Zero-Shot Object Goal Navigation with Multi-Scale Geometric-Affordance Guidance

Shuaihang Yuan, Hao huang, Yu Hao, Congcong Wen, Antony Tzes, and Yi Fang

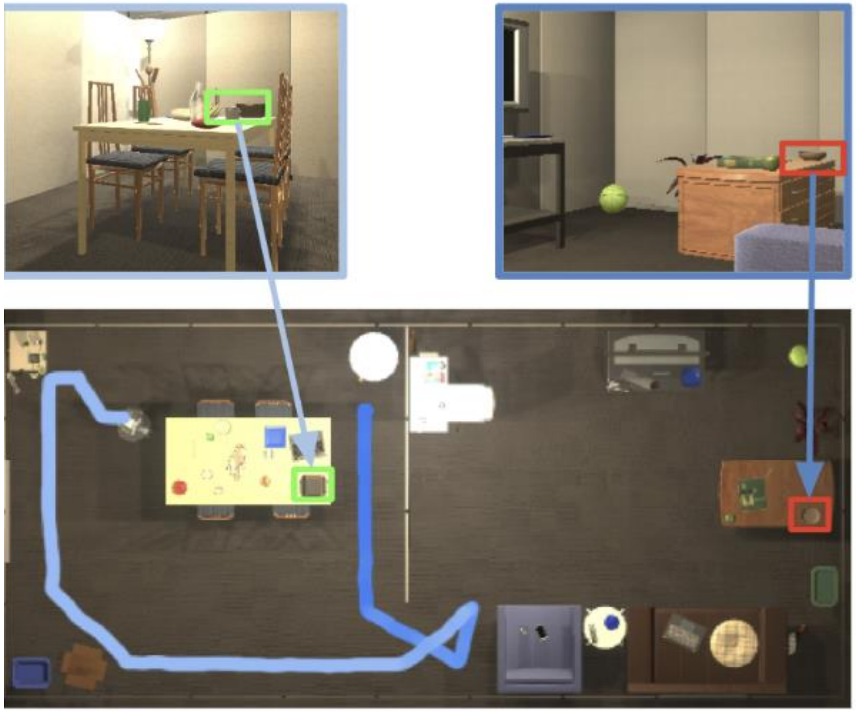
NYUAD Center for Artificial Intelligence and Robotics

Embodied AI&Robotics Lab

NYU Abu Dhabi

Object Goal Navigation (OGN)

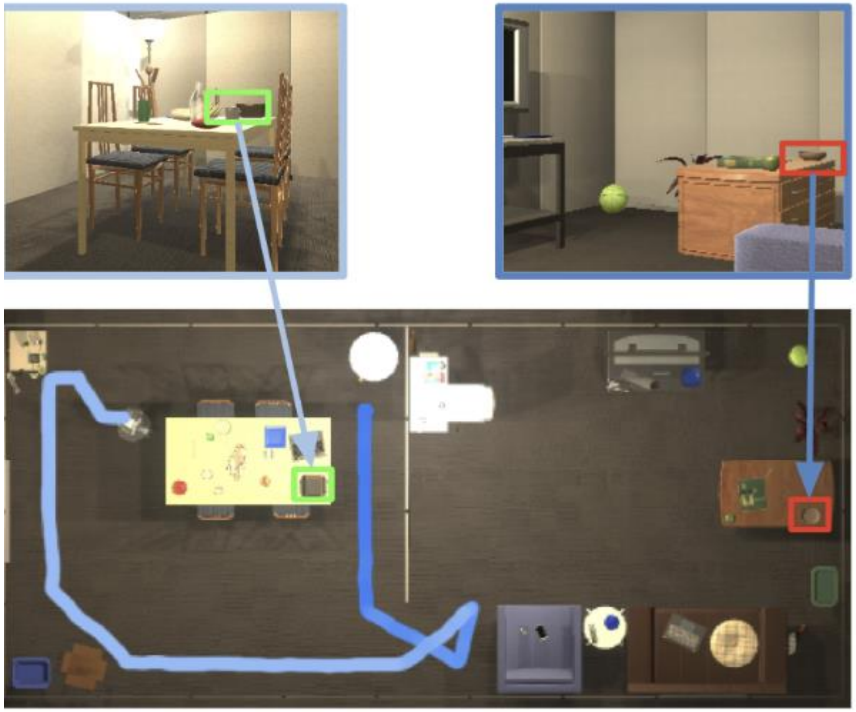
Object Goal Navigation requires both exploring the environment and identifying the semantic information of the scene to locate the desired object.



