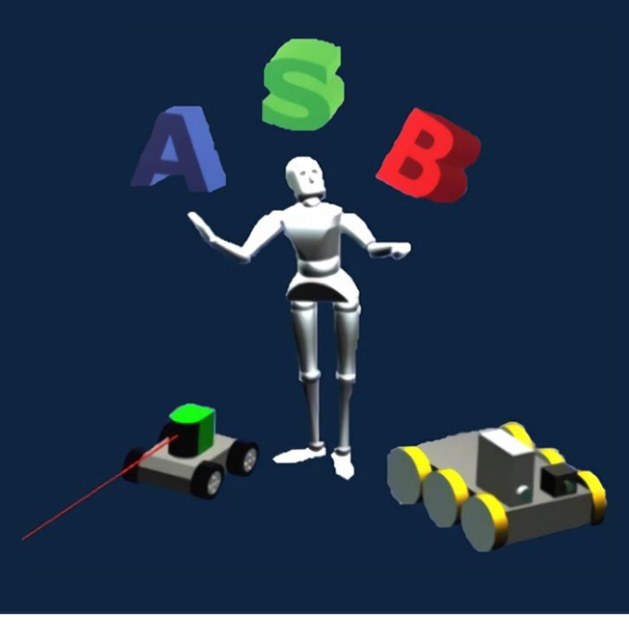
UNIVERSITY OF TORONTO
Robotics Institute



OLiVia-Nav: An Online Lifelong Vision Language Approach for Mobile Robot Social Navigation

LLHomeRobots @ Conference on Robot Learning
WORKSHOP ON LIFELONG LEARNING FOR HOME ROBOTS
9 NOVEMBER 2024 | TECHNICAL UNIVERSITY OF MUNICH | MUNICH, BAVARIA, GERMANY

Siddarth Narasimhan, Aaron Hao Tan*, Daniel Choi, Goldie Nejat

Introduction

Motivation: Mobile robots are often deployed in human-centered environments such as homes, hospitals or offices where they must navigate among humans while abiding by social norms. This is referred to as *robot social navigation***.**

Existing Approaches:

- Human Model Based (HMB): Human trajectories are explicitly predicted and integrated into a navigation policy using deep reinforcement learning (DRL)
- Human Model Free (HMF): Implicitly account for human trajectories using imitation learning (IL) or large foundation models (LLMs/VLMs)

Challenges: Integrating social context within the navigation policy, adapting to new social scenarios, real-time execution

Contributions

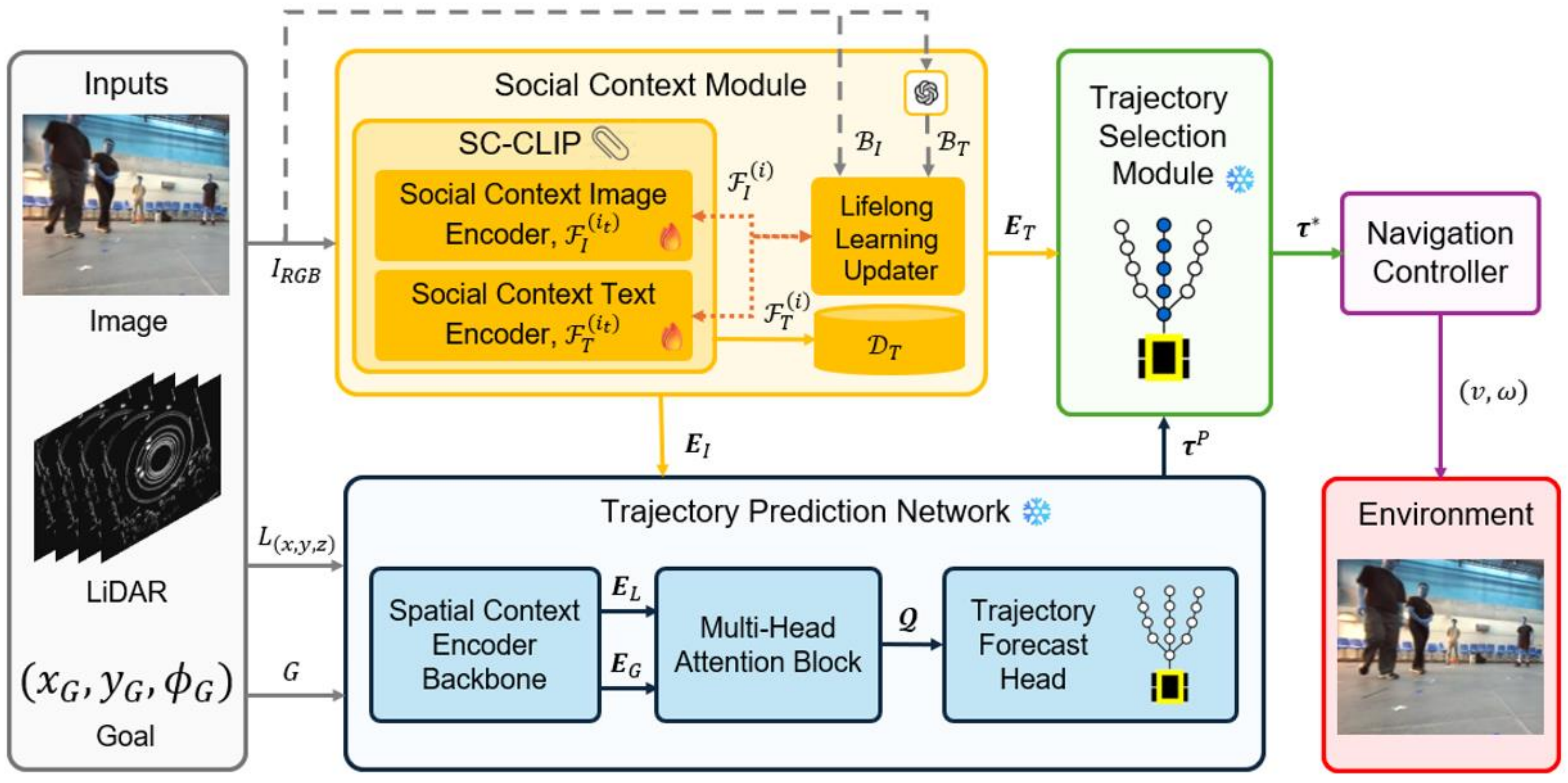
OLiVia-Nav is the first architecture which leverages the social reasoning of large VLMs, and the faster execution and lifelong learning capabilities of lightweight VLMs to generate social context embeddings to inform robot trajectory prediction.

