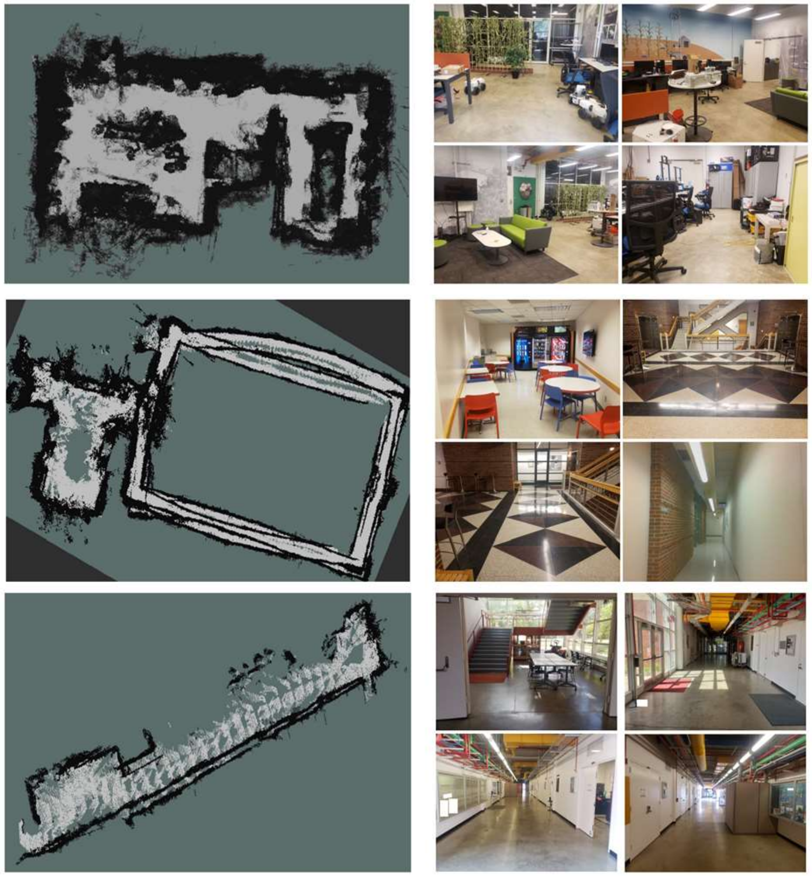
SLING transferred to a robot







SLING transferred to a robot



SLING transferred to a robot

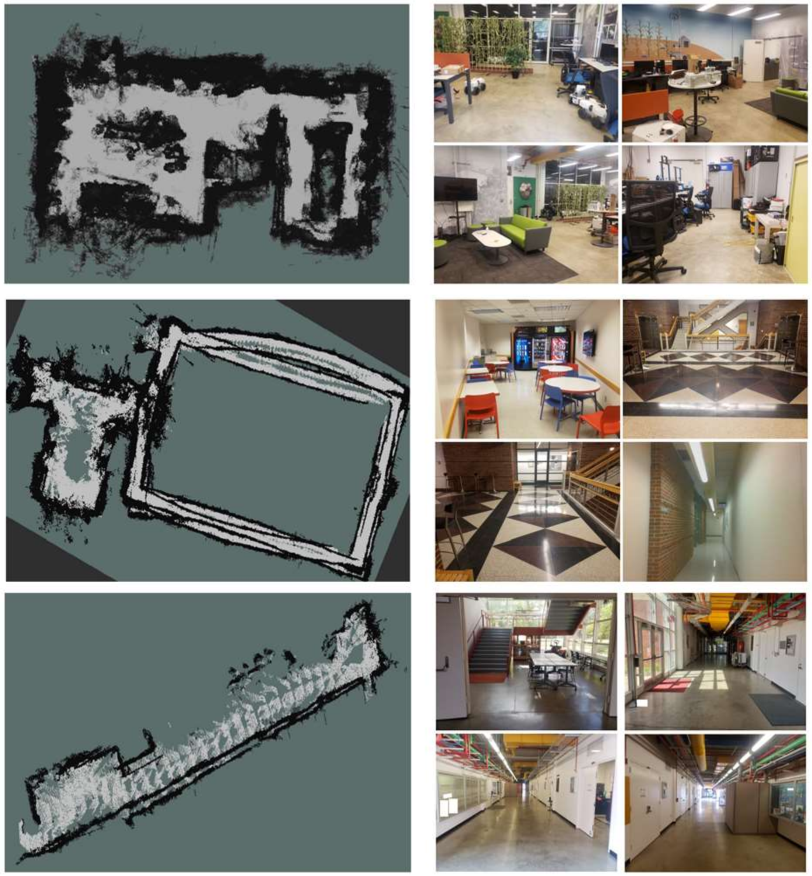
SPL

🟨 w/o SLING