Object Goal Navigation (OGN)

Object Goal Navigation requires both exploring the environment and identifying the semantic information of the scene to locate the desired object.

Traditional OGN Approaches:
Perform well in trained environment

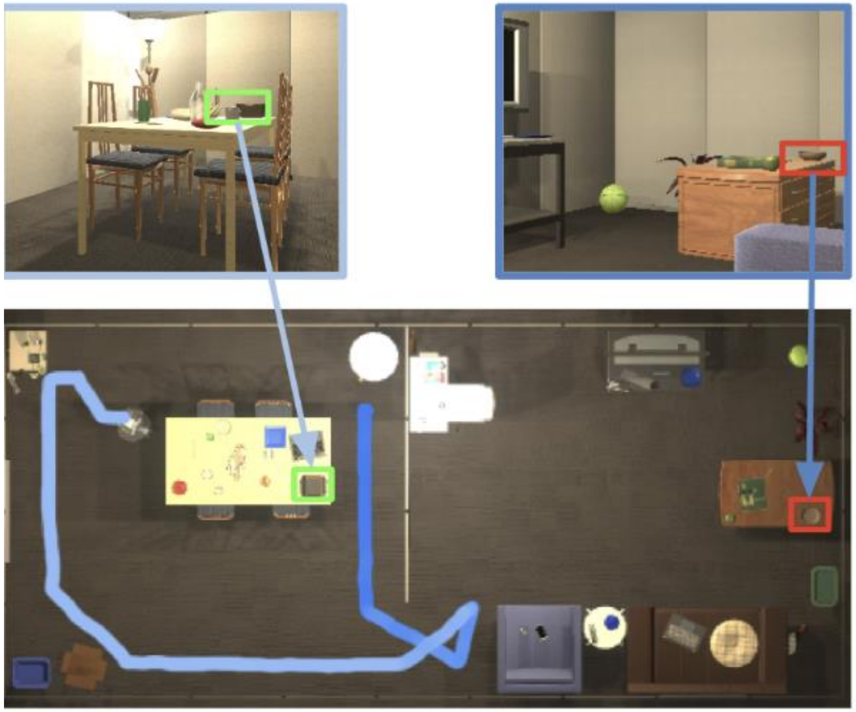


Object Goal Navigation (OGN)

Object Goal Navigation requires both exploring the environment and identifying the semantic information of the scene to locate the desired object.

Traditional OGN Approaches:
Perform well in trained environment

Zero-Shot OGN:
Able to navigate to unfamiliar objects in unknown environments without additional training.



Observation: Prior art relies the zero-shot detector for categorical information understanding which often fall shot when only partial observation are given

