Embodied Question Answering in Photorealistic Environments with Point Cloud Perception

Erik Wijmans*

Samyak Datta*

Oleksandr Maksymets*

Abhishek Das

Georgia Gkioxari

Stefan Lee

Irfan Essa

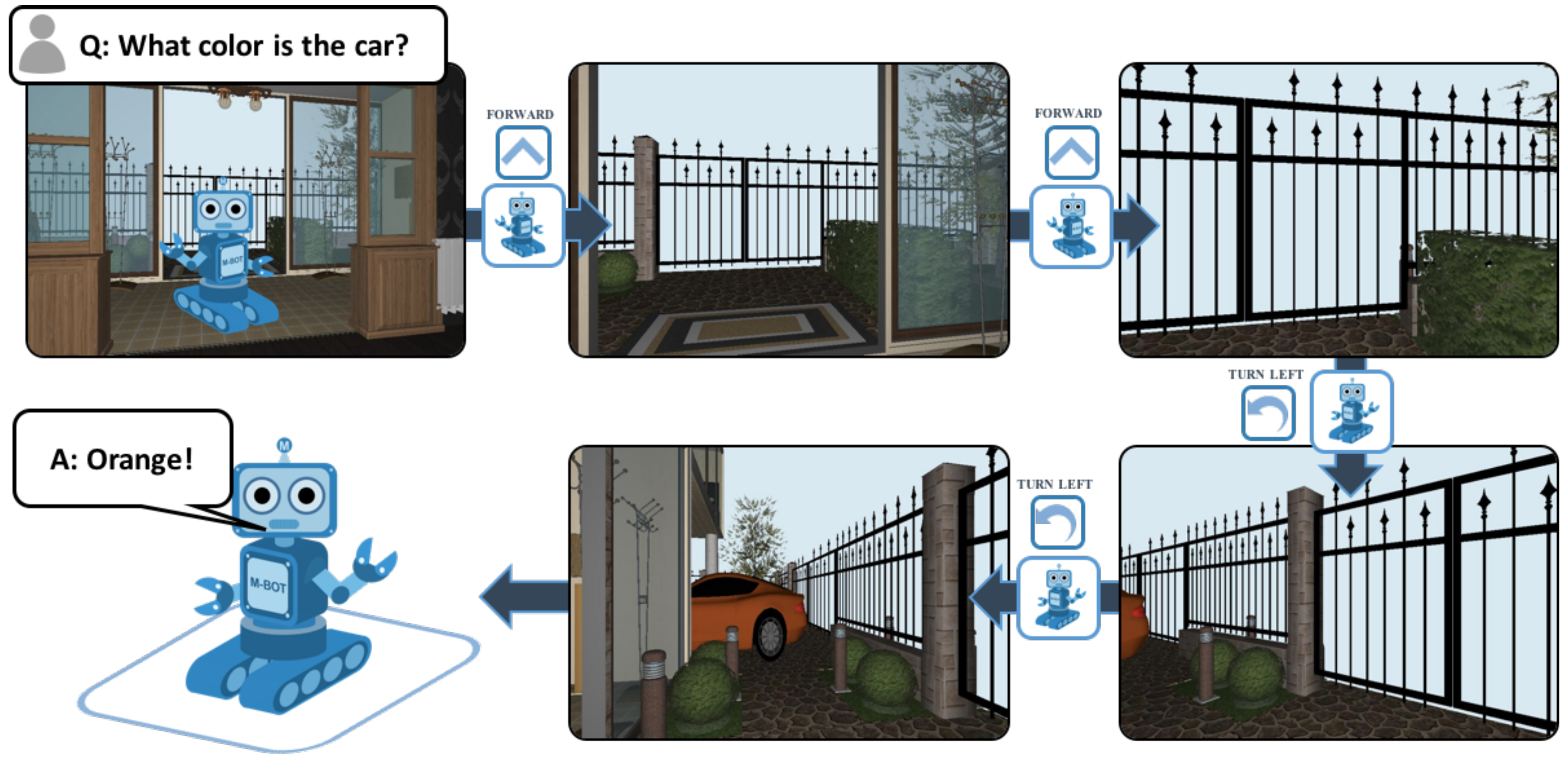
Devi Parikh

Dhruv Batra

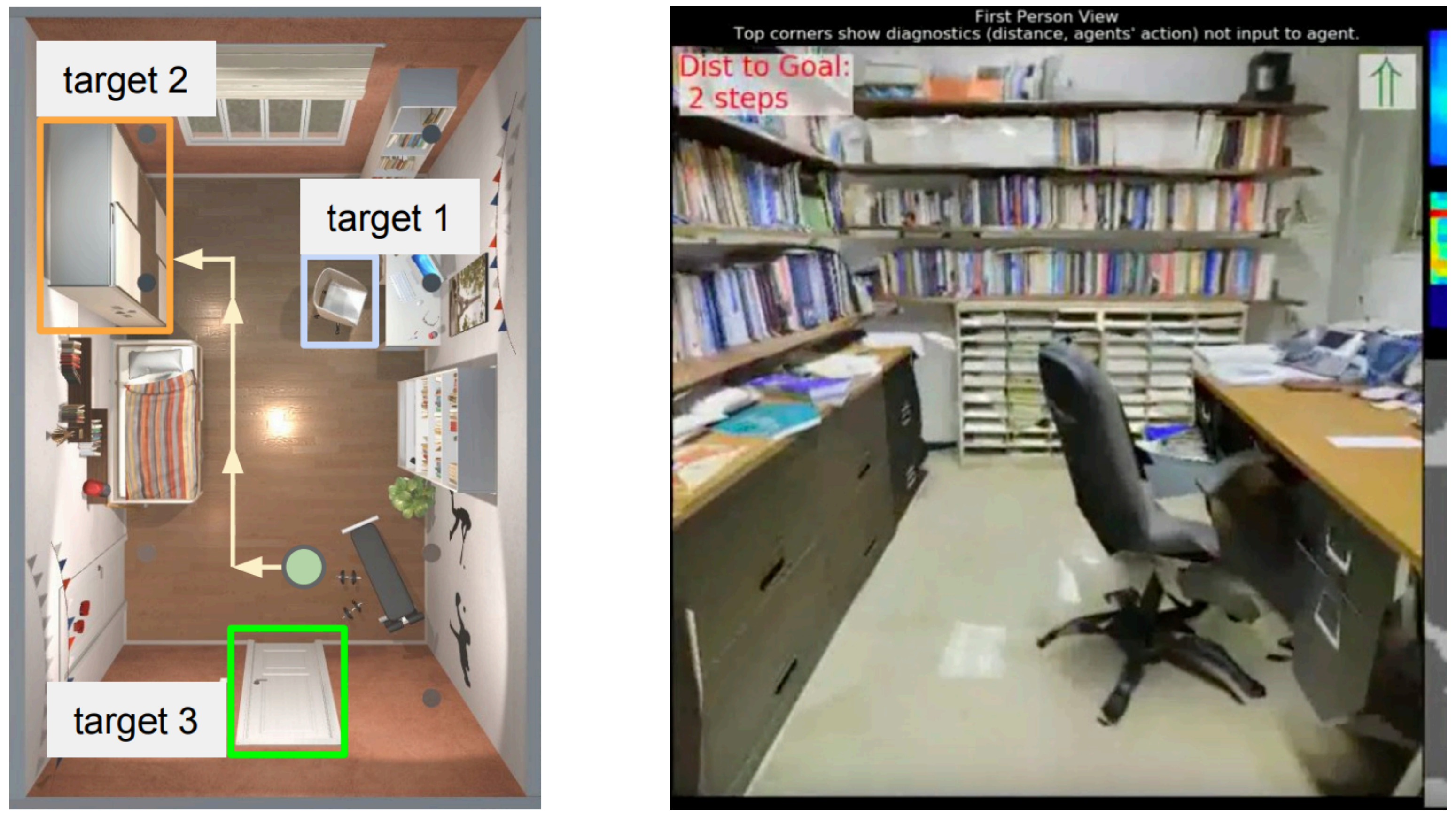
Georgia Tech

facebook
Artificial Intelligence Research

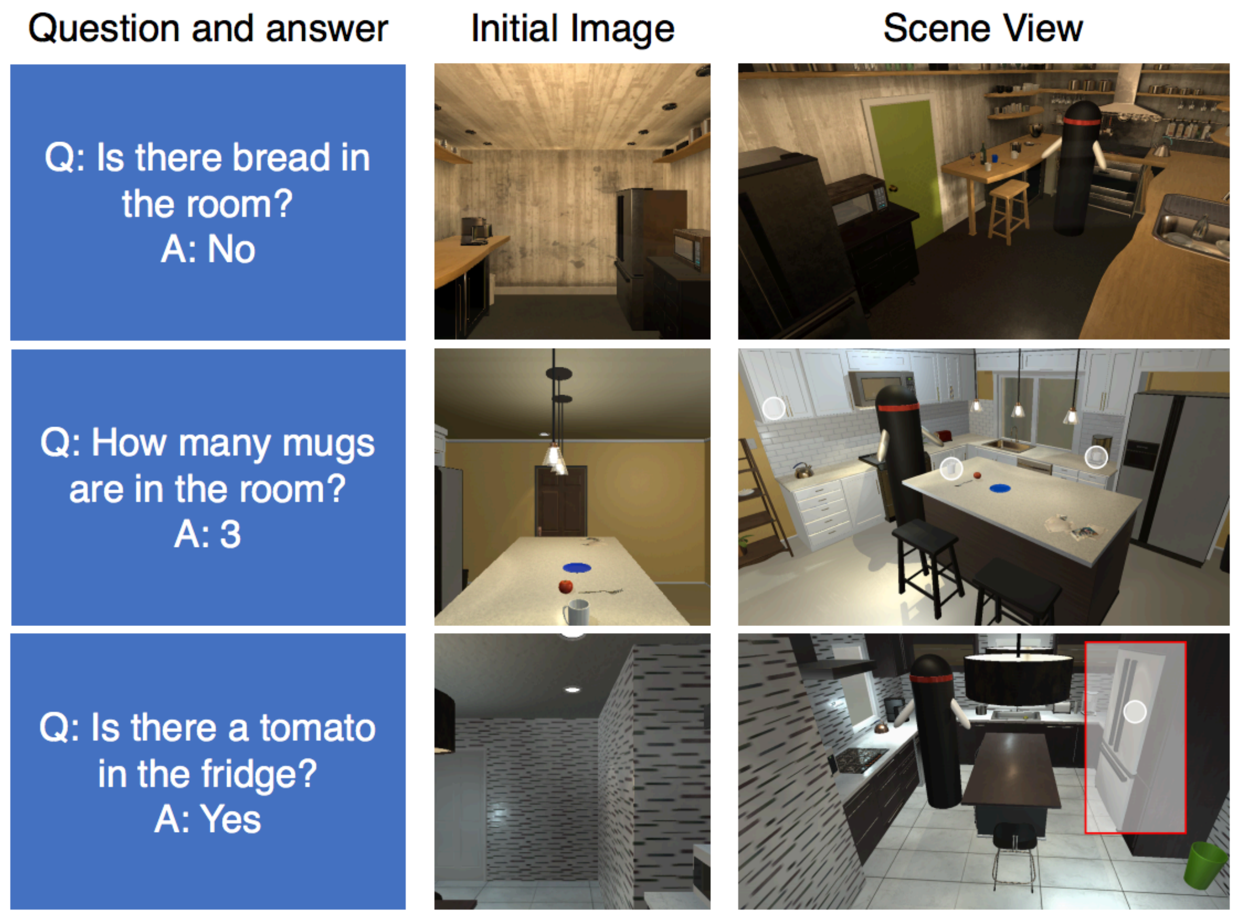
* Denotes equal contribution



EmbodiedQA
(Das et al., 2018)


target 2
target 1
target 3
First Person View
Top corners show diagnostics (distance, agents' action) not input to agent.
Dist to Goal:
2 steps

Visual Navigation
(Zhu et al., 2017, Gupta et al., 2017)



Question and answer | Initial Image | Scene View

Q: Is there bread in the room?
A: No

Q: How many mugs are in the room?
A: 3

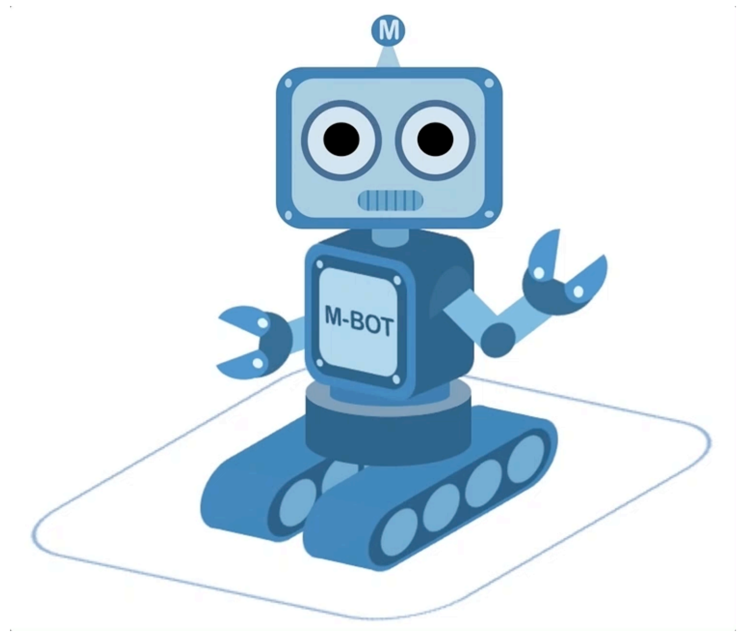
Q: Is there a tomato in the fridge?
A: Yes

Interactive QA
(Gordon et al., 2018)

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Vision-Language Navigation
(Anderson et al., 2018)