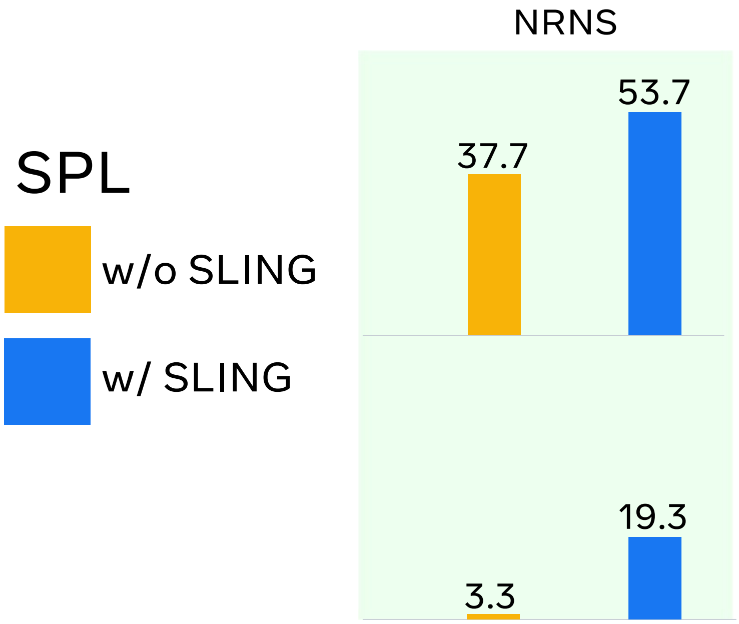
🟦 w/ SLING

Real-world, Easy
(1.5 – 3m)

Real-world, Hard
(5 – 10m)



SLING transferred to a robot



Real-world, Easy
(1.5 – 3m)

Real-world, Hard
(5 – 10m)