Contributions

1) Development of a novel distillation process (SC-CLIP) to transfer the social knowledge of large VLMs to two lightweight encoders, which are used to generate social context embeddings. These encoders can adapt to new social scenarios through lifelong learning.

2) Development of a trajectory prediction network that uniquely uses multi-head attention to account for the social context during trajectory prediction.

OLiVia-Nav



1) Social Context Module (SCM): Extracts social context image and text embeddings using the social context image and text encoders (SCIE and SCTE) respectively for trajectory prediction and selection. SCIE and SCTE update as the robot encounters new data during deployment using the *Lifelong Learning Updater*.

2) Trajectory Prediction Network (TPN): Generates socially compliant navigation trajectories using LiDAR data, goal and social context image embeddings

3) Trajectory Selection Module (TSM): Selects the trajectory that follows the high-level navigation action encoded in the social context text embeddings.

4) Navigation Controller (NC): PID control to follow trajectory.

Training

Datasets:

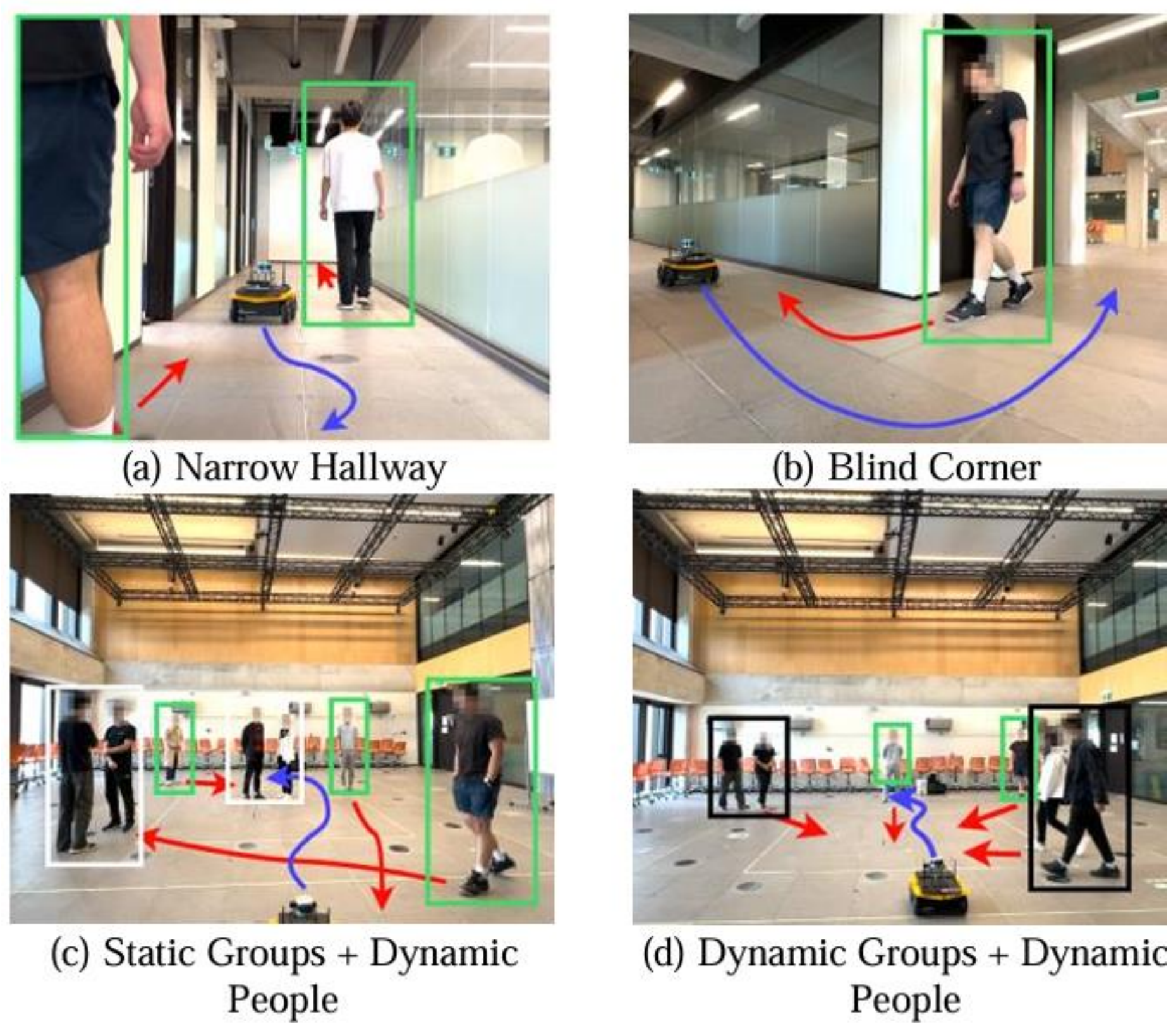
1) Social Context Dataset: 20K real-world RGB images, corresponding long and short text captions generated from GPT4o. SCIE and SCTE are trained using a cross-entropy loss function to align text captions with the images.

2) Trajectory Prediction Dataset: 5K expert trajectories, LiDAR point clouds and images. Used to train the TPN to predict diverse socially compliant trajectories.