Current differences from reality



Perfect actuations

Perfect odometry

RGB only perception

⋮

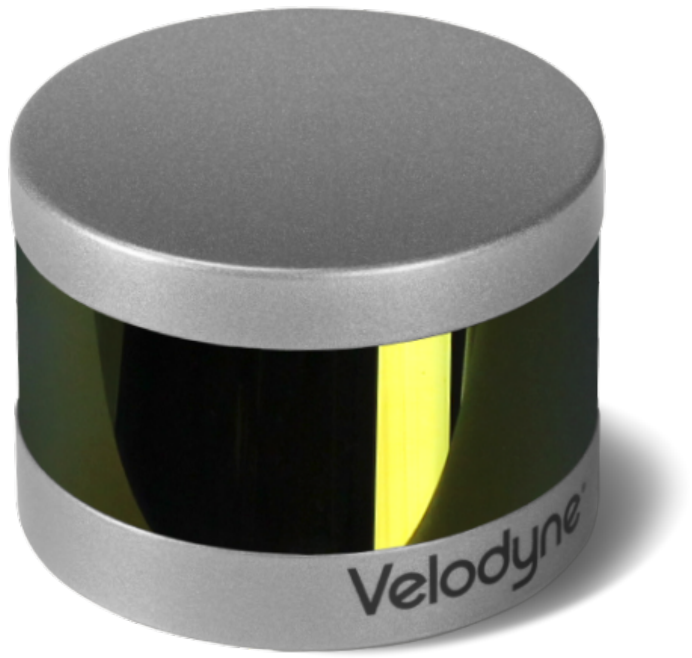
Noisy actuations

Noisy odometry

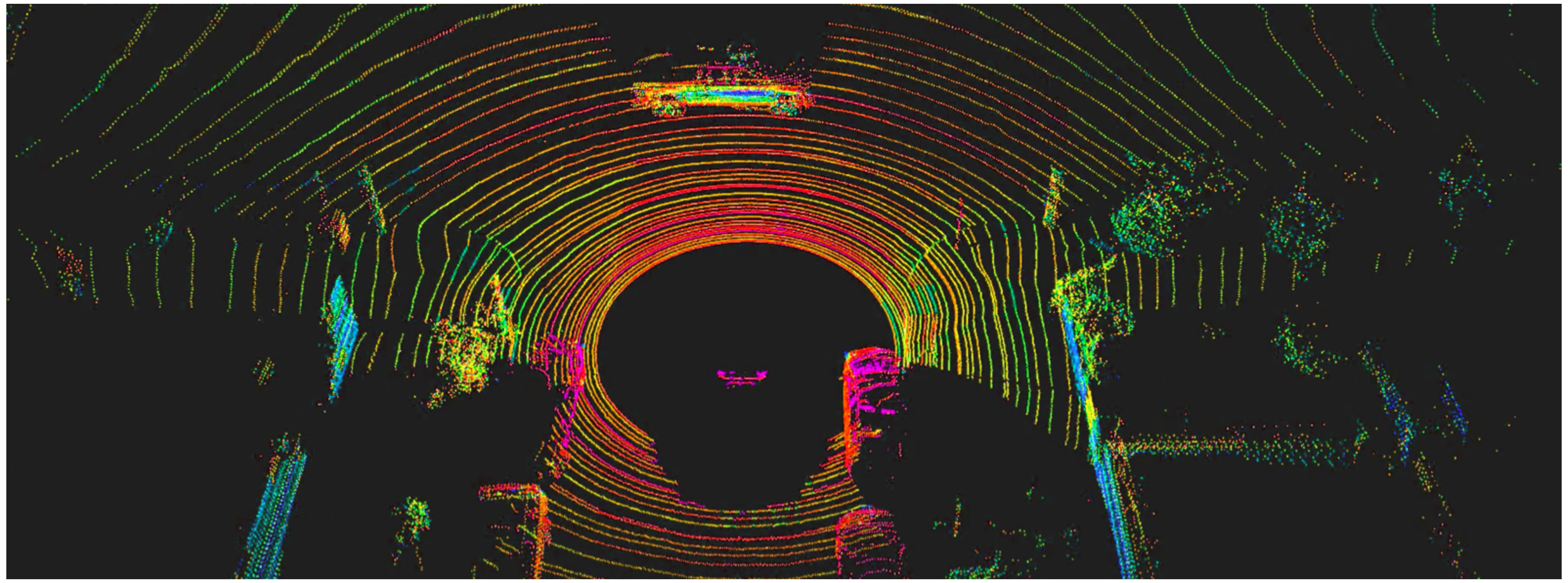
Multiple perceptual modalities

⋮

Depth information via point clouds



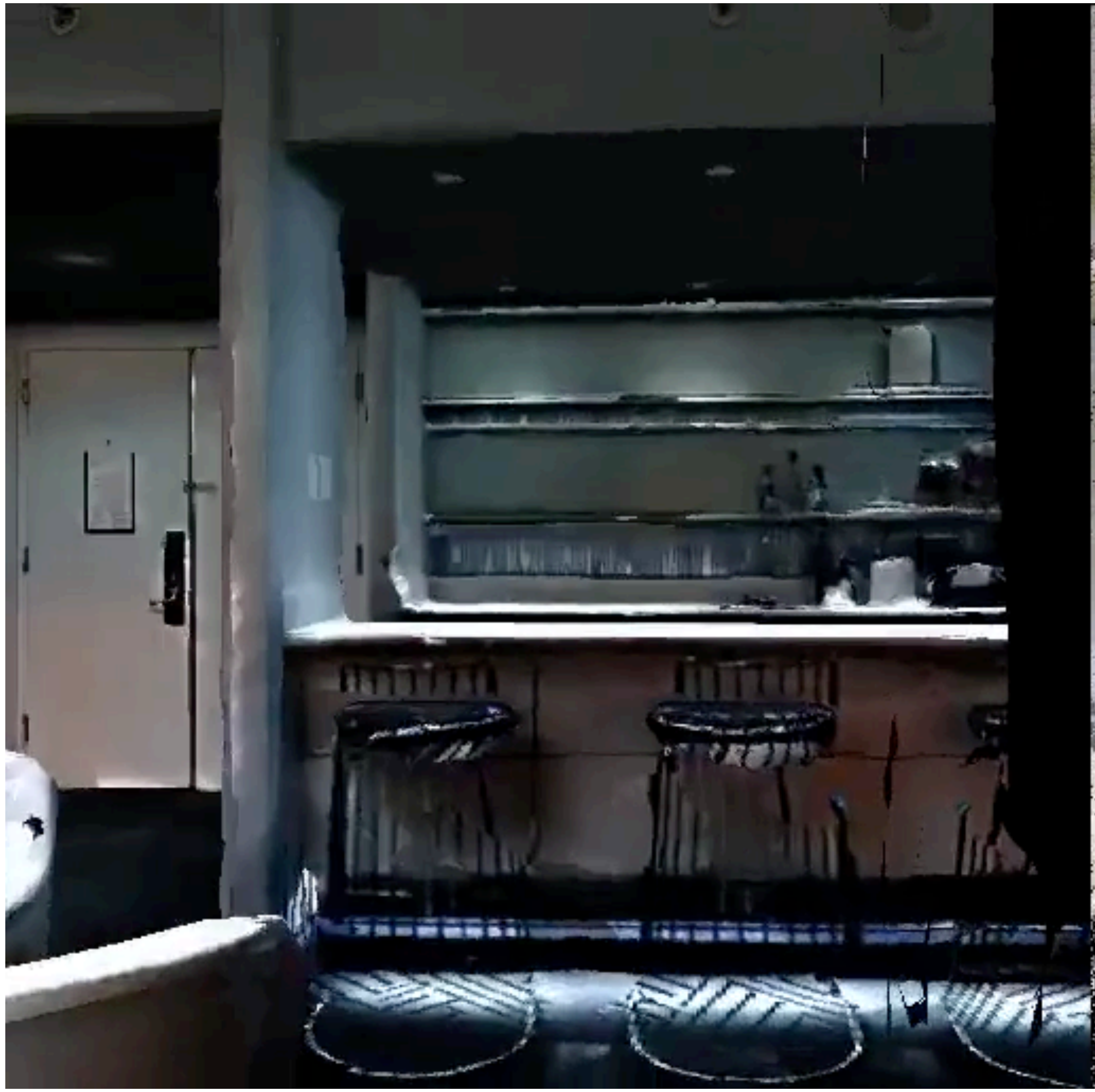
Lidar



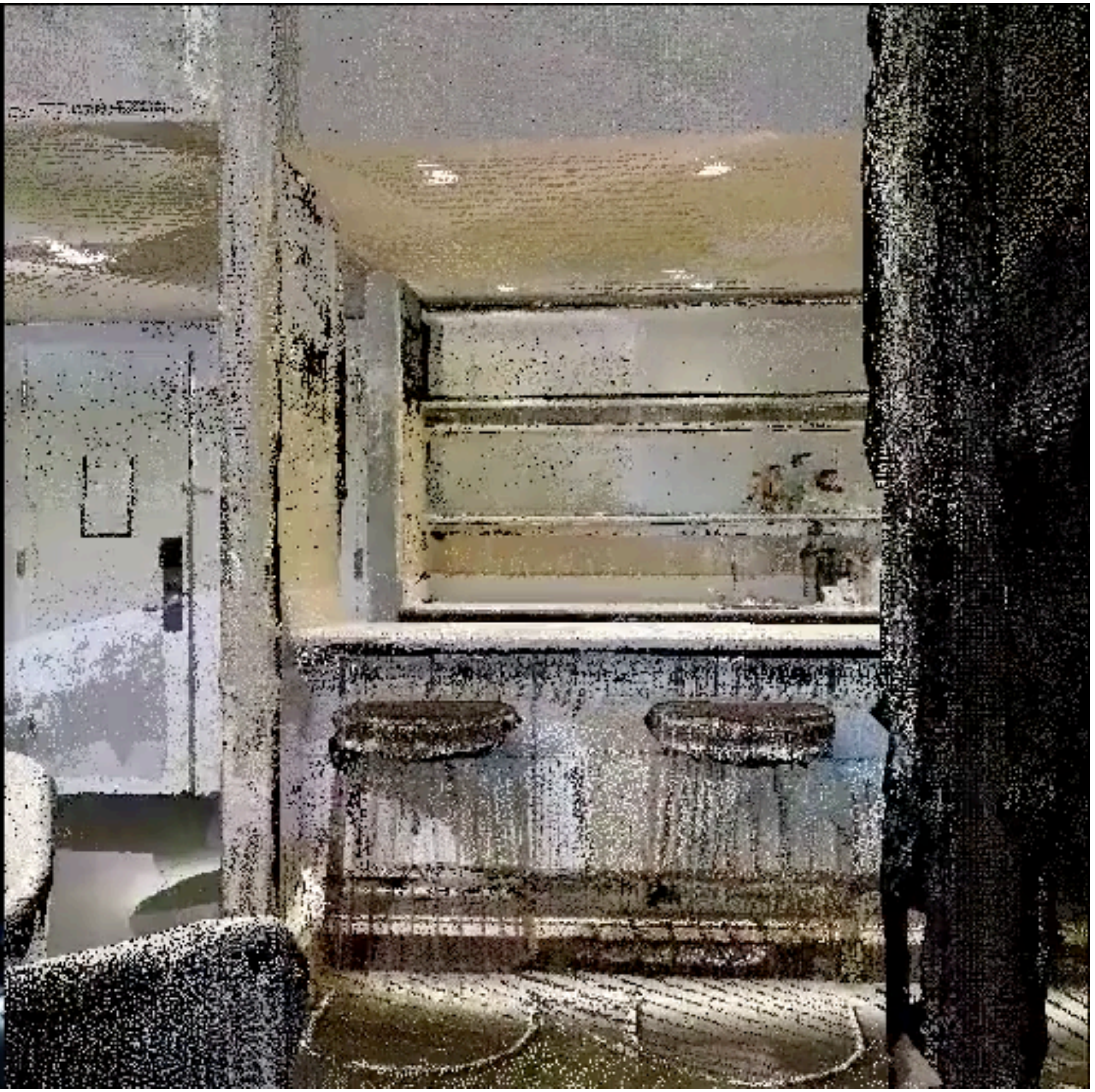
Structured Light + RGB

Perception for Embodied Agents

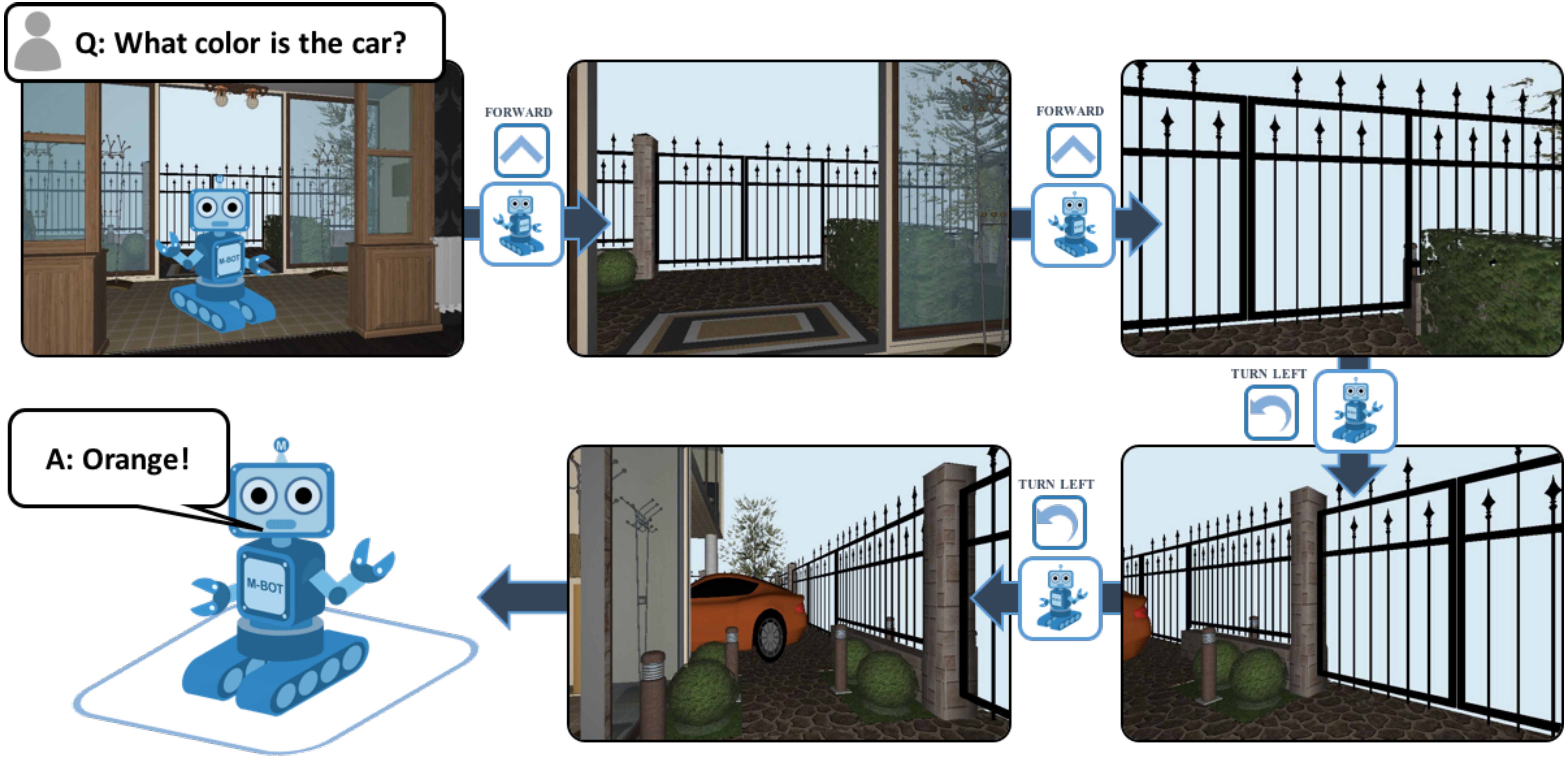
RGB Perception



Point Cloud Perception



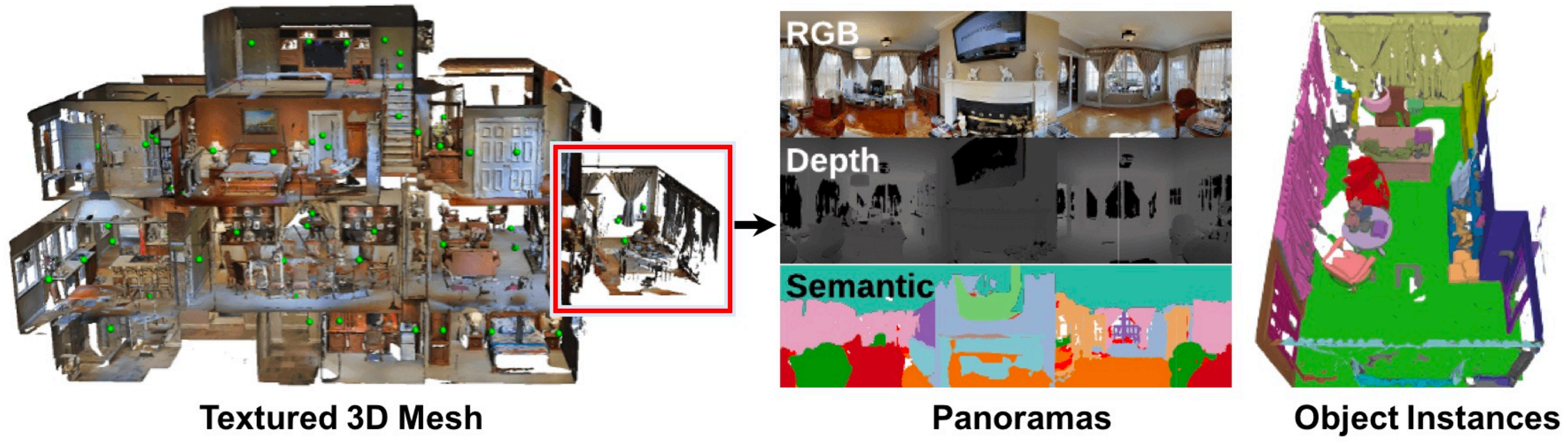
EmbodiedQA



EmbodiedQA: Das et. al., 2018

The EQA Matterport Dataset

Built using scenes from Matterport3D



1136 Questions

83 Houses

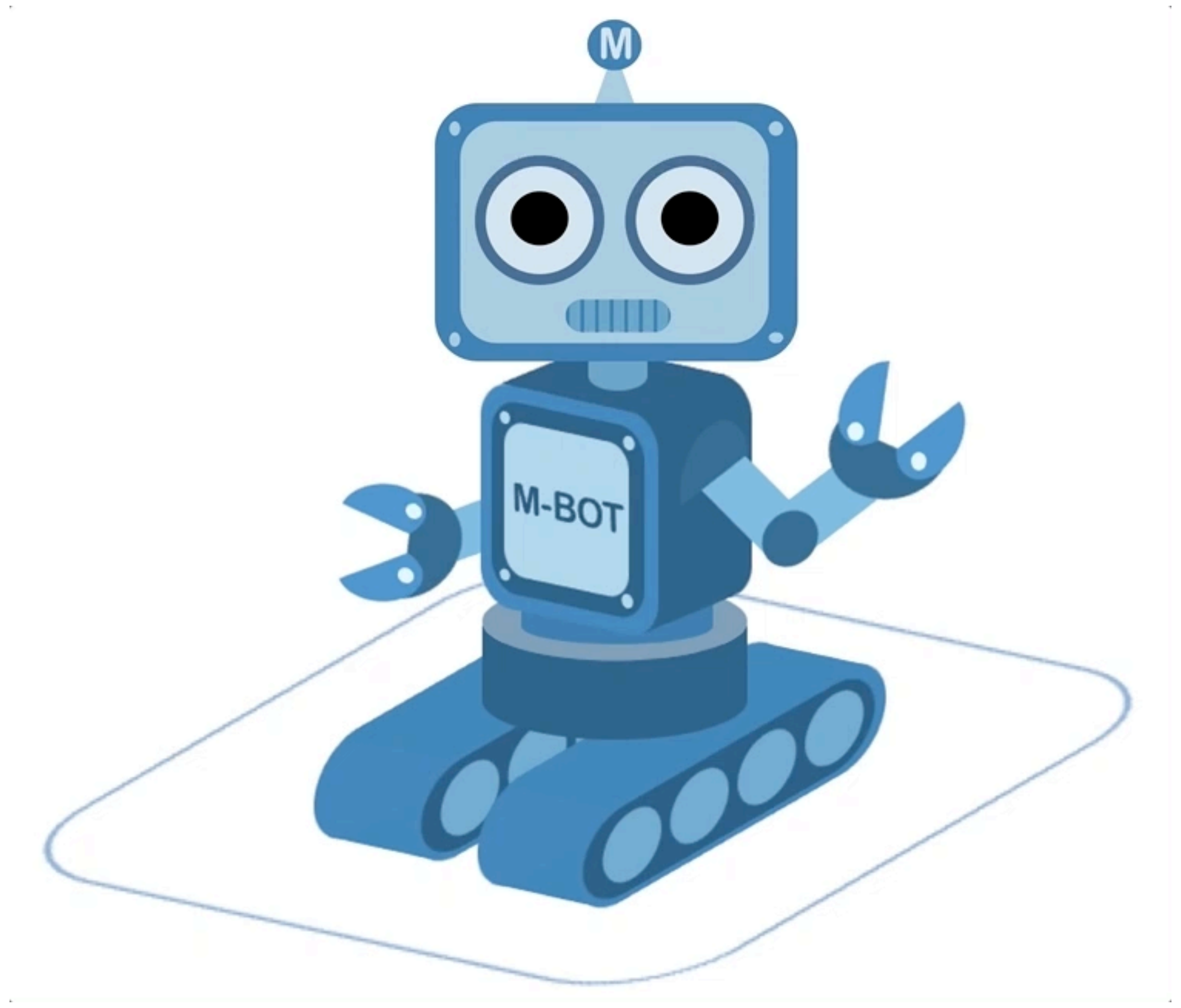
146 Floors

Matterport3D: Chang et. al., 2017

- Human color names for objects
- Careful selection of goal locations to ensure target of question is visible