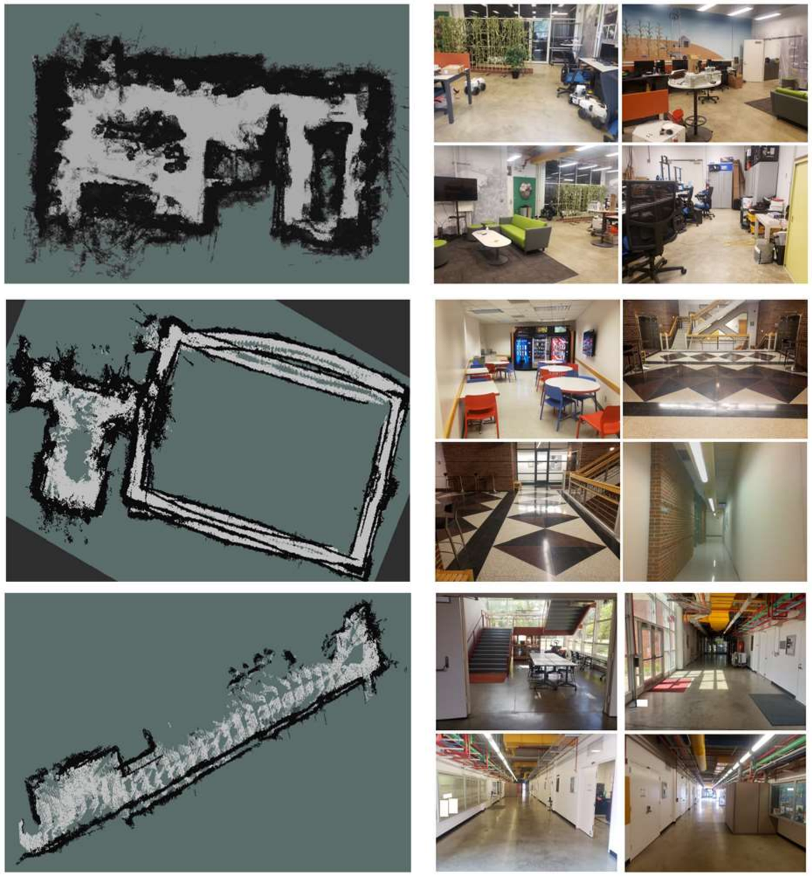


Image goal navigation in