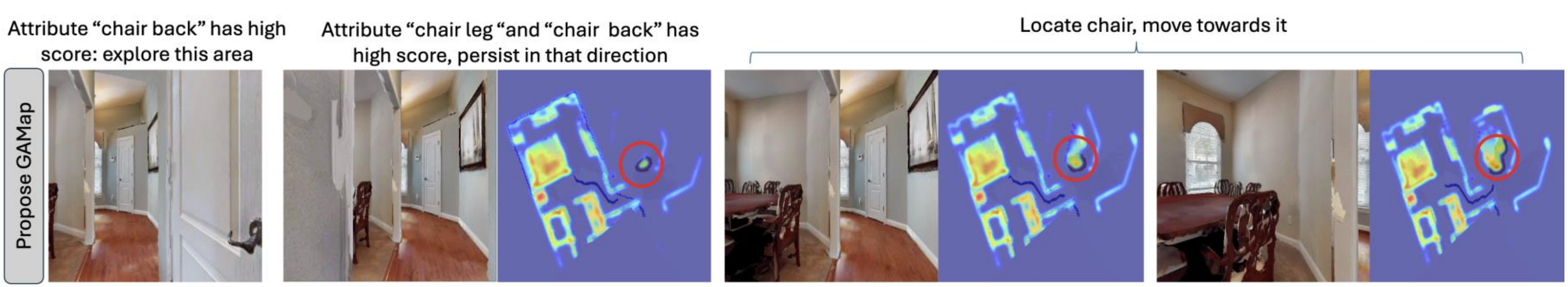
Observation: Prior art relies the zero-shot detector for categorical information understanding which often fall shot when only partial observation are given

Motivation: Human identify distinctive geometric parts or affordance attribute first when locating an object in an unfamiliar environment

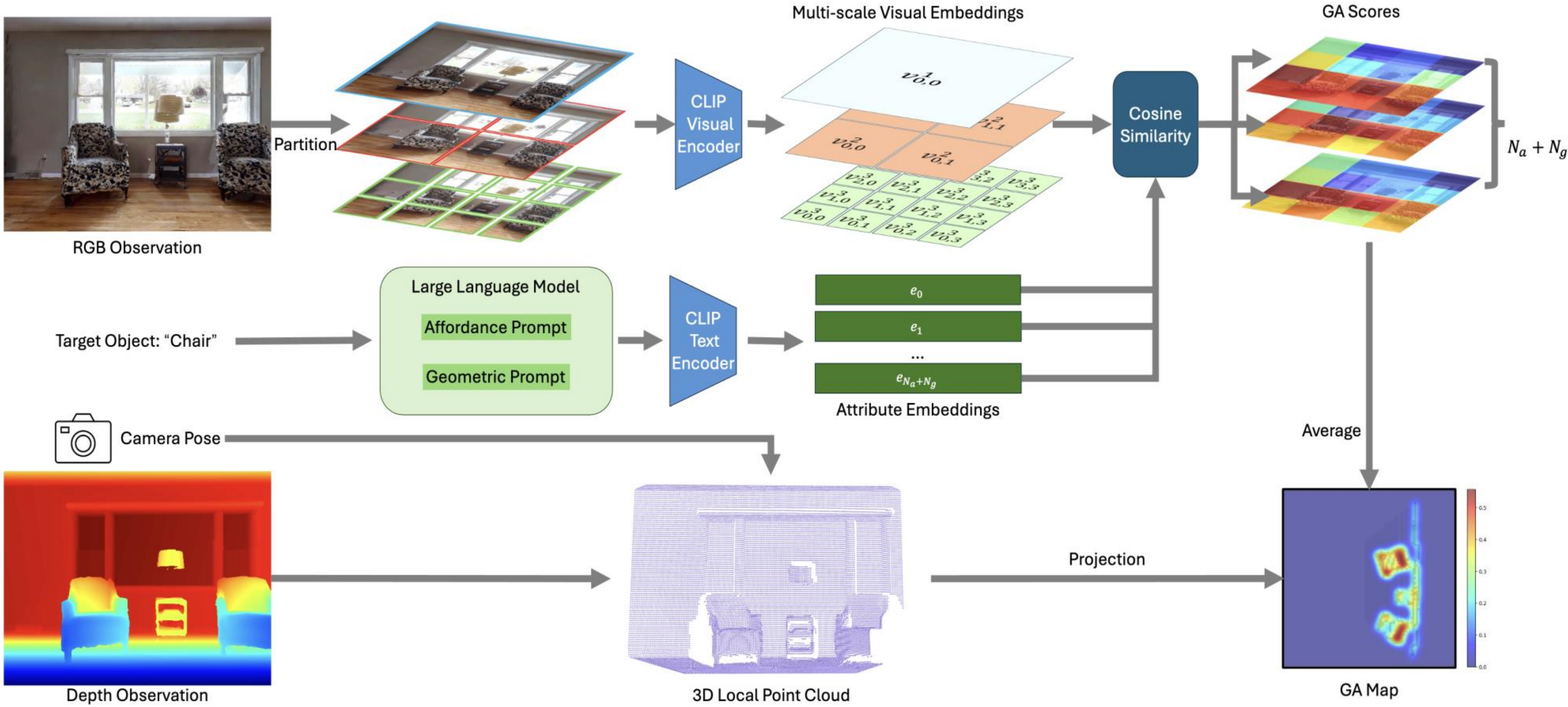
Observation: Prior art relies the zero-shot detector for categorical information understanding which often fall shot when only partial observation are given