- Generate point clouds from scanner data, not mesh

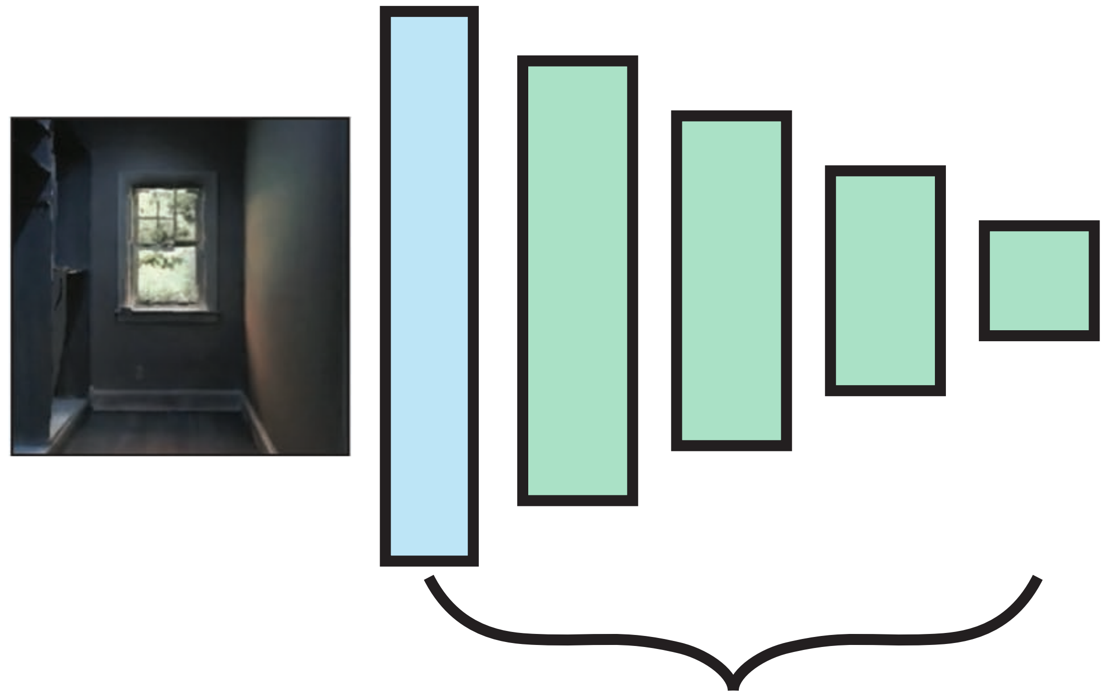
Agents for EmbodiedQA



Agents for EmbodiedQA

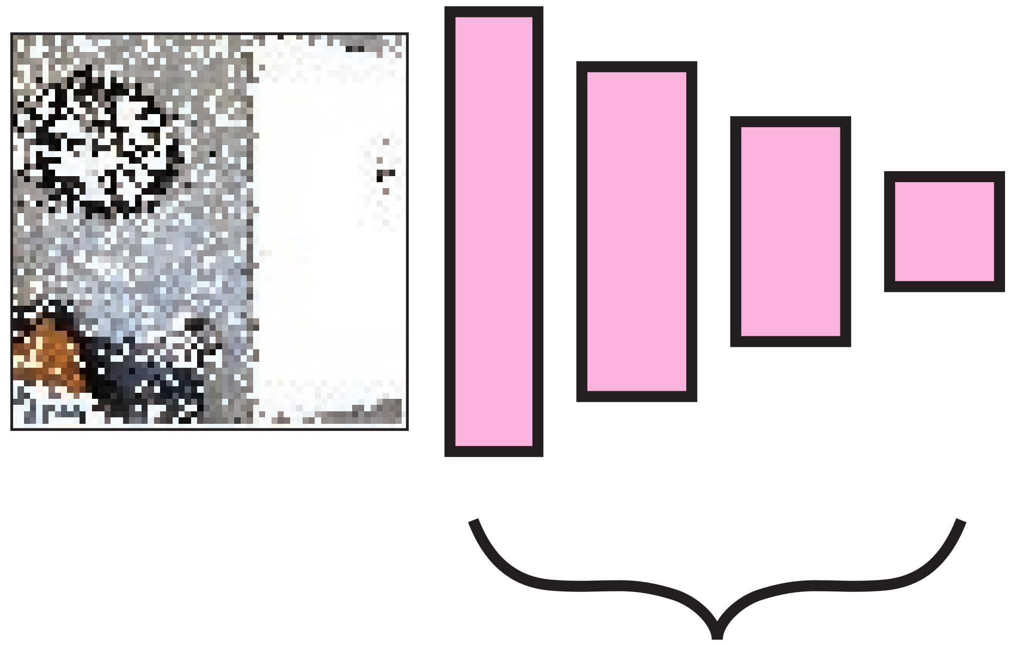
Vision, Navigation, and Answering

Perception for RGB

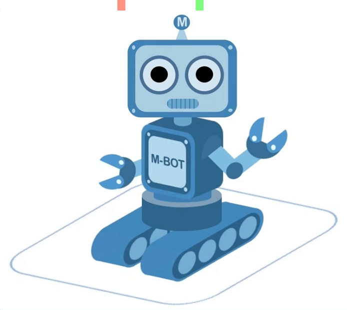


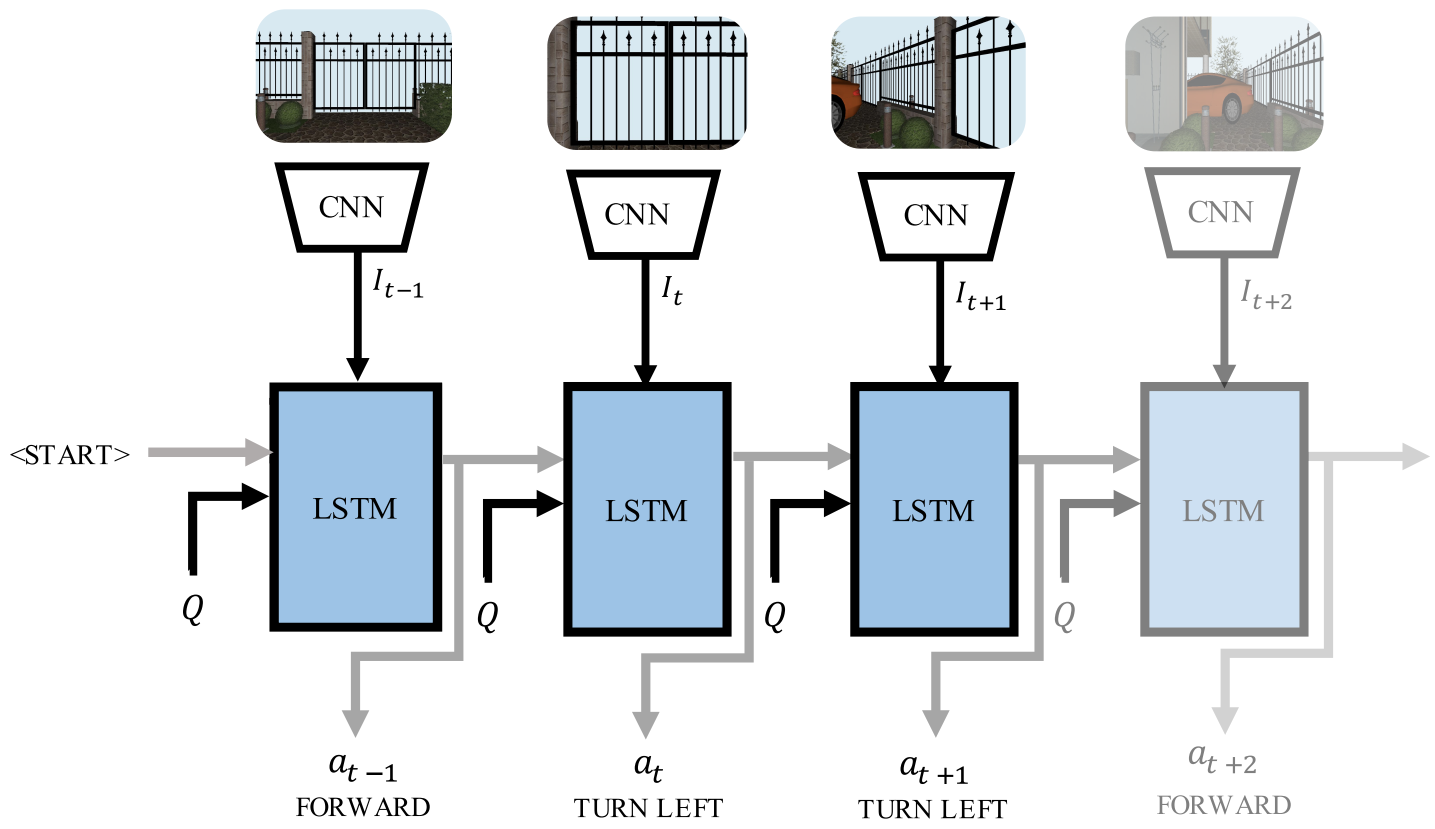
ResNet50 Encoder

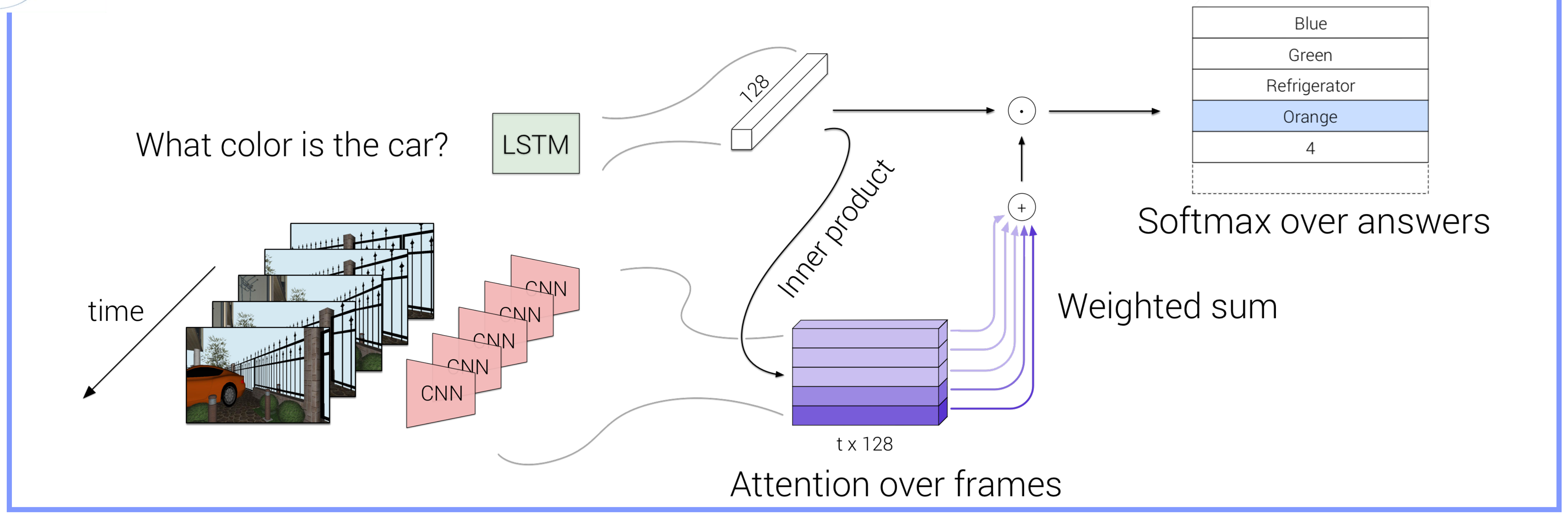
Perception for Point Clouds



PointNet++ Encoder



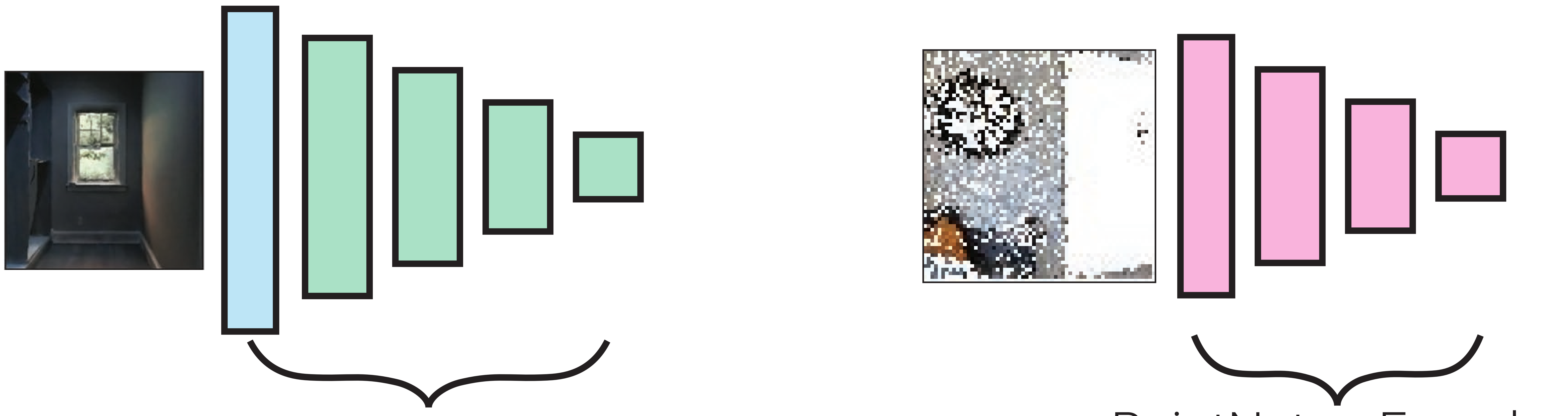




PointNet++: Qi et. al., 2017

Agents for EmbodiedQA

Vision



Perception for RGB

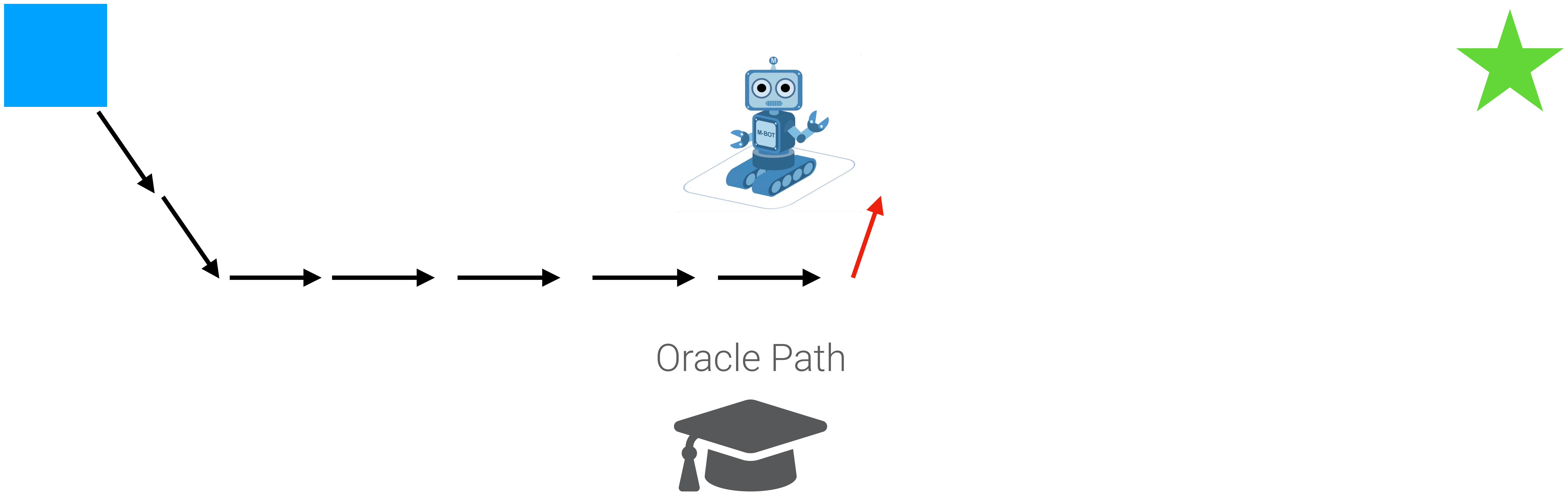
ResNet50 Encoder

Perception for Point Clouds

PointNet++ Encoder

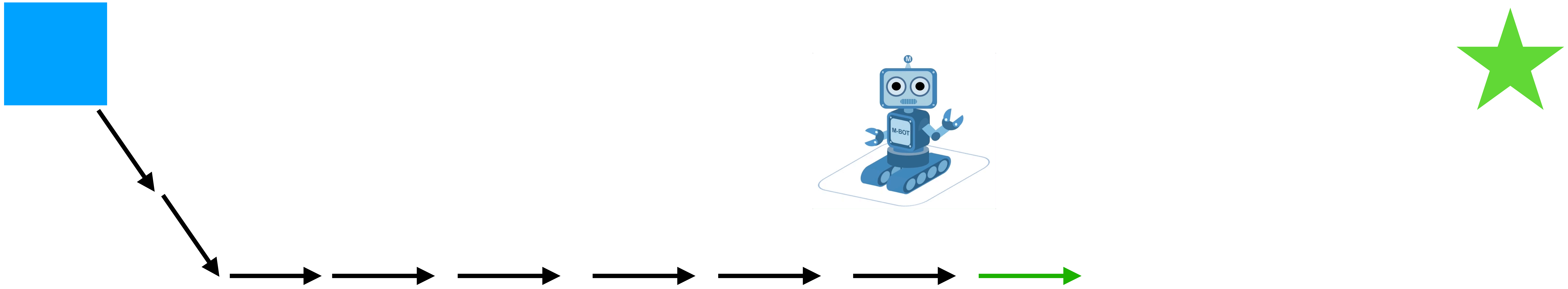
PointNet++: Qi et. al., 2017

Training for navigation



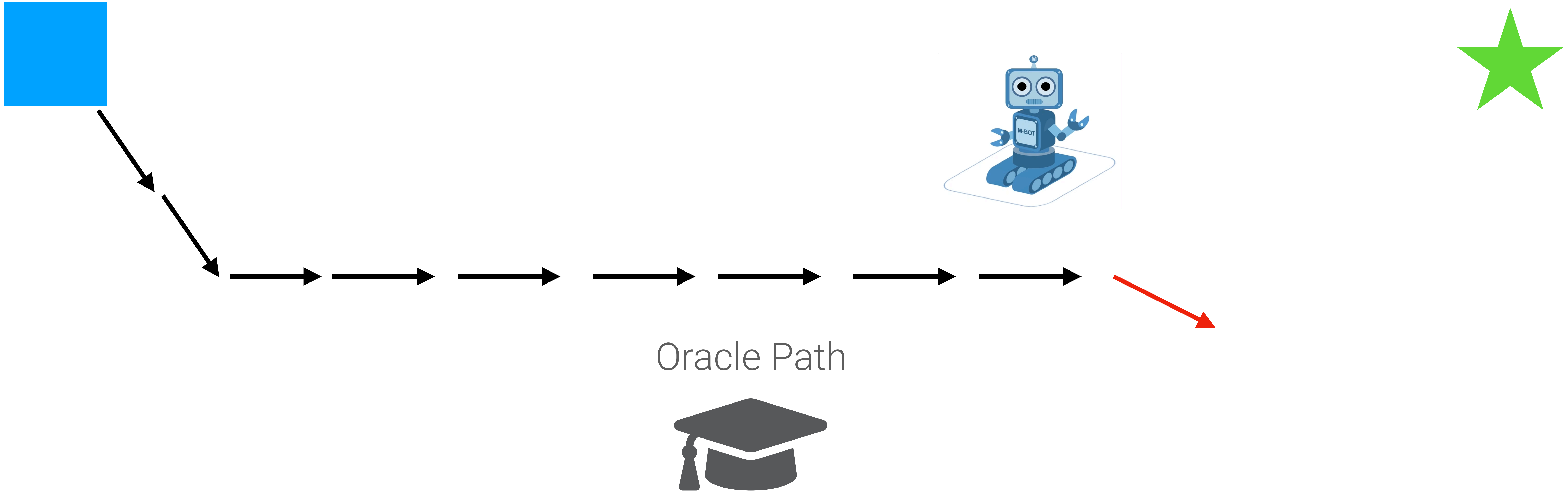
Oracle Path

Training for navigation