real life

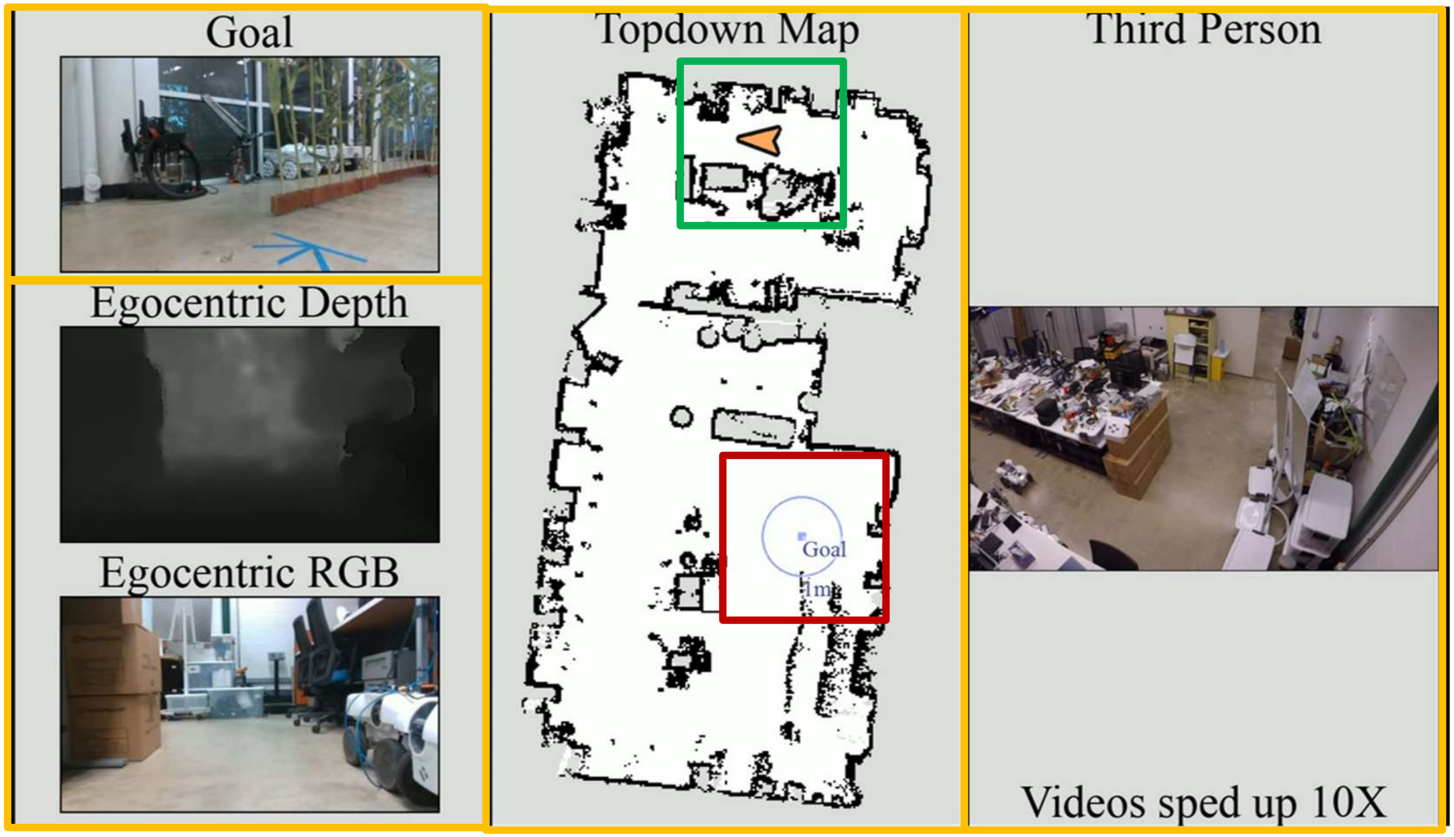