Motivation: Human identify distinctive geometric parts or affordance attribute first when locating an object in an unfamiliar environment

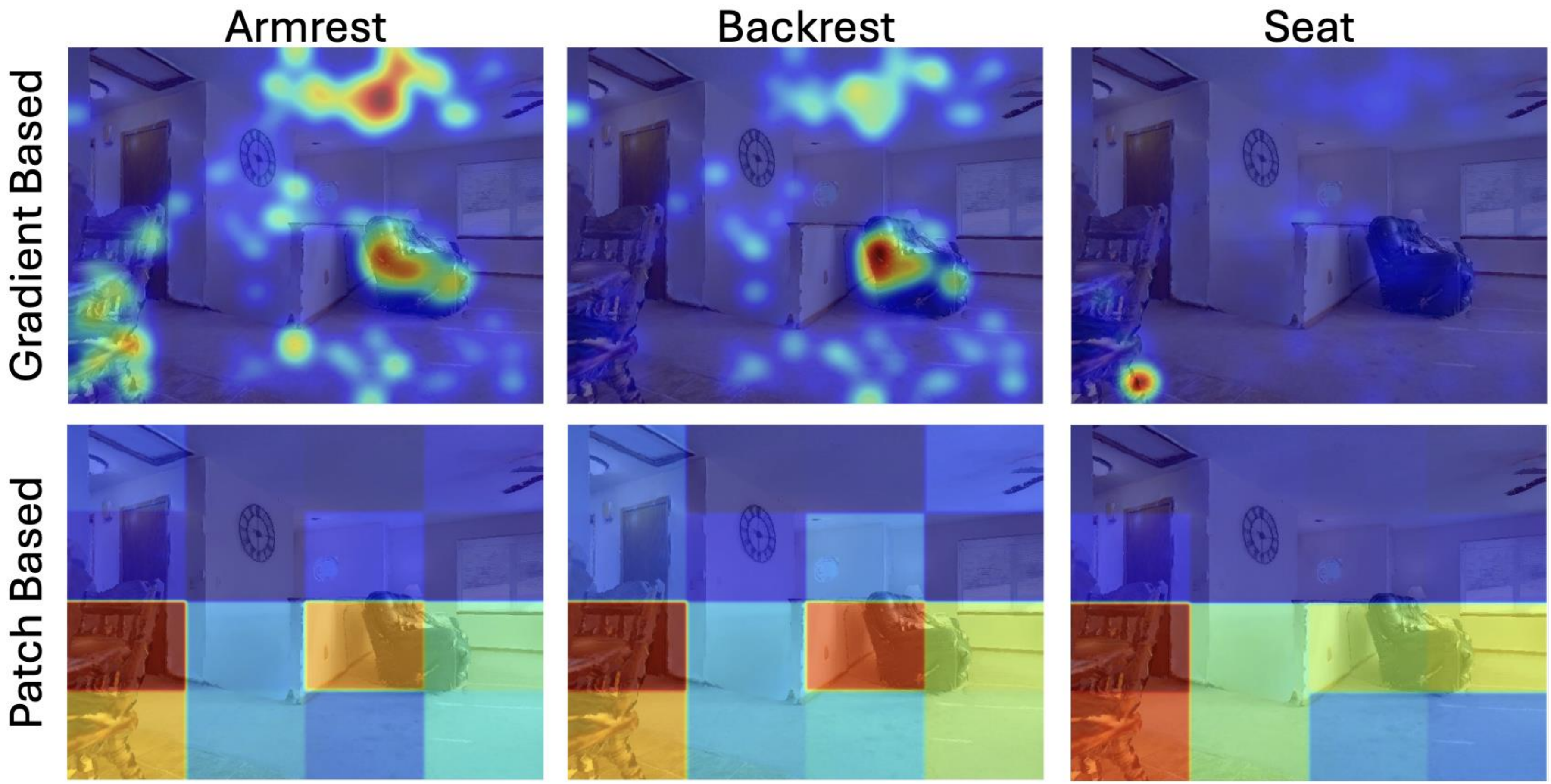
Our Solution: Multi-Scale Geometric Part and Affordance Map