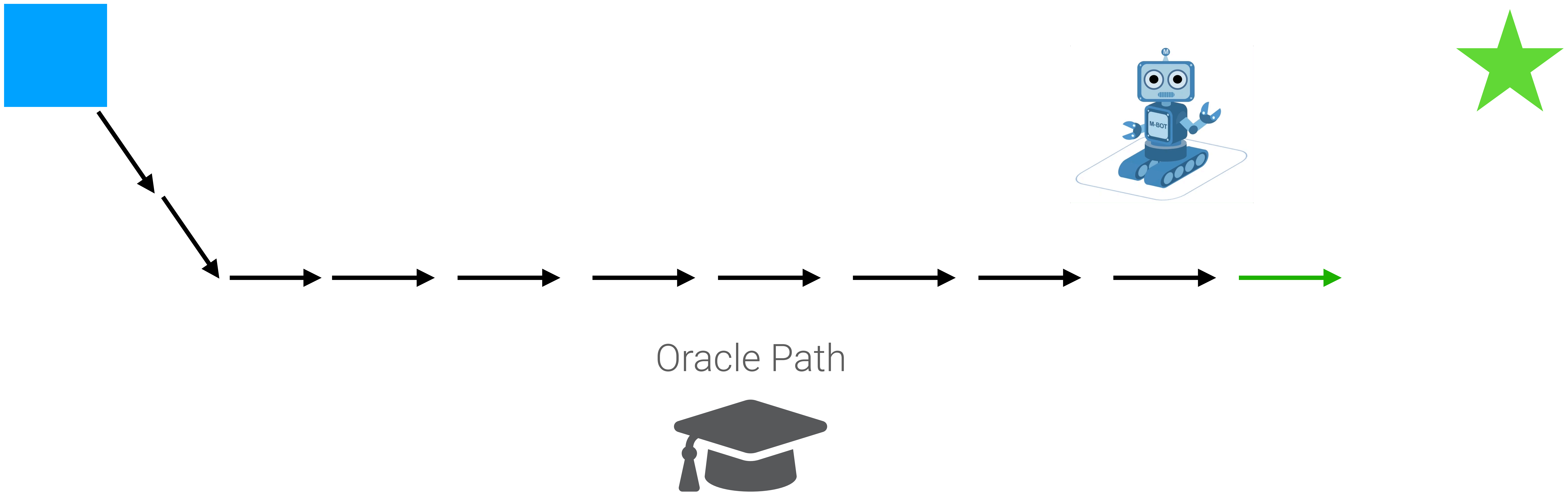
Oracle Path

Training for navigation

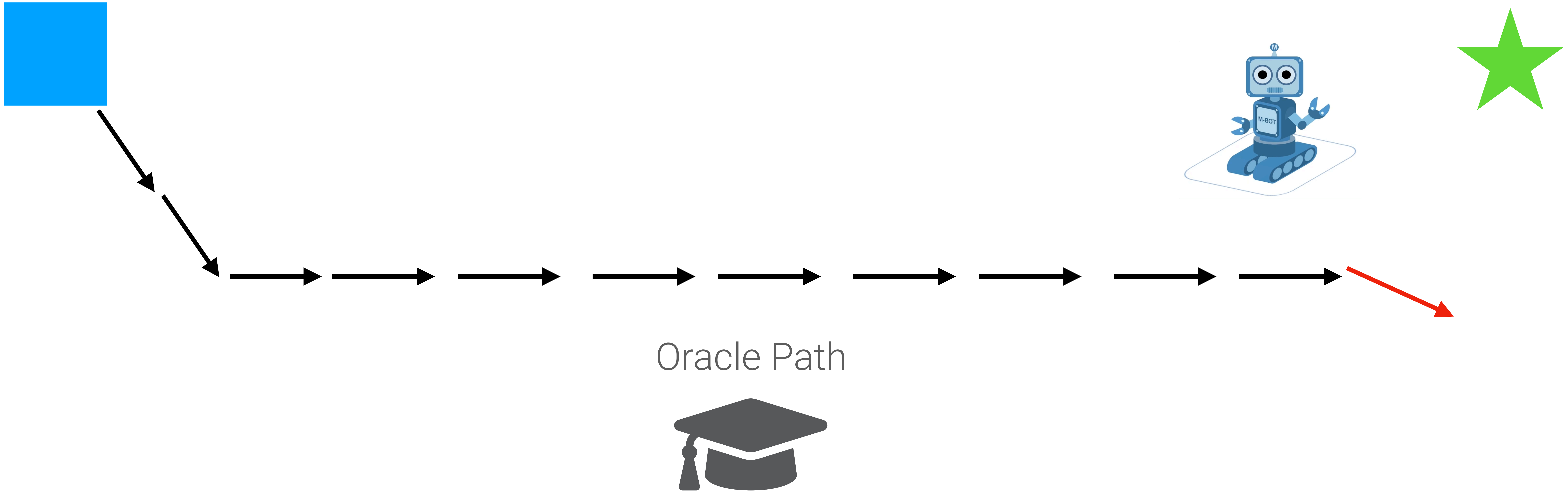


Oracle Path

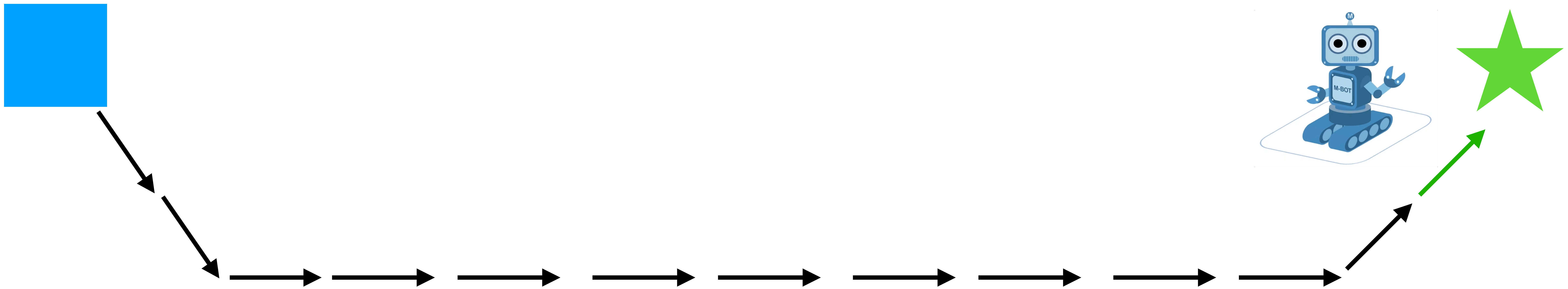
Training for navigation



Training for navigation



Training for navigation



Oracle Path

Navigator ablation

Visual Variations

{ Blind, RGB, PC, PC+RGB } ×

Language Variations

{ No Question, Question } ×

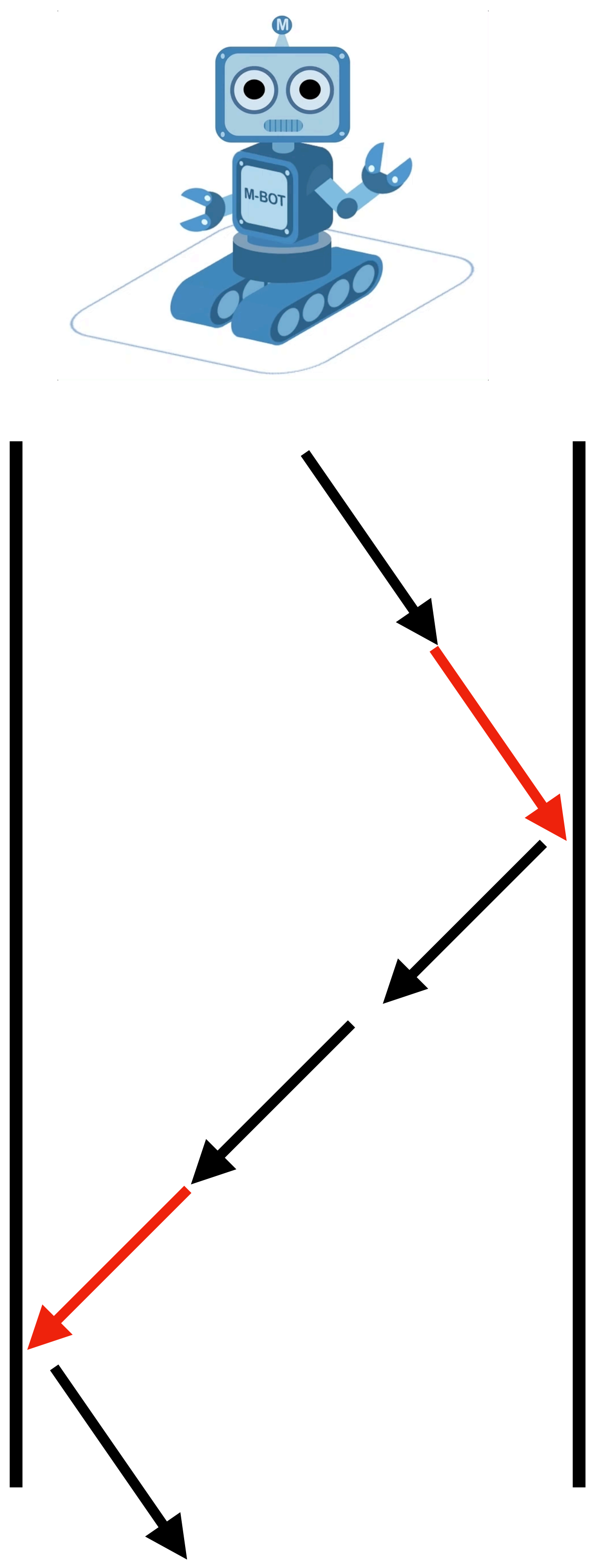
Memory Variations

{ Reactive, Memory }

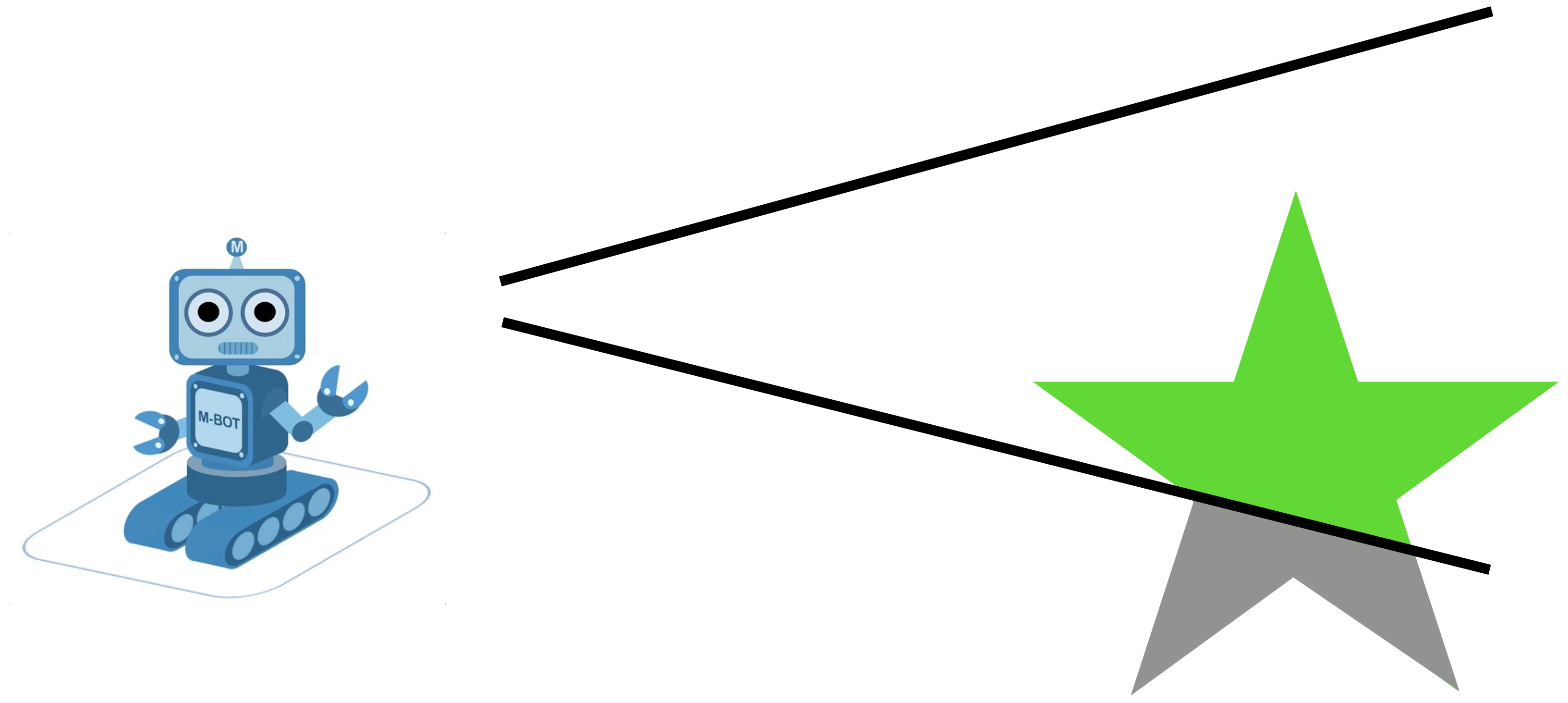
Shifting the Baseline: Thomason et. al., 2018

Metrics

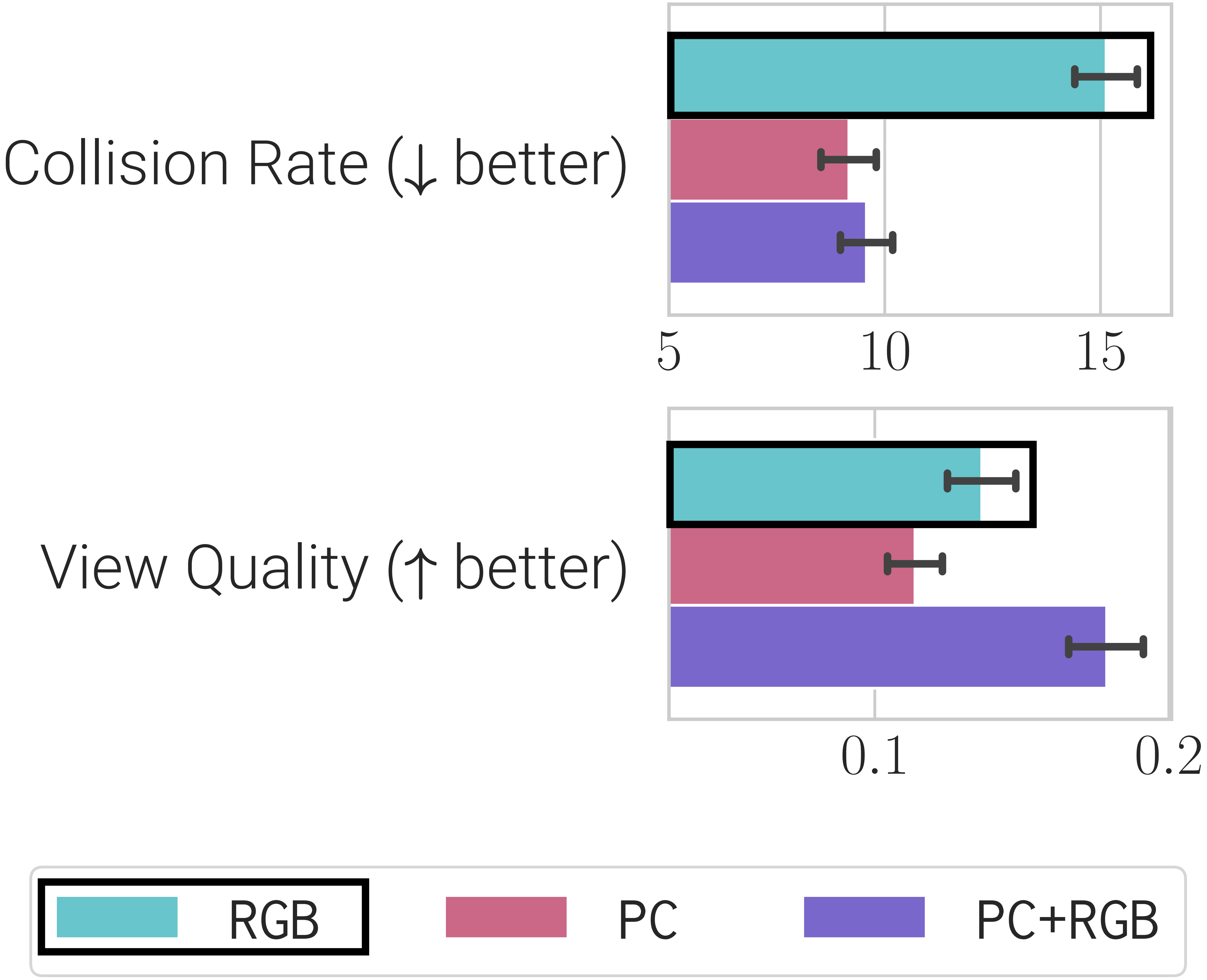
Collision Rate (↓ better)



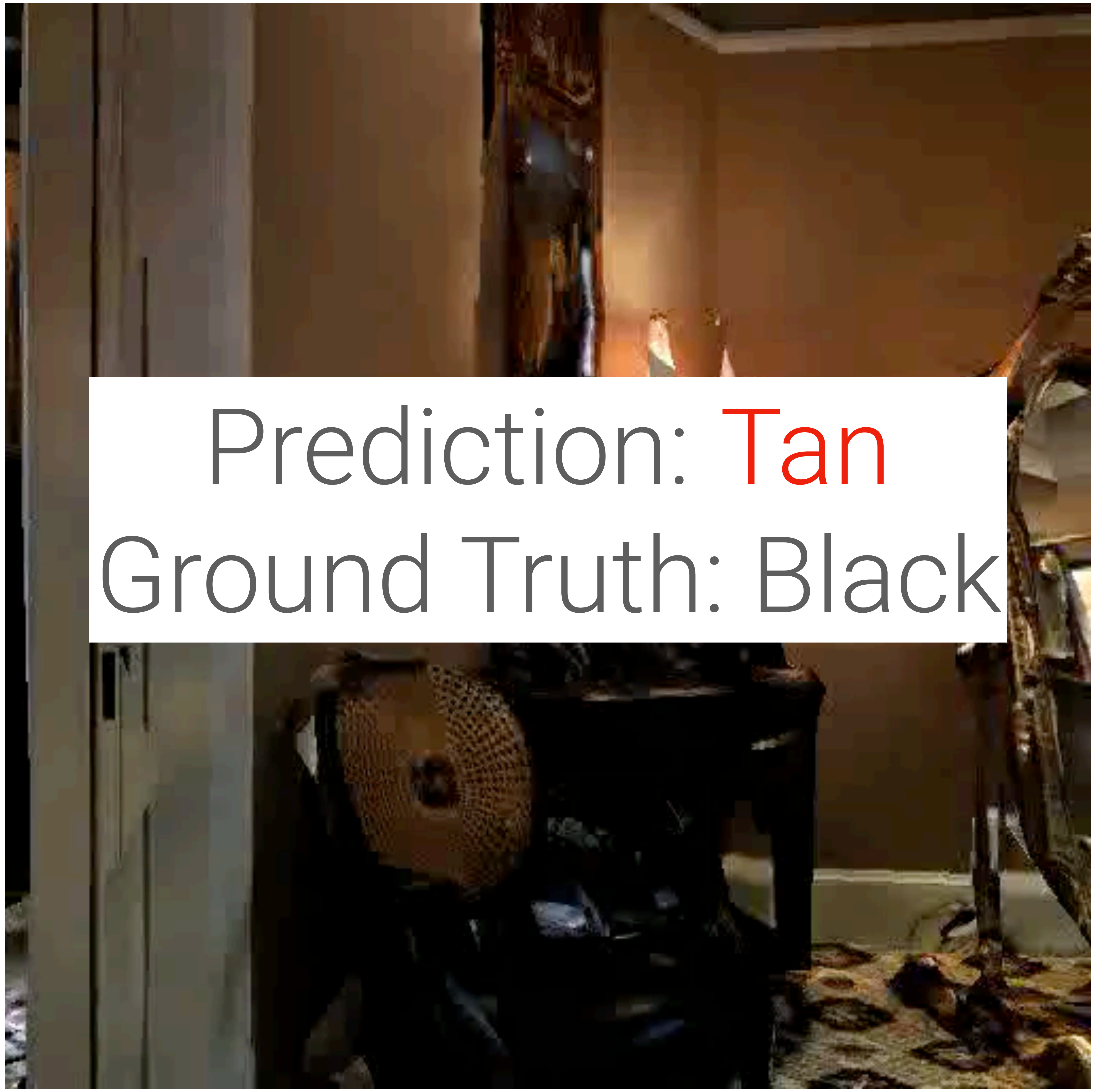
View Quality (↑ better)



Q: What color is the fireplace in the bedroom?

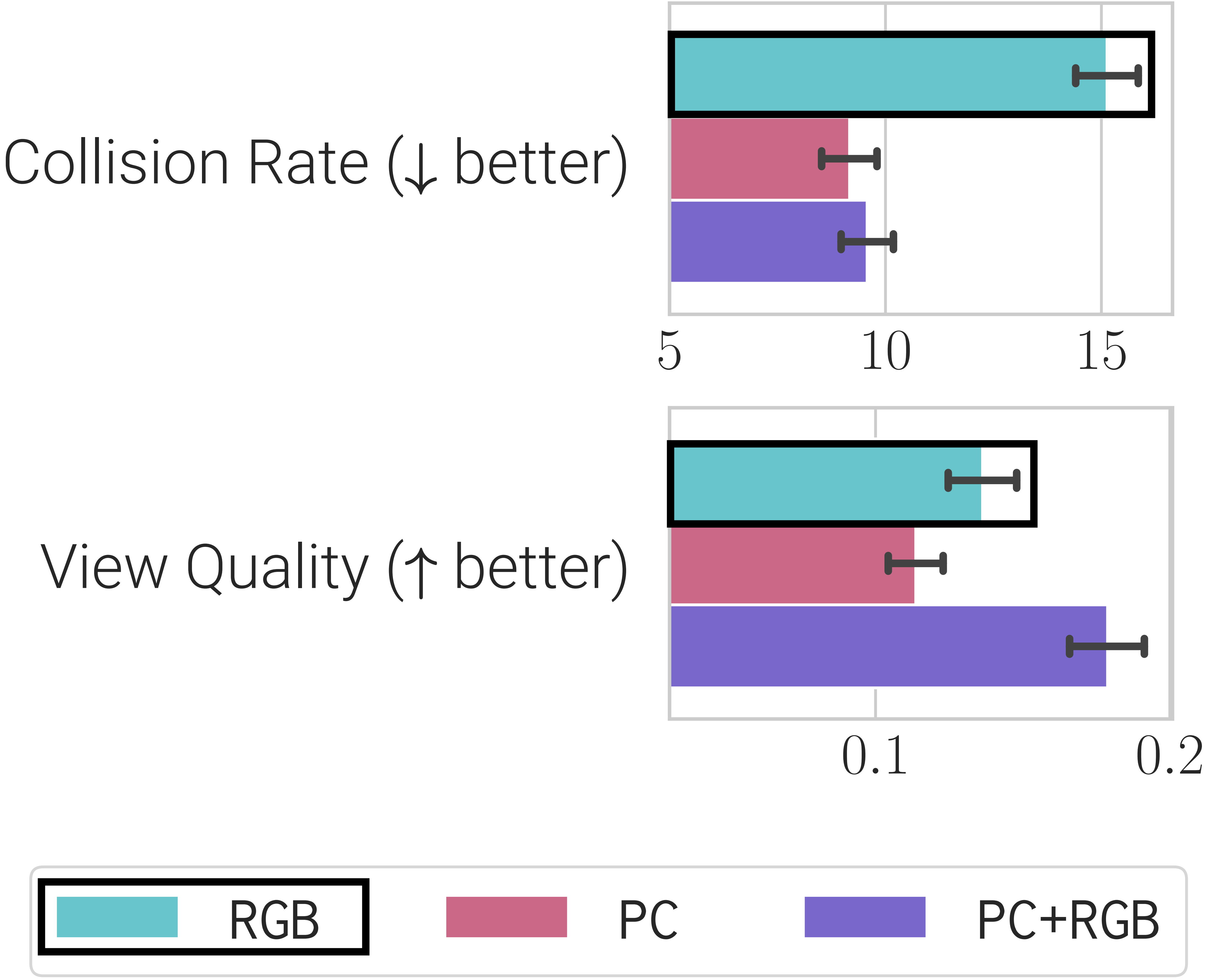


Q: What color is the fireplace in the bedroom?