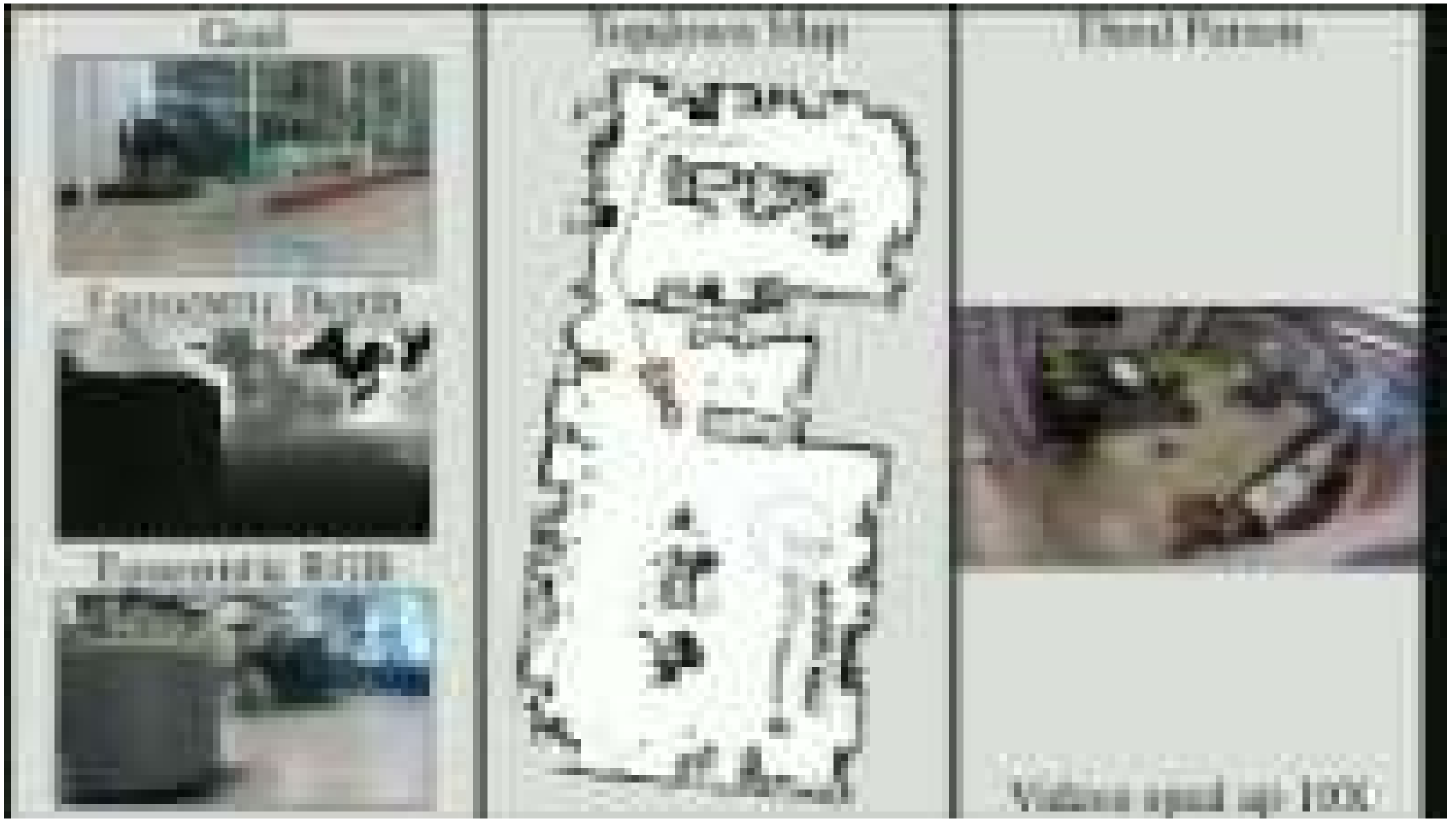
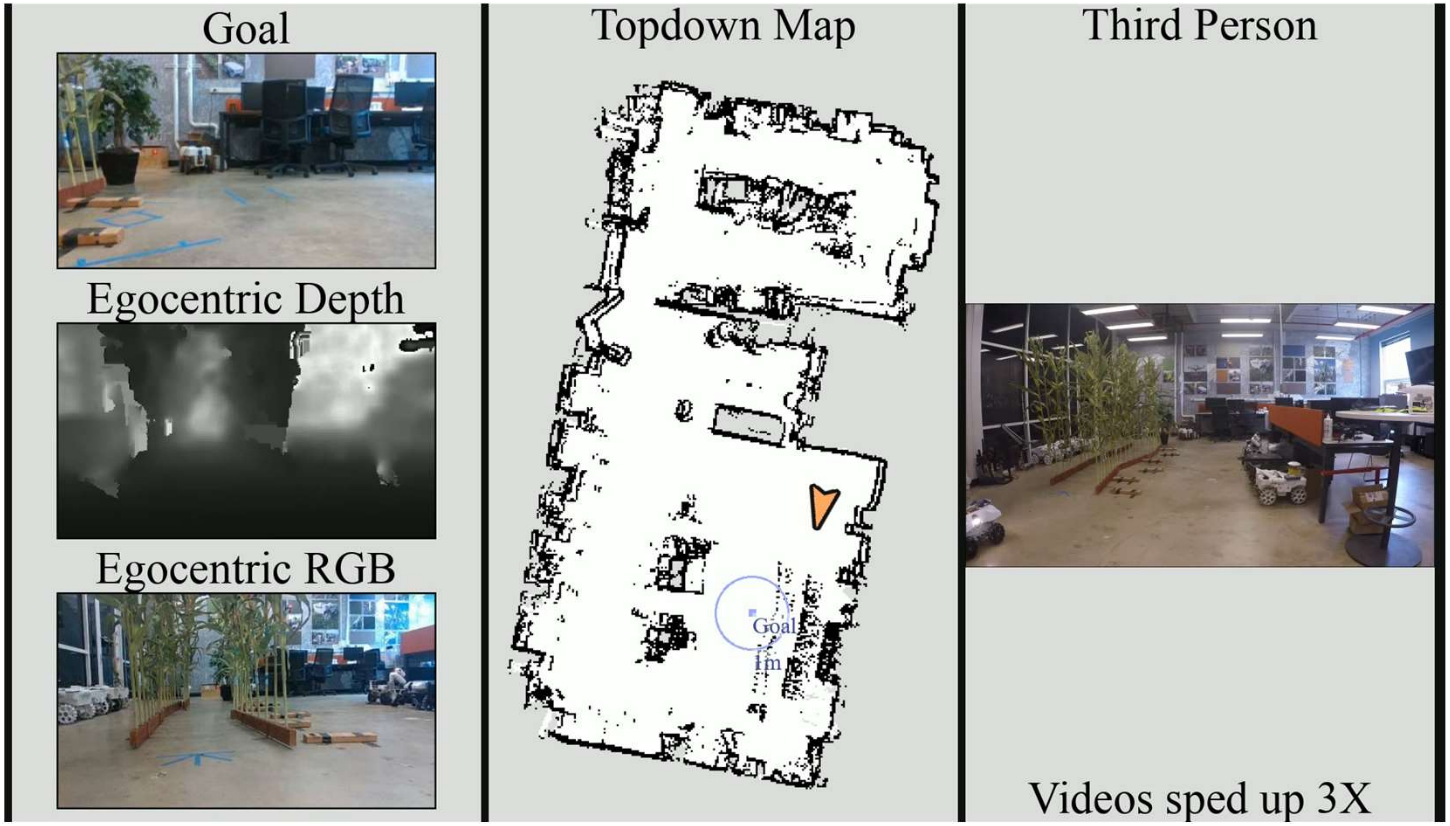