Our Solution: Multi-ScaleGeometric Part and Affordance Map



The first line of images shows our pro- posed method, where the multi-scale approach effectively captures objects at all scales, such as the sofa back in the background. The second line of images shows the results of PIVOT-Liked GPT- 4V



GA score visualization between gradient-based and patch-based methods for the armrest, backrest, and seat attributes of a target chair. The gradient-based method (top row) often attends to irrelevant areas, such as the ceiling, while the patch- based method (bottom row) accurately focuses on the relevant areas

Table a: The table illustrates the differences between our work and existing methods.

Method	Mapping	Multi-Scale	Zero-shot	Training	
				Locomotion	Semantic
SemExp	Categorical	×	×	✓	✓
ZSON	Categorical	×	×	✓	✓
PixNav	Categorical	×	✓	✓	×
VLFM	Categorical	×	✓	✓	×
PONI	Categorical	×	✓	×	✓
L3MVN	Categorical	×	✓	×	✓
CoW	Categorical	×	✓	×	×
ESC	Categorical	×	✓	×	×
VoroNav	Categorical	×	✓	×	×
GAMap	Affordance+Geometric	✓	✓	×	×

Method	Reference	Zero-shot	Training		HM3D		Gibson	
			Locomotion	Semantic	SR↑	SPL↑	SR↑	SPL↑
SemExp	NeurIPS 20 [2]	×	✓	✓	37.9	18.8	65.2	33.6
ZSON	NeurIPS 22 [26]	×	✓	✓	25.5	12.6	31.3	12.0
PixNav	ICRA 24 [11]	×	✓	×	37.9	20.5	-	-
VLFM	CoRL 23 [17]	✓	✓	×	*52.5*	**30.4**	*84.0*	*52.2*
PONI	CVPR 22 [20]	×	×	✓	-	-	73.6	41.0
FBE	-	✓	×	✓	23.7	12.3	41.7	21.4
L3MVN	IROS 23 [21]	✓	×	✓	50.4	23.1	76.1	37.7
Random	-	✓	×	×	0.0	0.0	3.0	3.0
CoW	CVPR 23 [30]	✓	×	×	32.0	18.1	-	-
ESC	ICML 23 [5]	✓	×	×	38.5	22.0	-	-
SemUtil	RSS 23 [16]	✓	×	×	-	-	69.3	40.5
VoroNav	ICML 24 [18]	✓	×	×	42.0	26.0	-	-
GAMap	Proposed	✓	×	×	**53.1** (↑26.4%)	*26.0*	**85.7** (↑23.7%)	**55.5** (↑37.0%)