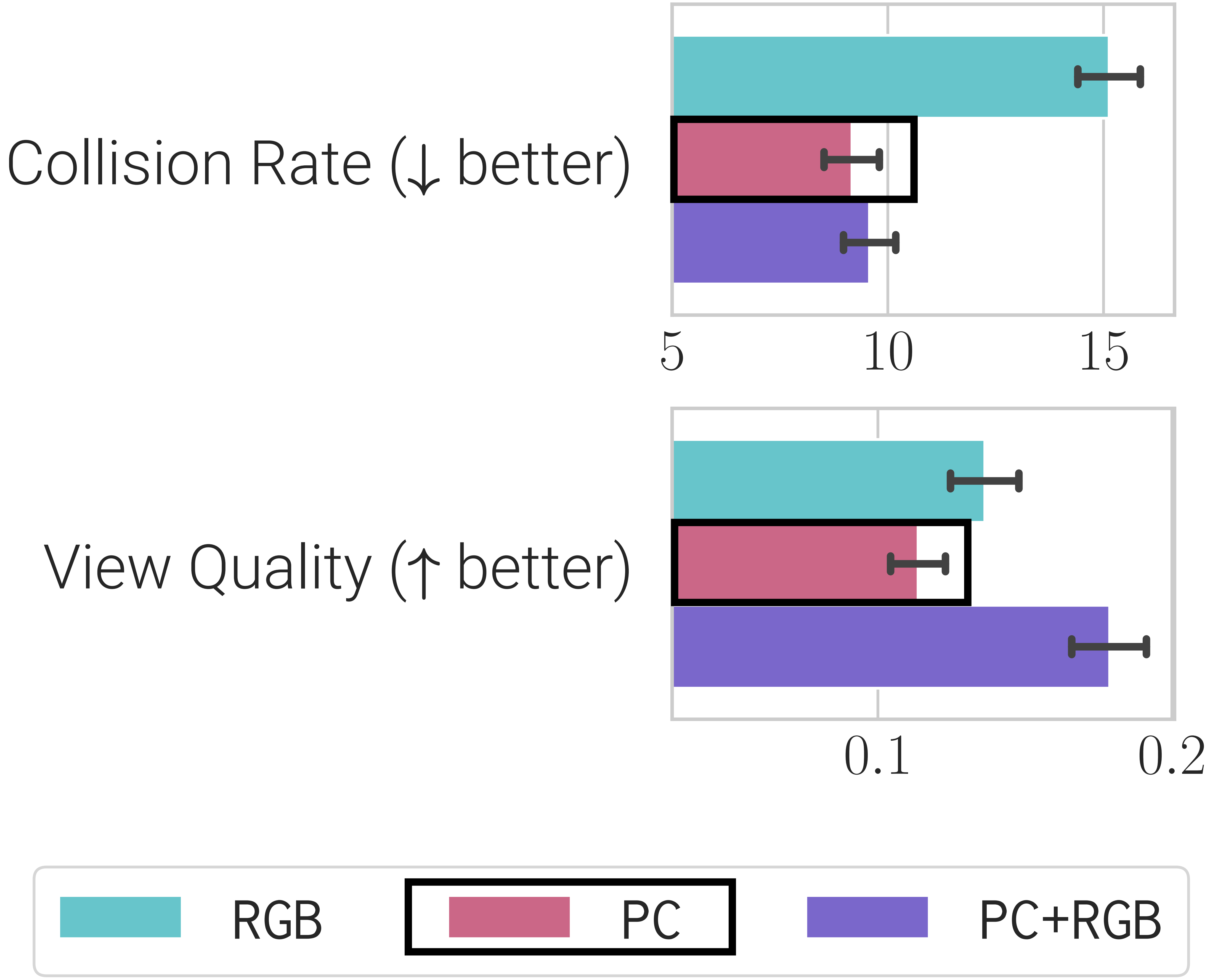


Prediction: Tan

Ground Truth: Black



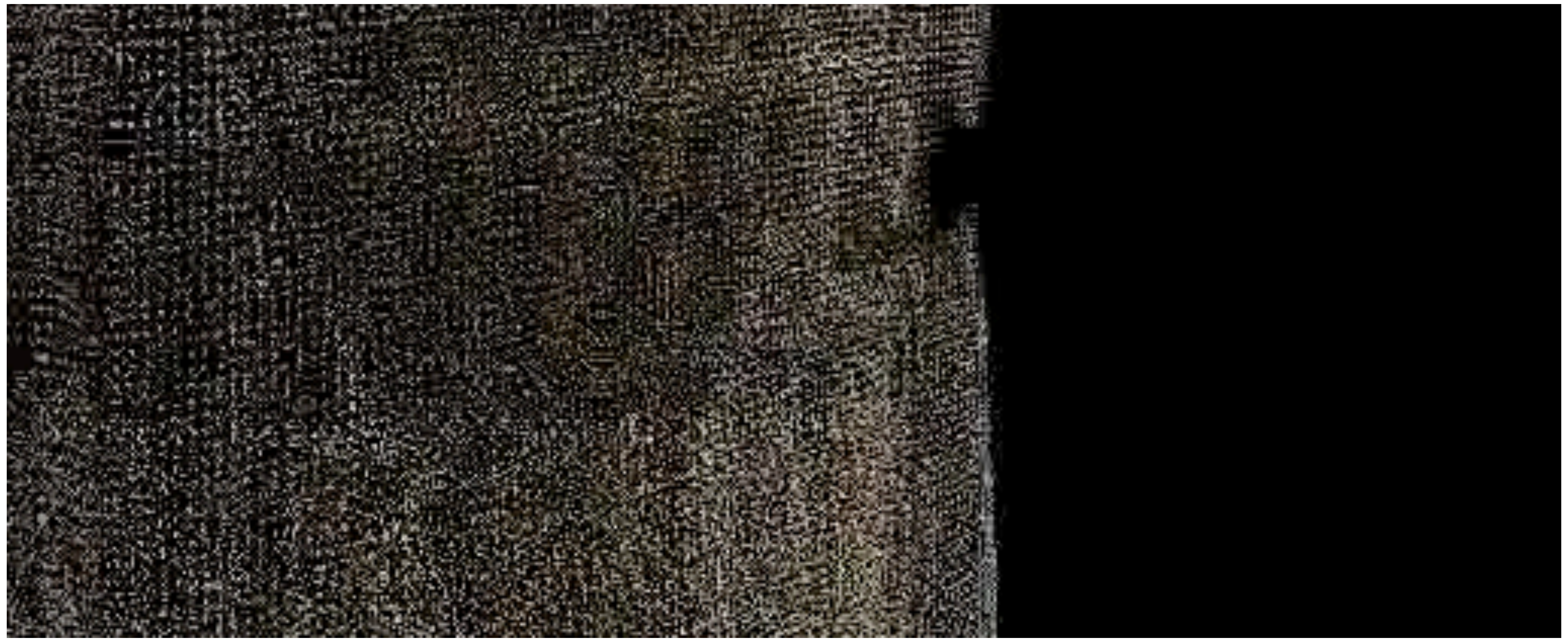
Q: What room is the wardrobe located in?

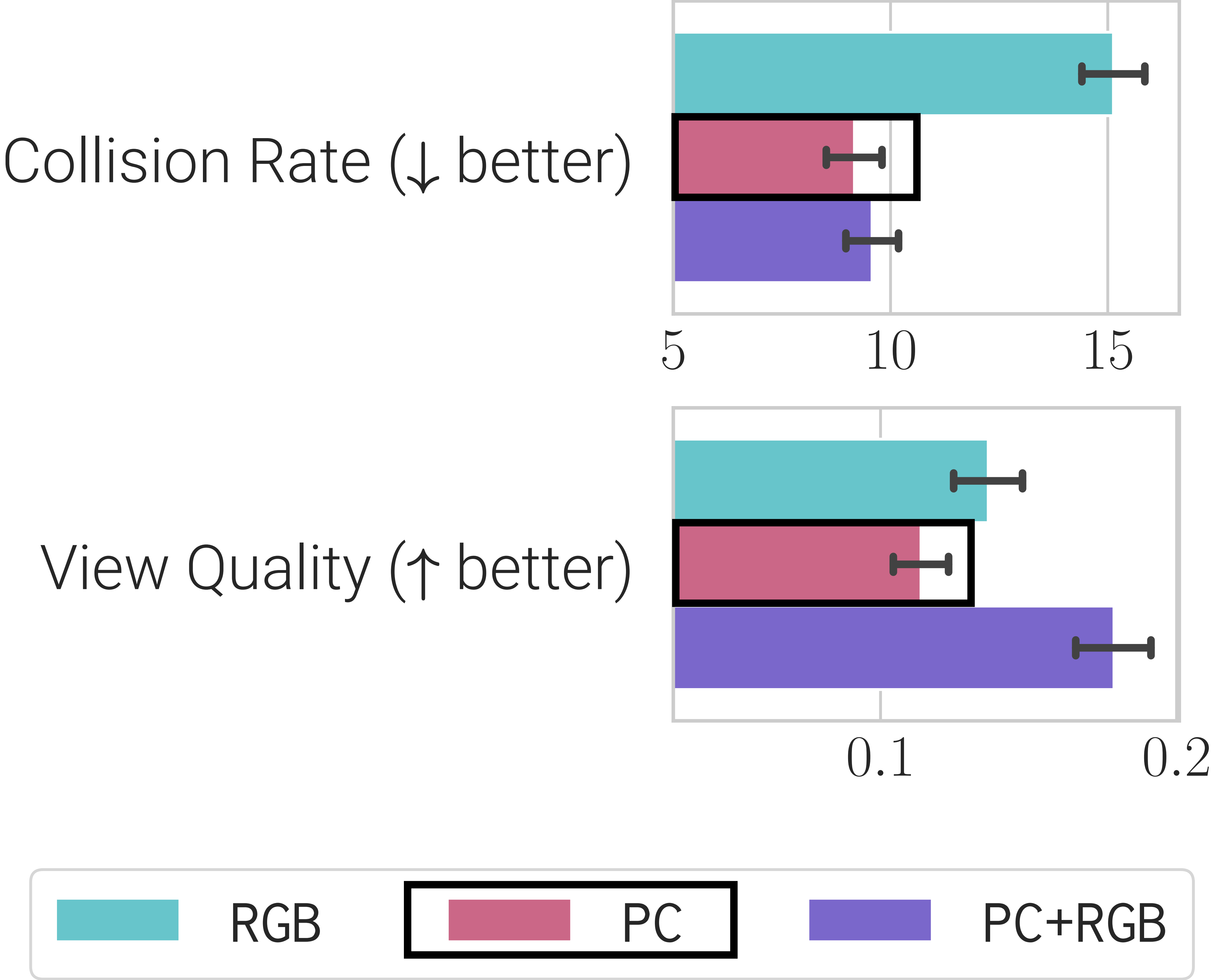


Q: What room is the wardrobe located in?

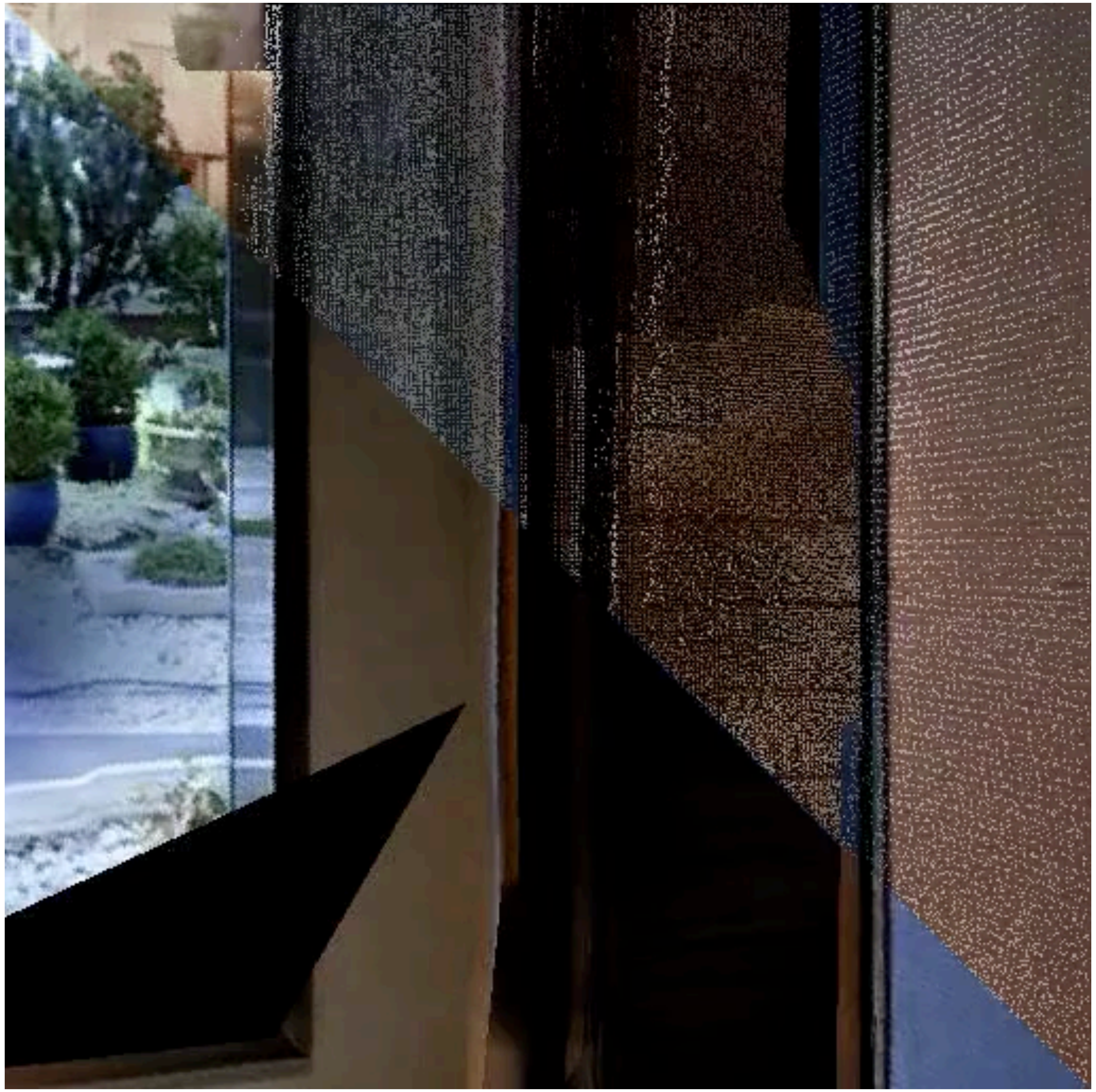
Prediction: Bathroom

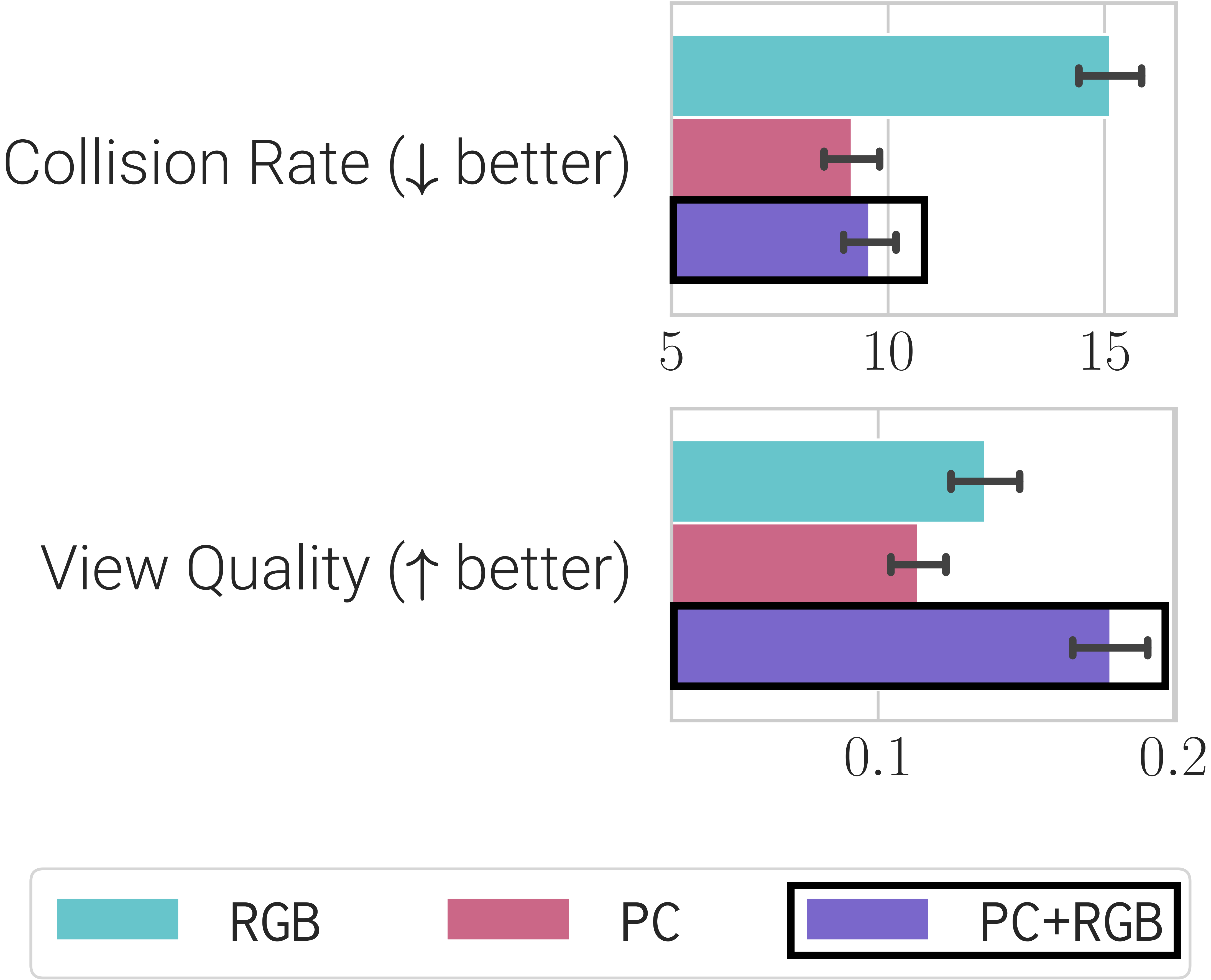
Ground Truth: Bedroom





Q: What color is the counter in the hallway?

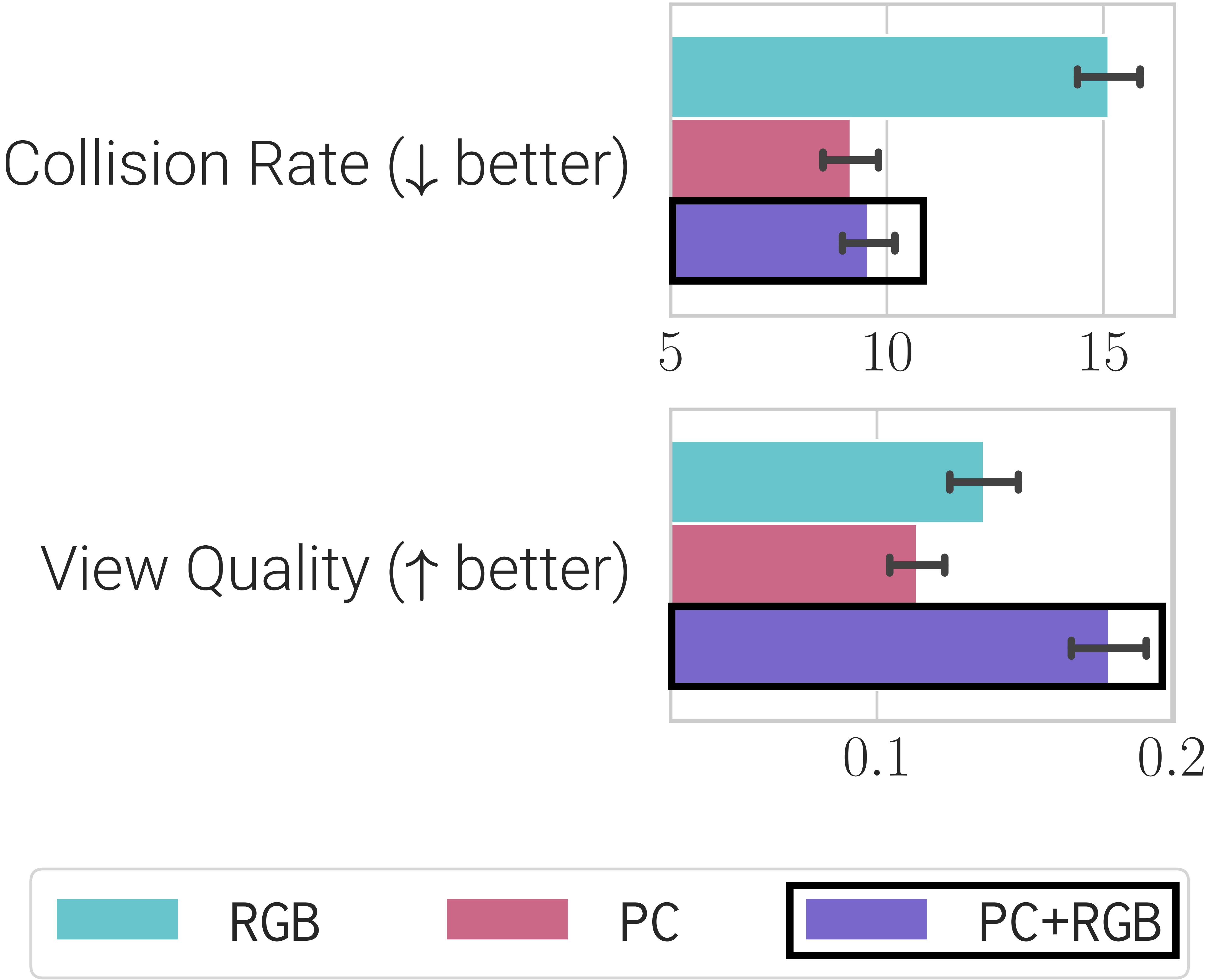




Q: What color is the counter in the hallway?