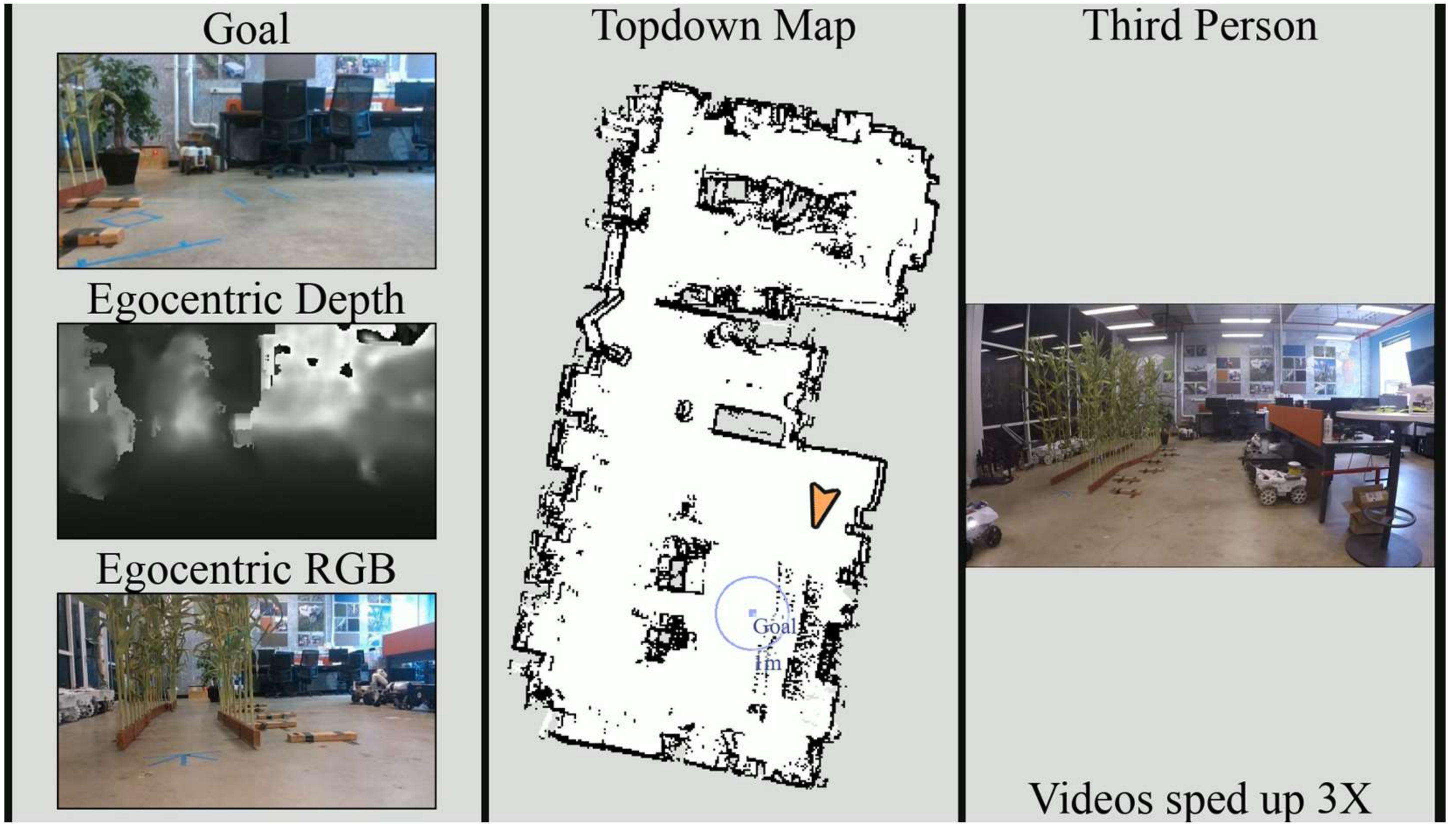
Image goal navigation in real life



NRNS for the last mile



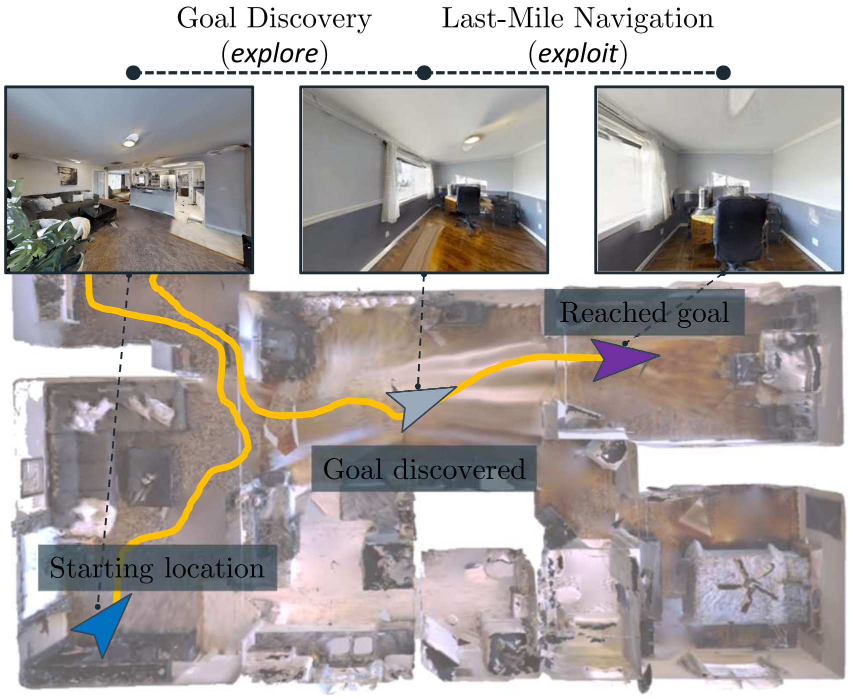
SLING for the last mile



Summary