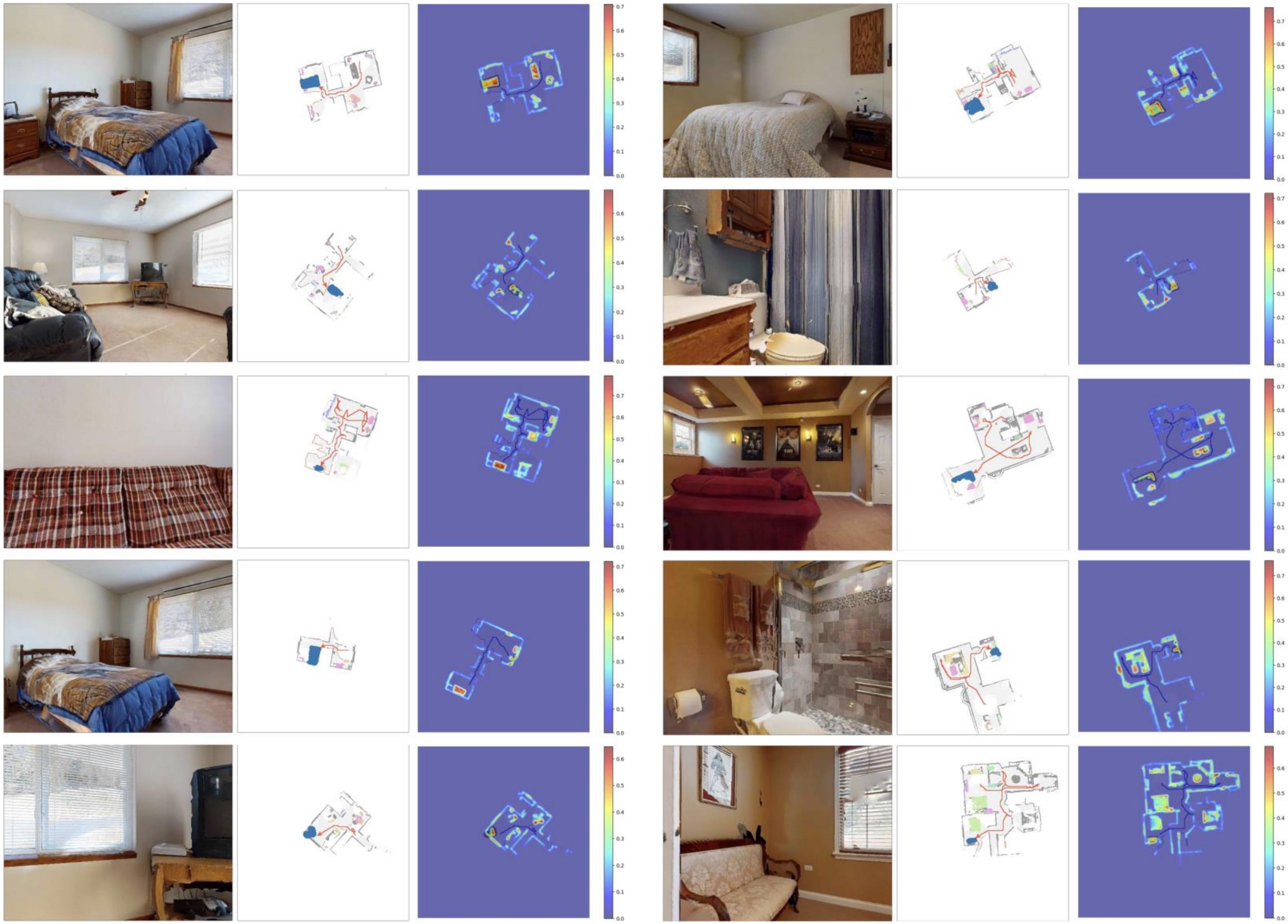


Figure 7: Visualized results of last observation frame, navigation path, and GAMap.

Thank you!