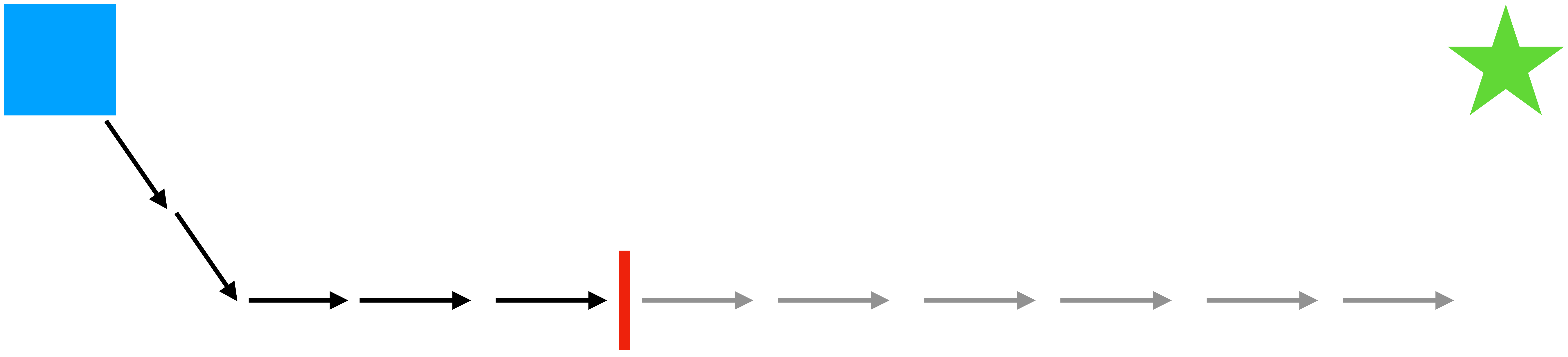
Prediction: White

Ground Truth: White



And lots more!

Forward only works well

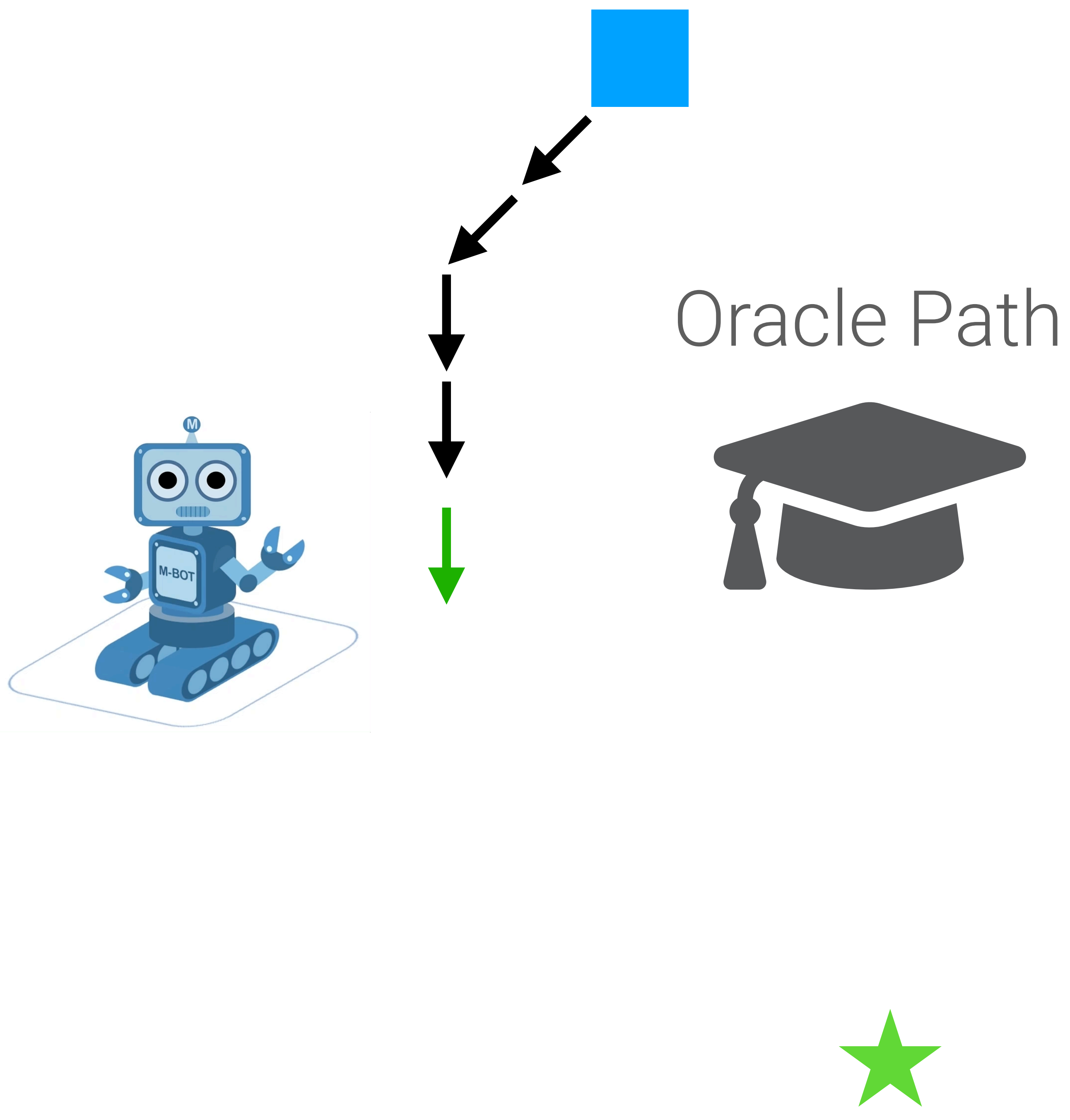


Agent given control

EmbodiedQA: Das et. al., 2018

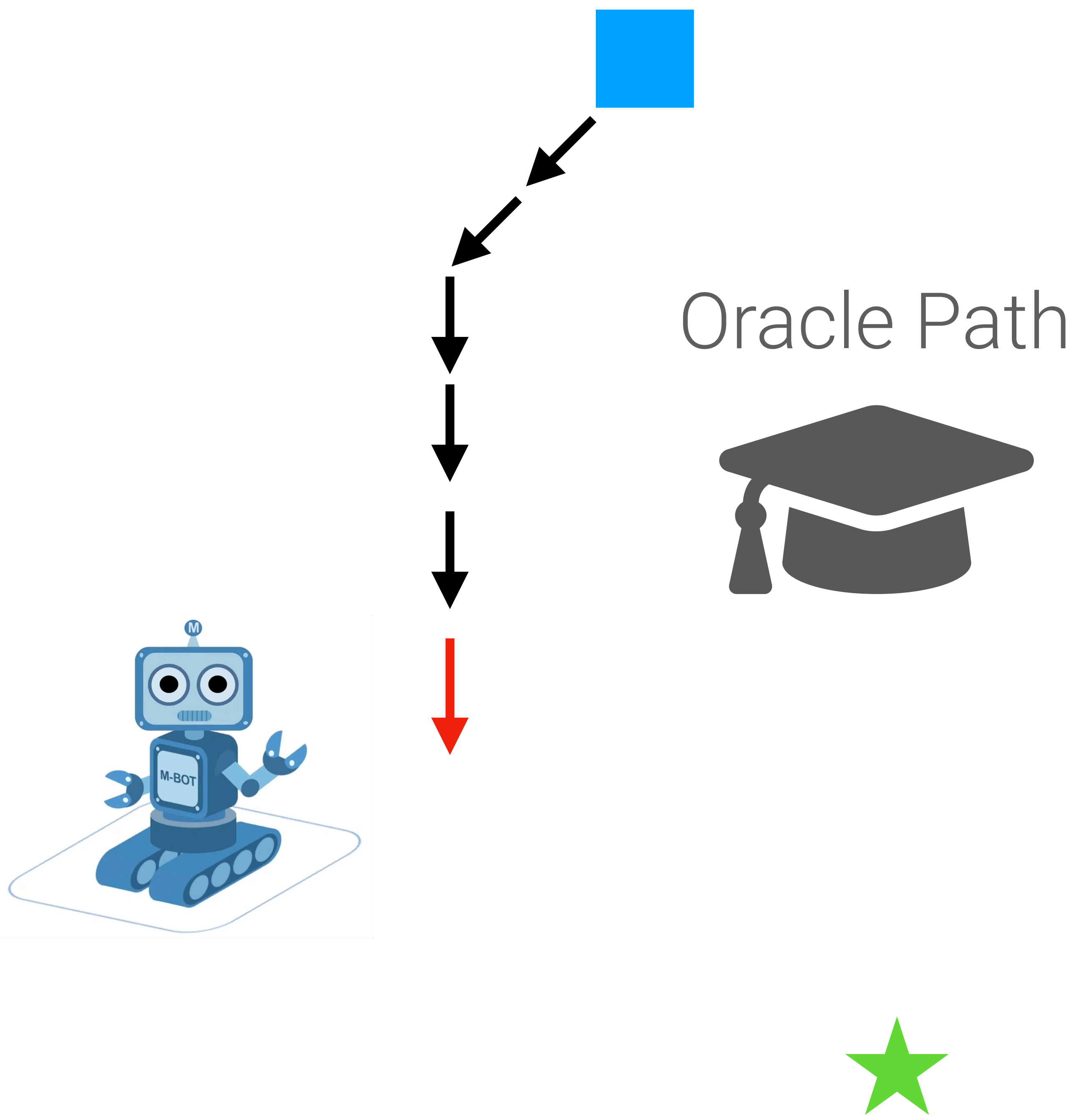
Repeat last action works well during training but fails during evaluation

During training



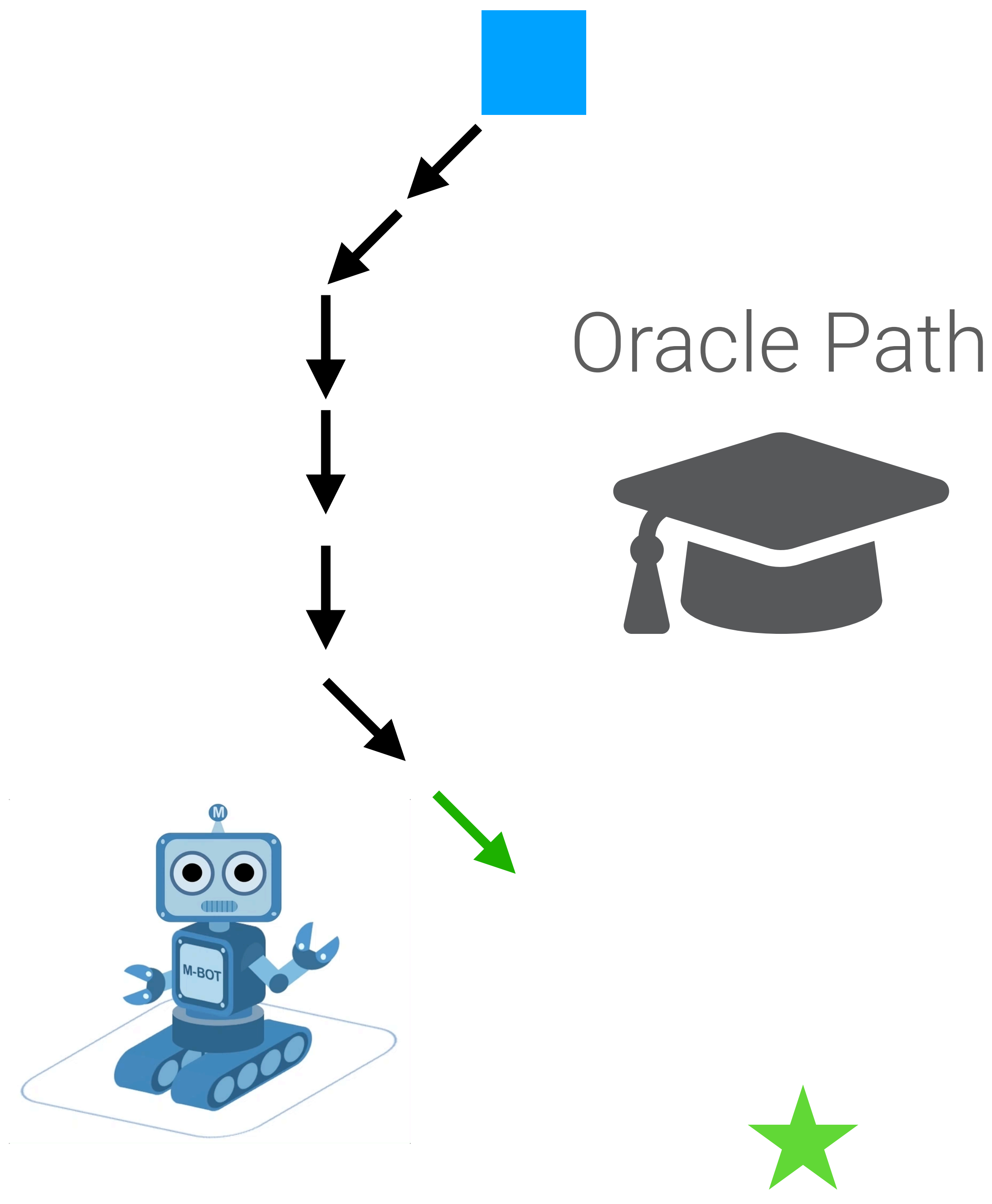
Repeat last action works well during training but fails during evaluation

During training



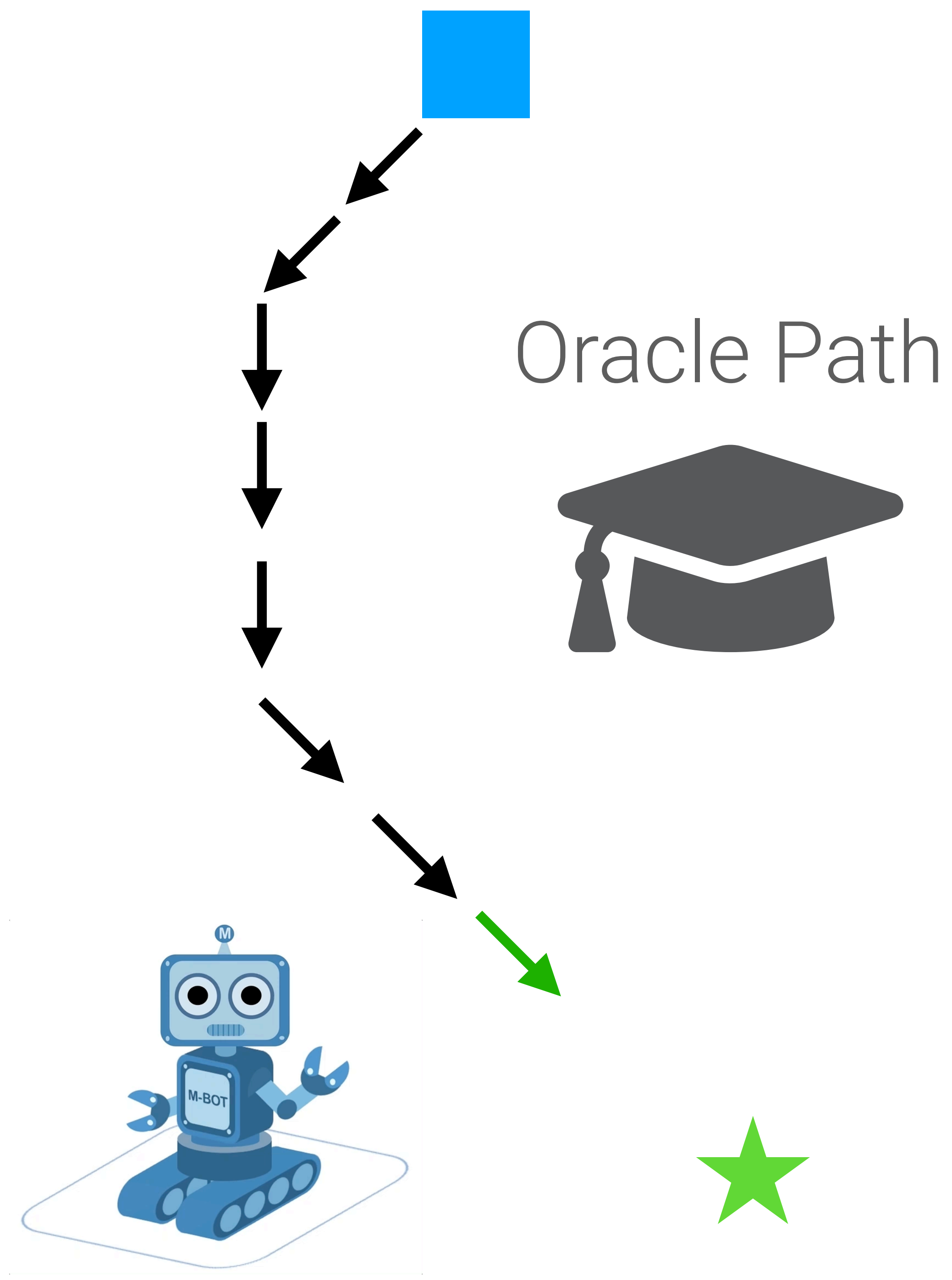
Repeat last action works well during training but fails during evaluation