Summary

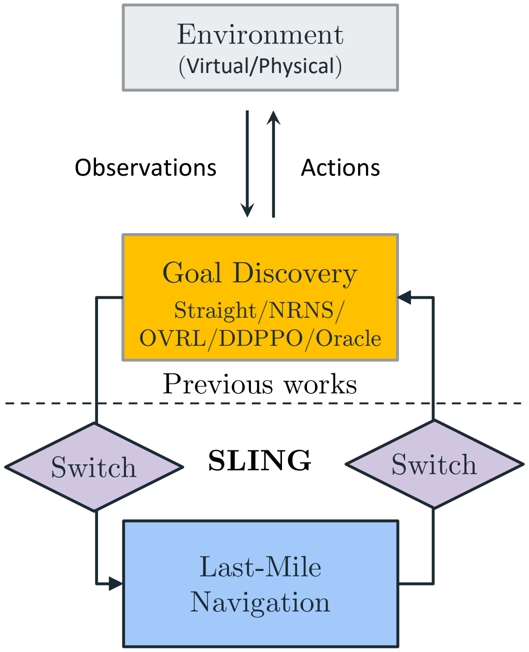
- ❖ Goal discovery and last mile navigation



Summary

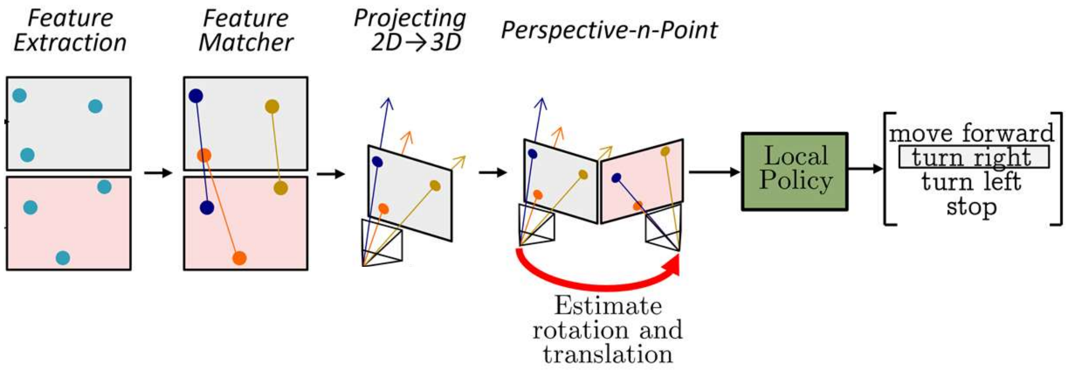
- Goal discovery and last mile navigation