3) Trajectory Selection Dataset: 20K predicted trajectories, and RGB images and text captions. Used to train the TSM.

Training GPU: NVIDIA H100 GPU (cloud) 80GB of VRAM, NVIDIA RTX 3090 32GB of VRAM

Experiments



(a) Narrow Hallway
(b) Blind Corner
(c) Static Groups + Dynamic People
(d) Dynamic Groups + Dynamic People

Benchmarks:

1) VLM-Social Nav: HMF method that uses a VLM with a cost-based planner.

2) MultiSoc: HMB method that uses DRL with attention.

Metrics:

1) Mean Squared Error

2) Hausdorff Distance

3) Personal Space Violation (PSV)

Results

Table 1: Comparison Results for the Four Social Scenarios

Scenario	Method	MSE ↓	Haus ↓	PSV (s) ↓
Narrow Hallway	**OLiVia-Nav**	**0.1075**	**0.7348**	**1.2**
	VLM-Social-Nav	0.1915	0.8785	1.9
	MultiSoc	0.2968	0.9095	2.1
Blind Corner	**OLiVia-Nav**	**0.0236**	**0.2572**	**0.4**
	VLM-Social-Nav	0.0755	0.5384	2.6
	MultiSoc	0.1021	0.4596	2.8
Static Groups + Dynamic People	**OLiVia-Nav**	**0.2195**	**0.7563**	**2.1**
	VLM-Social-Nav	0.4361	1.5579	3.5
	MultiSoc	0.2747	1.1081	3.2
Dynamic Groups + Dynamic People	**OLiVia-Nav**	**0.0733**	**0.4813**	**3.3**
	VLM-Social-Nav	0.1459	0.6832	4.7
	MultiSoc	0.1154	0.7929	4.5

↓ indicates that lower values are better.

Conclusion

Future Work

- Investigate the performance of OLiVia-Nav in larger environments with real crowds
- Evaluate the impact of the Lifelong Learning Update over longer periods of time

Full Paper + Video

Author Website

Presenter Website*