During training



Repeat last action works well during training but fails during evaluation

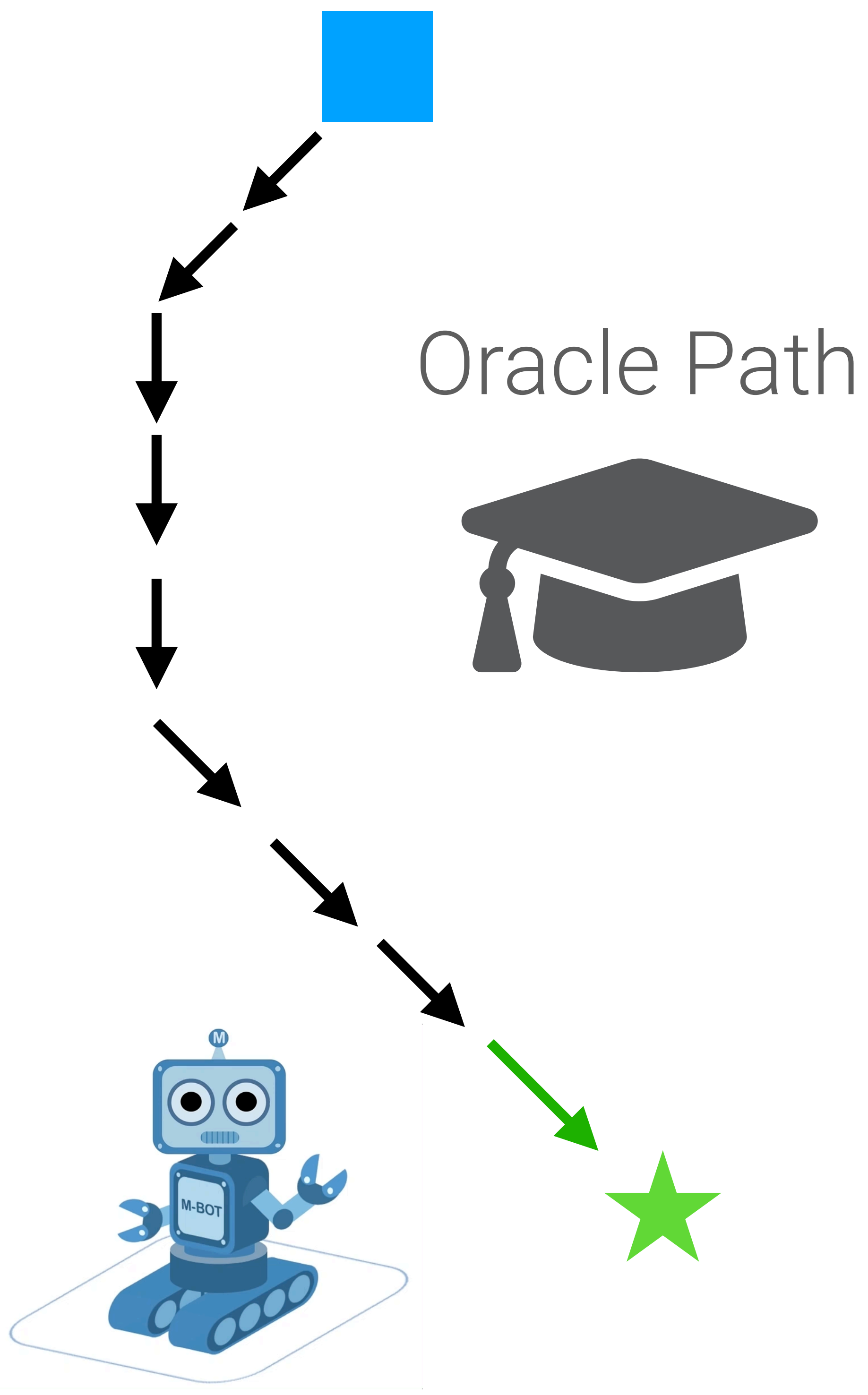
During training



Oracle Path

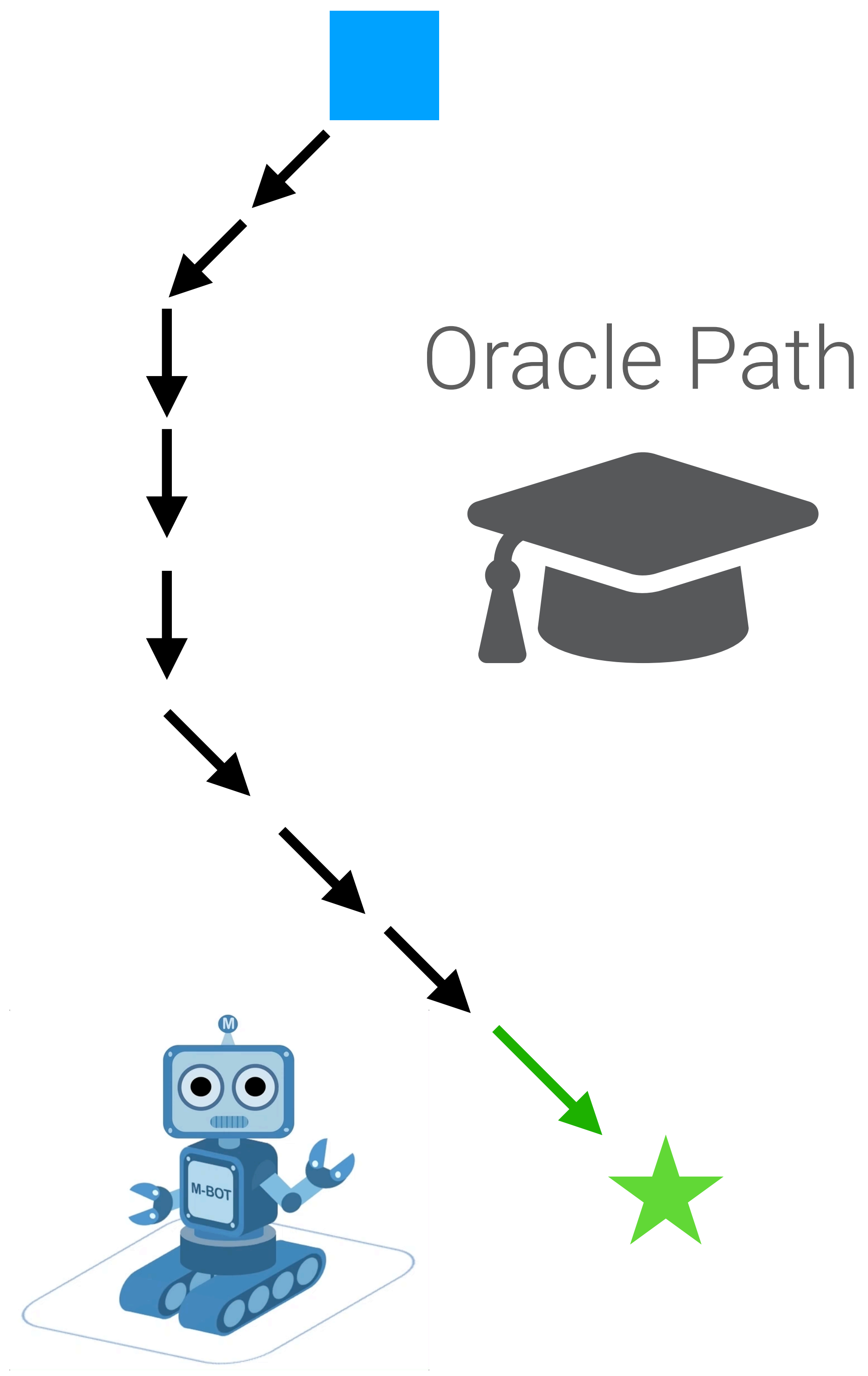
Repeat last action works well during training but fails during evaluation