- Switches to attach to prior baselines



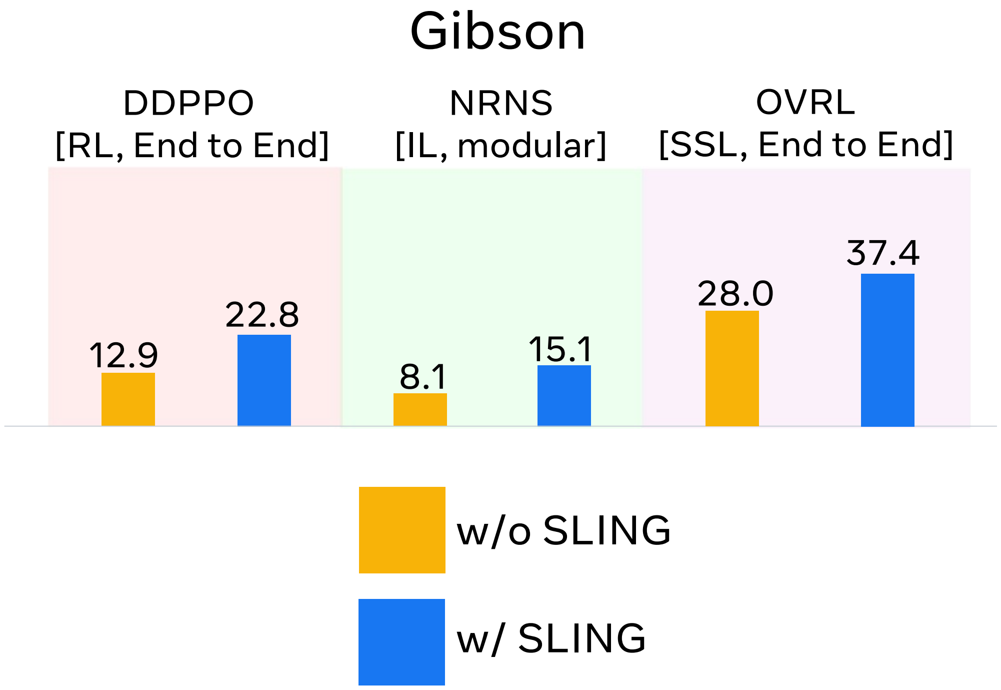
Summary

- Goal discovery and last mile navigation
- Switches to attach to prior baselines
- Principled 3D vision approach for last mile navigation



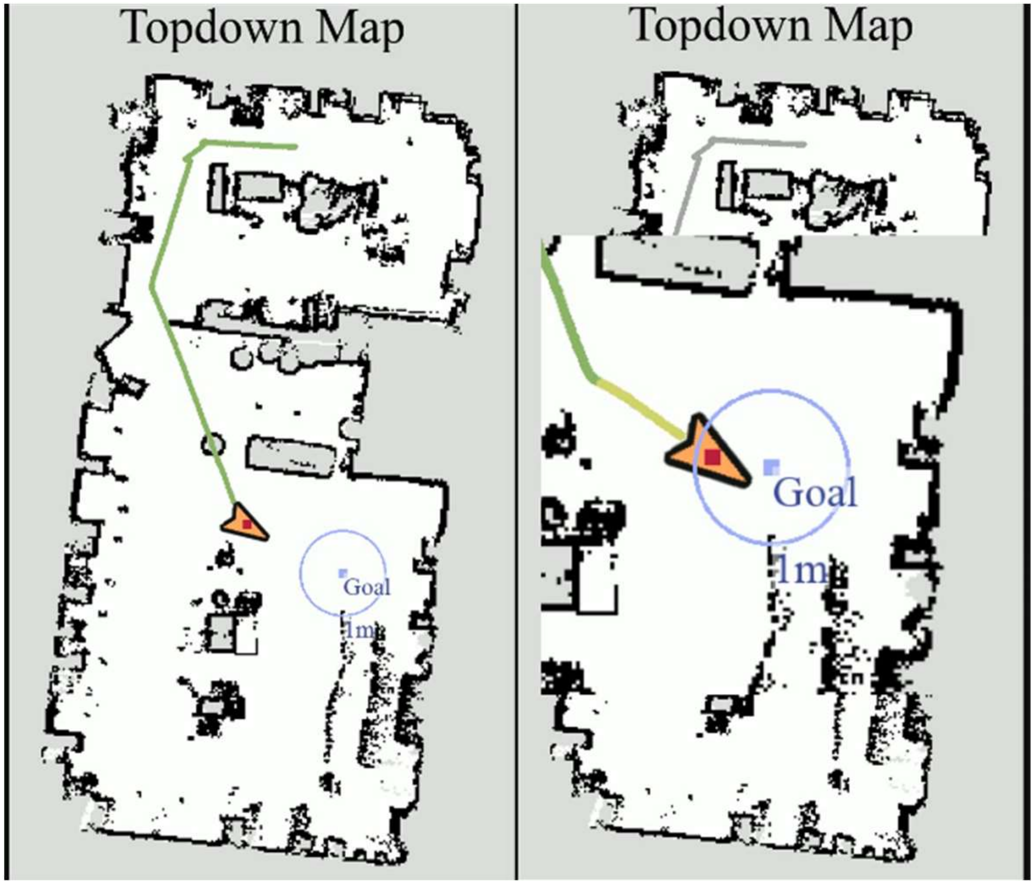
Summary

- Goal discovery and last mile navigation
- Switches to attach to prior baselines
- Principled 3D vision approach for last mile navigation
- Improved results across 5 baselines



Summary

- Goal discovery and last mile navigation
- Switches to attach to prior baselines
- Principled 3D vision approach for last mile navigation
- Improved results across 5 baselines
- Transfer to the real world


Topdown Map
Goal
1m
Topdown Map
Goal
1m