During training

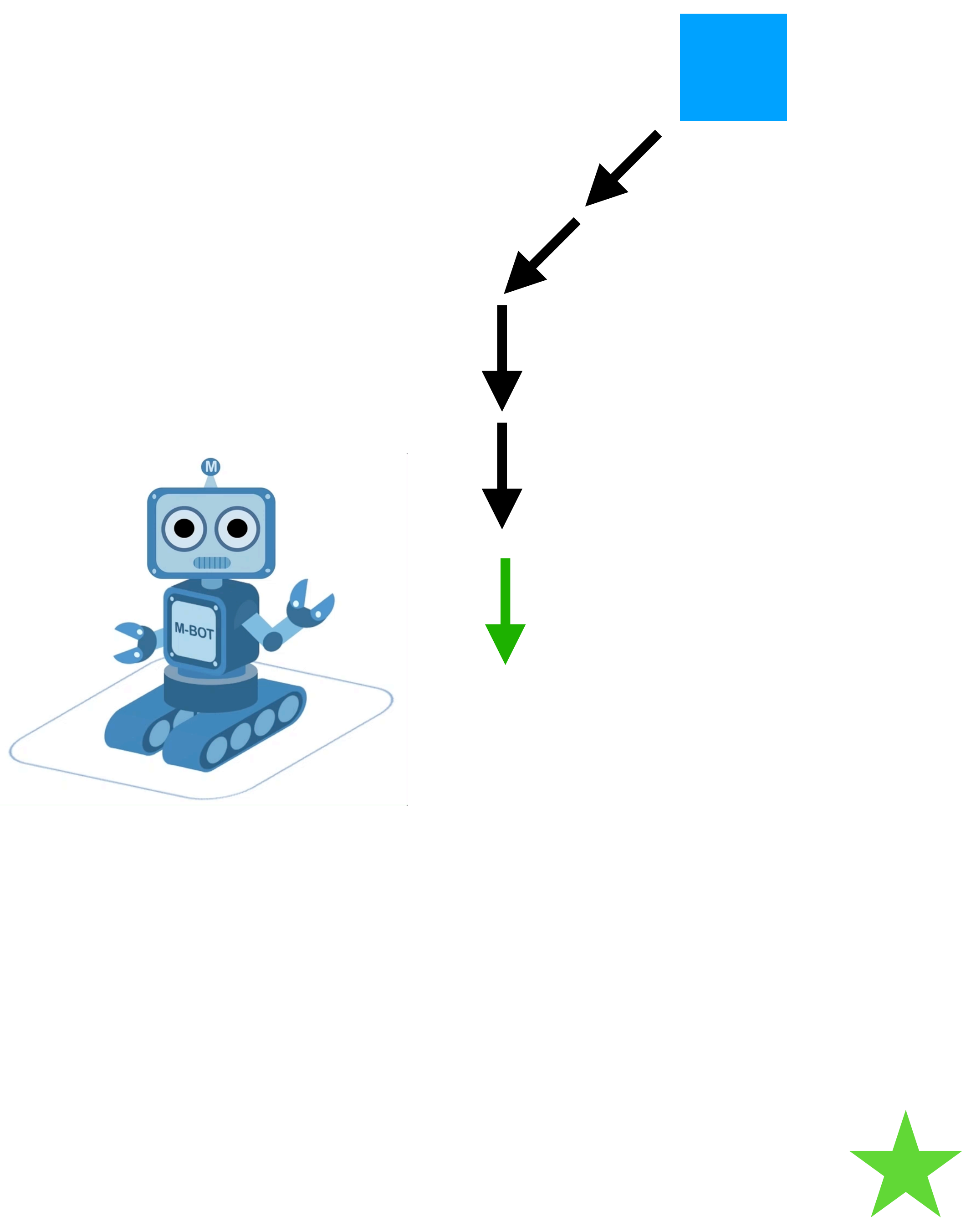


Repeat last action works well during training but fails during evaluation

During training

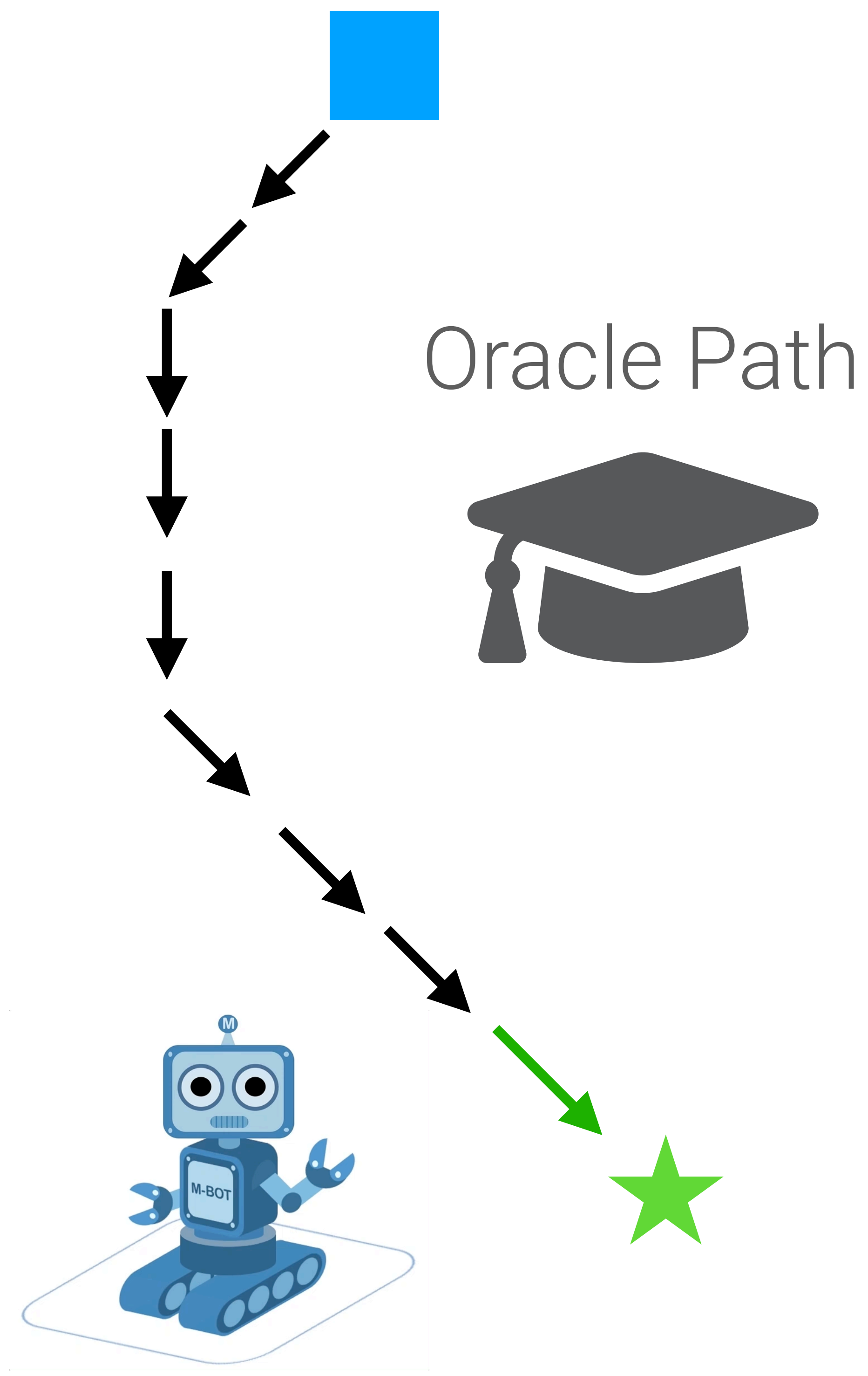


During evaluation

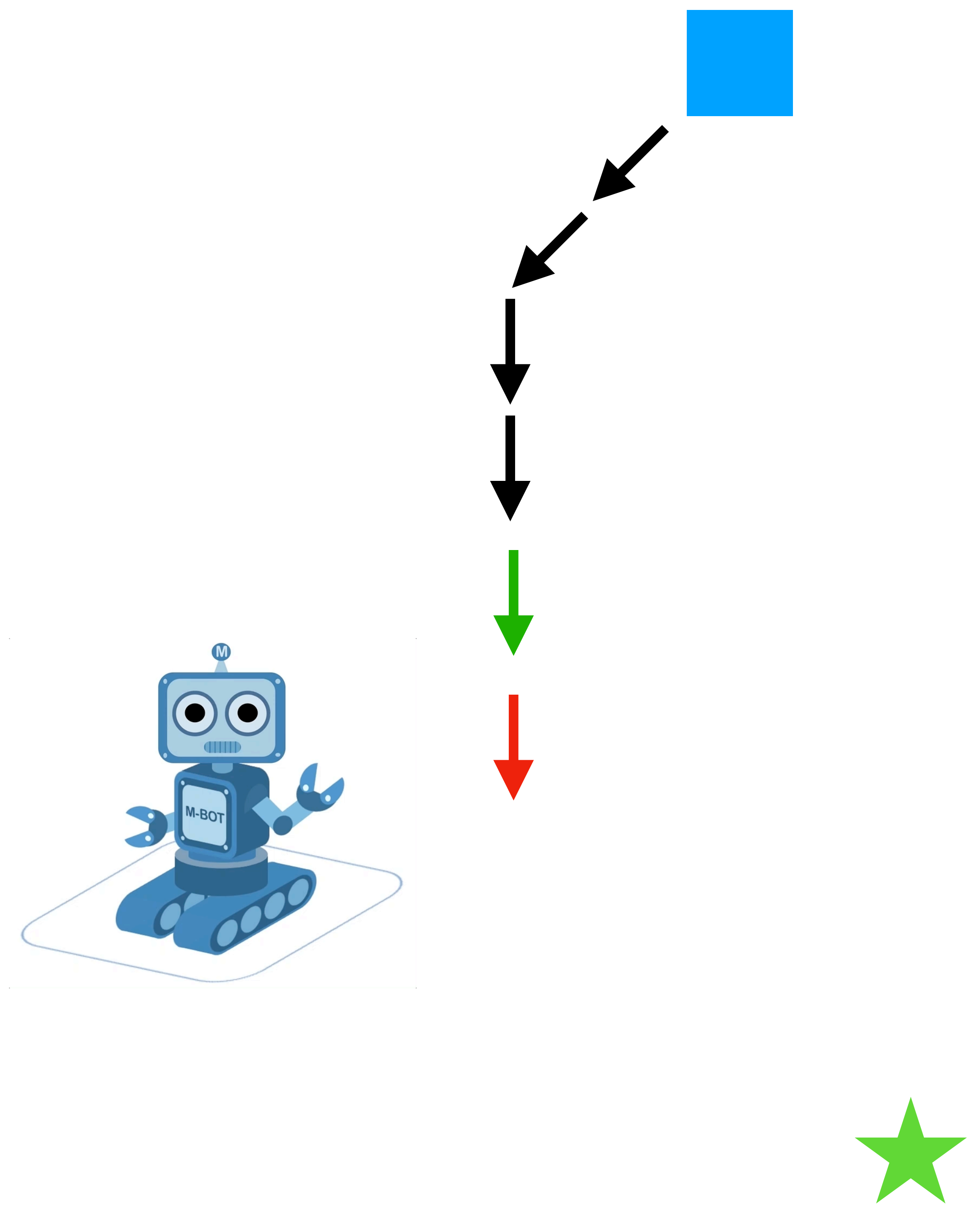


Repeat last action works well during training but fails during evaluation

During training

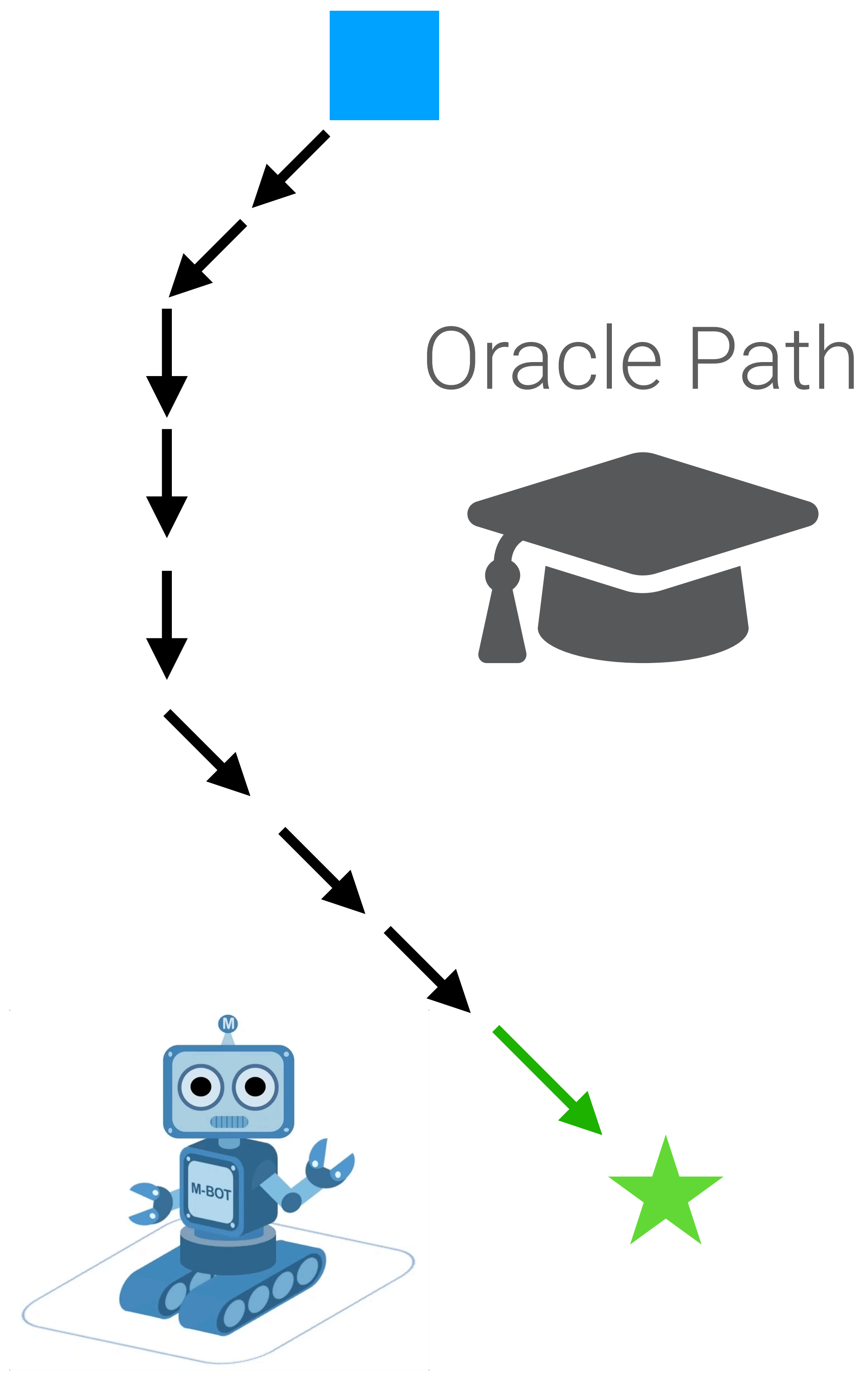


During evaluation

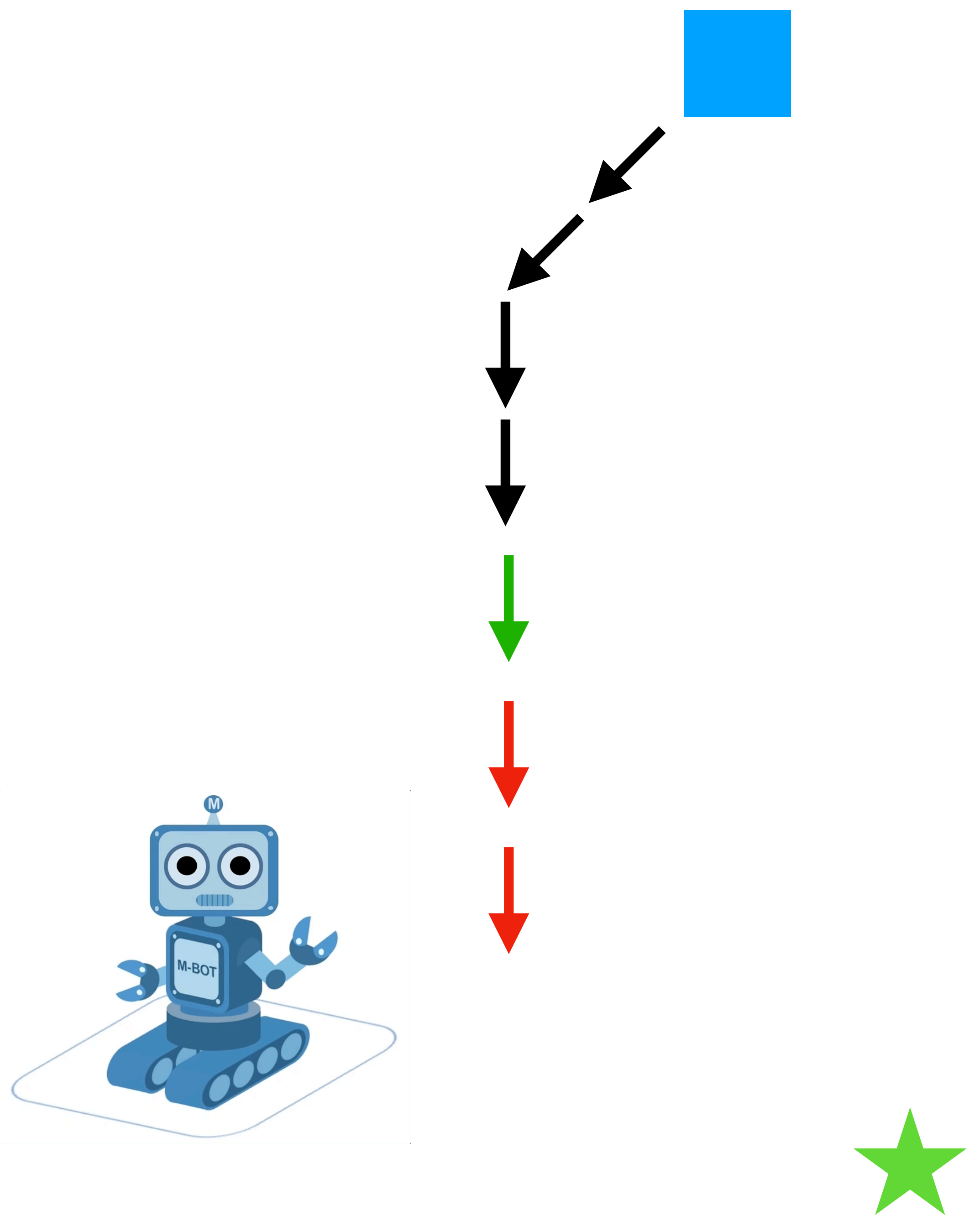


Repeat last action works well during training but fails during evaluation

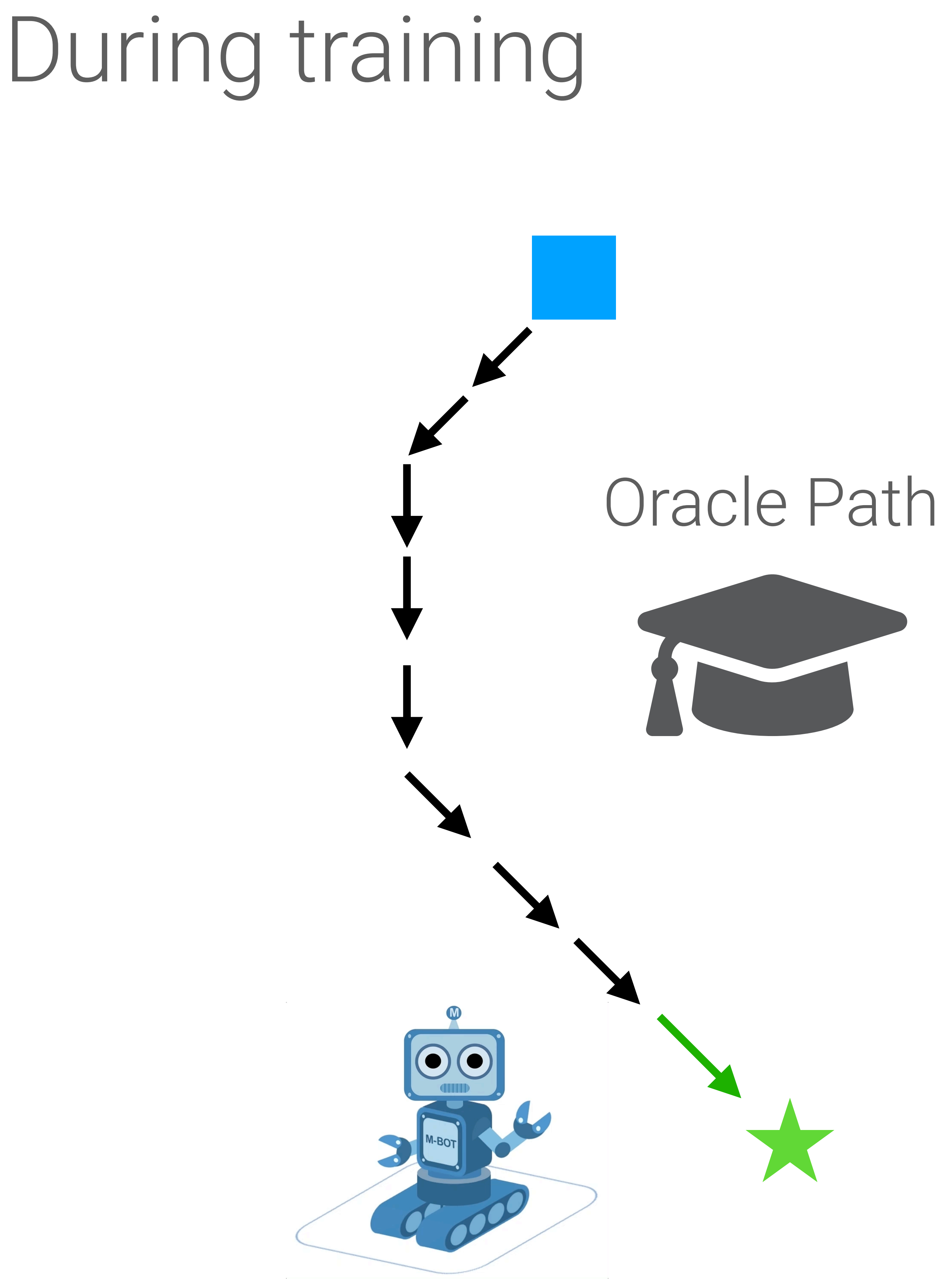
During training

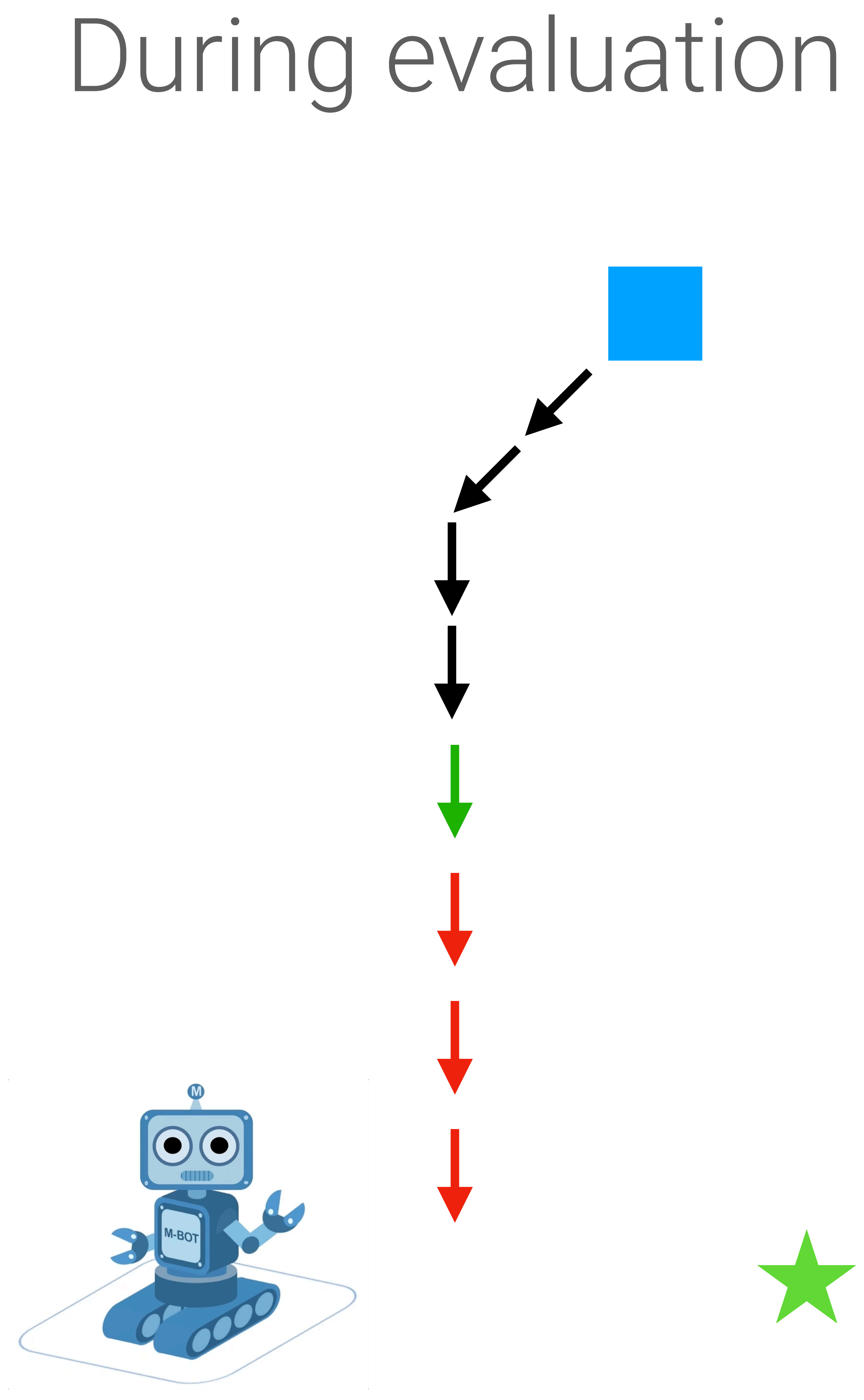


During evaluation

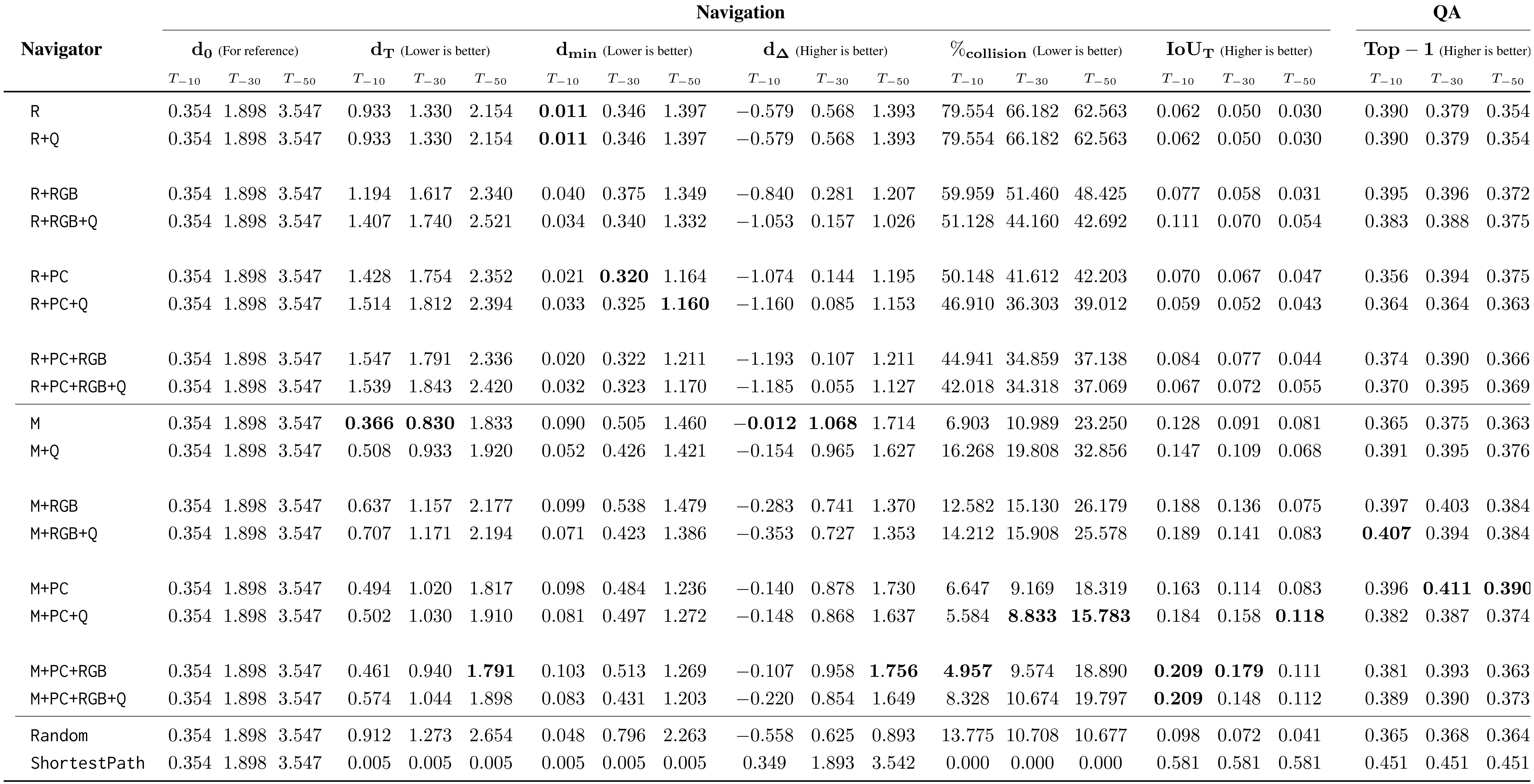


Repeat last action works well during training but fails during evaluation





Navigator ablation



	Navigation																		**QA**		
Navigator	$\mathbf{d_0}$ (For reference)			$\mathbf{d_T}$ (Lower is better)			$\mathbf{d_{min}}$ (Lower is better)			$\mathbf{d_\Delta}$ (Higher is better)			$\%_{\mathbf{collision}}$ (Lower is better)			$\mathbf{IoU_T}$ (Higher is better)			$\mathbf{Top-1}$ (Higher is better)		
	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}	T_{-10}	T_{-30}	T_{-50}
`R`	0.354	1.898	3.547	0.933	1.330	2.154	**0.011**	0.346	1.397	−0.579	0.568	1.393	79.554	66.182	62.563	0.062	0.050	0.030	0.390	0.379	0.354
`R+Q`	0.354	1.898	3.547	0.933	1.330	2.154	**0.011**	0.346	1.397	−0.579	0.568	1.393	79.554	66.182	62.563	0.062	0.050	0.030	0.390	0.379	0.354
`R+RGB`	0.354	1.898	3.547	1.194	1.617	2.340	0.040	0.375	1.349	−0.840	0.281	1.207	59.959	51.460	48.425	0.077	0.058	0.031	0.395	0.396	0.372
`R+RGB+Q`	0.354	1.898	3.547	1.407	1.740	2.521	0.034	0.340	1.332	−1.053	0.157	1.026	51.128	44.160	42.692	0.111	0.070	0.054	0.383	0.388	0.375
`R+PC`	0.354	1.898	3.547	1.428	1.754	2.352	0.021	**0.320**	1.164	−1.074	0.144	1.195	50.148	41.612	42.203	0.070	0.067	0.047	0.356	0.394	0.375
`R+PC+Q`	0.354	1.898	3.547	1.514	1.812	2.394	0.033	0.325	**1.160**	−1.160	0.085	1.153	46.910	36.303	39.012	0.059	0.052	0.043	0.364	0.364	0.363
`R+PC+RGB`	0.354	1.898	3.547	1.547	1.791	2.336	0.020	0.322	1.211	−1.193	0.107	1.211	44.941	34.859	37.138	0.084	0.077	0.044	0.374	0.390	0.366
`R+PC+RGB+Q`	0.354	1.898	3.547	1.539	1.843	2.420	0.032	0.323	1.170	−1.185	0.055	1.127	42.018	34.318	37.069	0.067	0.072	0.055	0.370	0.395	0.369
`M`	0.354	1.898	3.547	**0.366**	**0.830**	1.833	0.090	0.505	1.460	**−0.012**	**1.068**	1.714	6.903	10.989	23.250	0.128	0.091	0.081	0.365	0.375	0.363
`M+Q`	0.354	1.898	3.547	0.508	0.933	1.920	0.052	0.426	1.421	−0.154	0.965	1.627	16.268	19.808	32.856	0.147	0.109	0.068	0.391	0.395	0.376
`M+RGB`	0.354	1.898	3.547	0.637	1.157	2.177	0.099	0.538	1.479	−0.283	0.741	1.370	12.582	15.130	26.179	0.188	0.136	0.075	0.397	0.403	0.384
`M+RGB+Q`	0.354	1.898	3.547	0.707	1.171	2.194	0.071	0.423	1.386	−0.353	0.727	1.353	14.212	15.908	25.578	0.189	0.141	0.083	**0.407**	0.394	0.384
`M+PC`	0.354	1.898	3.547	0.494	1.020	1.817	0.098	0.484	1.236	−0.140	0.878	1.730	6.647	9.169	18.319	0.163	0.114	0.083	0.396	**0.411**	**0.390**
`M+PC+Q`	0.354	1.898	3.547	0.502	1.030	1.910	0.081	0.497	1.272	−0.148	0.868	1.637	5.584	**8.833**	**15.783**	0.184	0.158	**0.118**	0.382	0.387	0.374
`M+PC+RGB`	0.354	1.898	3.547	0.461	0.940	**1.791**	0.103	0.513	1.269	−0.107	0.958	**1.756**	**4.957**	9.574	18.890	**0.209**	**0.179**	0.111	0.381	0.393	0.363
`M+PC+RGB+Q`	0.354	1.898	3.547	0.574	1.044	1.898	0.083	0.431	1.203	−0.220	0.854	1.649	8.328	10.674	19.797	**0.209**	0.148	0.112	0.389	0.390	0.373
`Random`	0.354	1.898	3.547	0.912	1.273	2.654	0.048	0.796	2.263	−0.558	0.625	0.893	13.775	10.708	10.677	0.098	0.072	0.041	0.365	0.368	0.364
`ShortestPath`	0.354	1.898	3.547	0.005	0.005	0.005	0.005	0.005	0.005	0.349	1.893	3.542	0.000	0.000	0.000	0.581	0.581	0.581	0.451	0.451	0.451

Habitat: Where AI Agents Live

Manolis Savva

Abhishek Kadian

Oleksandr Maksymets

Yili Zhao

Erik Wijmans

Bhavana Jain

Julian Straub

Jia Liu

Vladlen Koltun

Jitendra Malik

Devi Parikh

Dhruv Batra

Modular high-level API | Fast and reliable simulator | Headless rendering

Support for generic tasks | Support for generic datasets | SLAM and RL baselines

aihabitat.org

Summary

Poster ID: 185

- **Comparison of point cloud vs. RGB perception**
  We take a step toward closing the gap between simulation and reality by examine how depth via point clouds affects the task of EmbodiedQA

- **Large scale navigator ablation**
  We perform a large ablation study to examine how various visual modalities, language modalities, and access to memory affect agents for EQA

- **Inflection Weighting**
  We propose a simple yet effective technique to help improve generalize from teacher forcing training to freeform evaluation

Paper and slides: embodiedqa.org

Slide Credits: Stefan Lee, Abhishek Das, Manolis Savva, Samyak Datta, and others