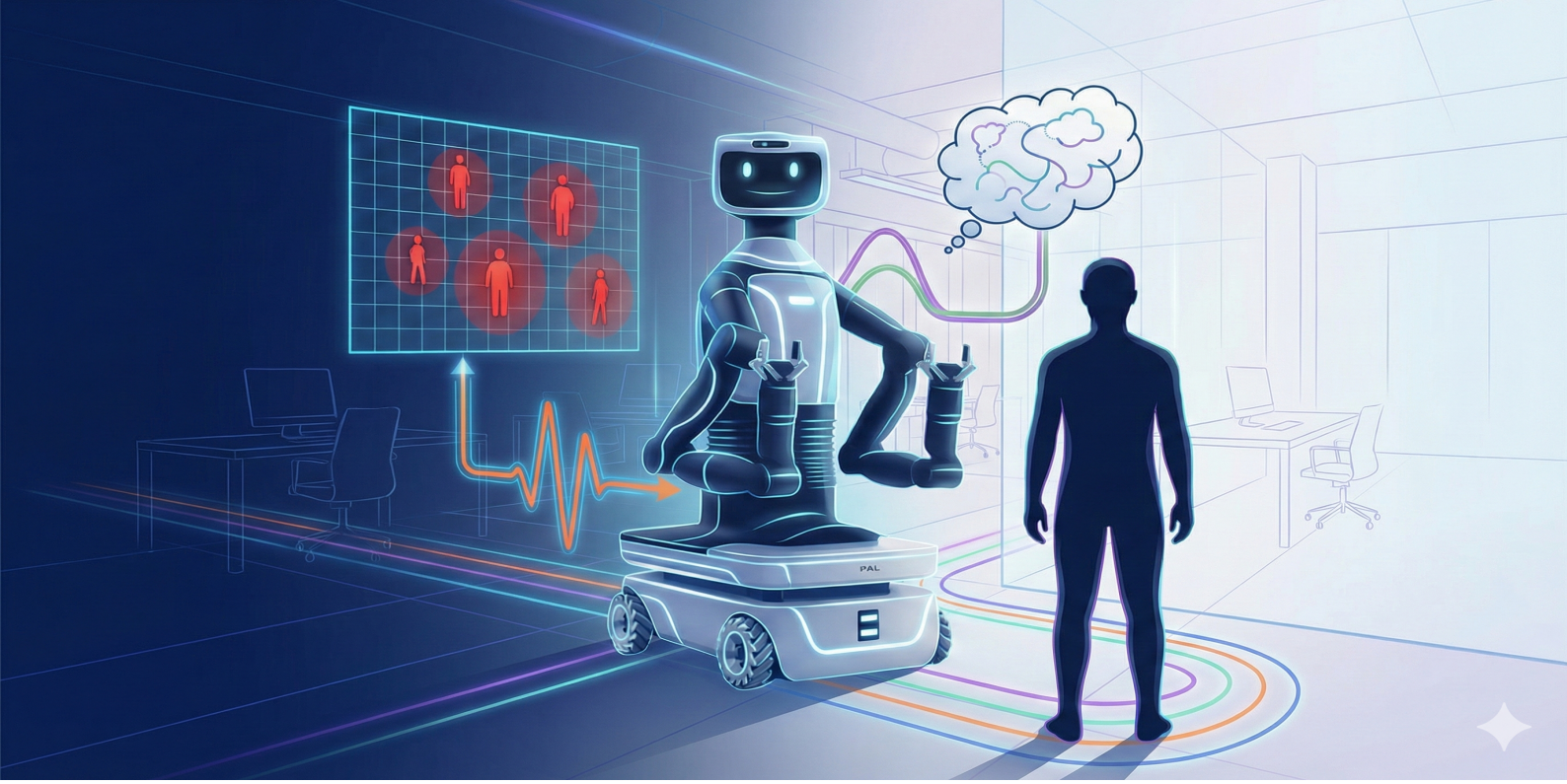

Social Navigation on the Tiago Pro

Using VLMs and YOLO for end-to-end Social Navigation

Endri Dibra, Daniel Jonatan Bank Dharampal, Nicklas Alexander Hougaard Deding

AAU - Department of Electronic Systems - Robotics Msc - ROB7-163 - November 2025

7th Semester mini Project

STUDENT REPORT

AALBORG UNIVERSITY

Electronic Systems
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Social Navigation on the Tiago Pro

Theme:
Using VLMs and YOLO for end-to-end Social Navigation

Project Period:
Fall Semester 2025

Project Group:
163

Participant(s):
Endri Dibra, Joel Thamby, Daniel Jonatan Bank Dharampal, Nicklas Alexander Hougaard Deding

Supervisor(s):
Professor Dimitris Chrysostomou

Copies: 1

Page Numbers: 8

Date of Completion:
27th of November 2025

Abstract:

As mobile robots increasingly operate in shared human environments, the ability to navigate with social compliance is critical for acceptance and safety. This project implements a hybrid social navigation stack for the Tiago Pro robot, integrating Vision-Language Models (VLMs) and YOLO-based detection within a ROS 2 framework. Adopting a "System 1 vs. System 2" paradigm, the architecture combines high-frequency, proxemic-based costmap updates for immediate safety with low-frequency VLM queries for adaptive parameter tuning and natural language goal translation. The system is evaluated against a standard Nav2 baseline in simulation and real-world experiments across four dynamic scenarios, including frontal approaches and narrow doorways, to assess Personal Space Compliance (PSC) and navigational efficiency.

Introduction

As mobile robots enter the social space designed for humans, the need for reliable social navigation becomes increasingly pervasive. Social navigation is the incorporation of social dynamics into robot navigation. This will enable robots to navigate among humans in an efficient, trustworthy, and safe way. Various studies have shown that robots that incorporate some level of social navigation are much more accepted than robots that do not [1].

The fundamental challenges of implementing social navigation lie in quantifying socially compliant navigation and being able to validate and compare navigation policies on a large scale with diverse scenarios and edge cases [1].

One of the core concepts in social navigation is how humans occupy space and how spatial distance can influence human behaviour. This concept is referred to as proxemics [2]. There are two main distances to be aware of: the intimate distance and the personal distance. Ignoring some complexity for a more simplified model, these distances can be estimated at 0.45m and 1.2m, respectively [1].

Related Works

The SOTA in social navigation has shifted away from handcrafted algorithms, often relying on explicitly defined rules and cost functions (e.g., DWA, Social Force) and standard learning-based policies (e.g., HSAC-LLM) toward **Foundation Models**, specifically Vision-Language-Action (VLA) models, which overcome previous limitations in generalisation and contextual understanding [3]. Leveraging massive pretraining data, foundation models can generalise better to unseen scenarios [4].

Leading examples include: VLM-Social-Nav [5], NavFoM [3], and SmolVLA [6].

VLM-Social-Nav is a VLA approach to social navigation. They use GPT-4V [7] to generate symbolic actions in line with how a human would navigate, given a set of constraints in the prompt. They incorporate these actions into an optimisation-based planner for the robot [5].

SmolVLA is a small end-to-end pre-trained VLA model by LeRobot, built on top of SmolVLM with half the layers truncated. They then added an action expert, which allows low-latency async action inference [6].

Methods

The main contribution of this paper is a ROS 2 Package that integrates social navigation on the Tiago Pro. The ROS 2 simulation stack provided by Pal Robotics has been used as a good default stack without much need for modification [8].

3.1 Implementation Details

To enable socially compliant navigation, this project integrates 3 core topics into the stack.

The 1st one is high-frequency, proxemic-based human update to the costmaps on the robot.

The 2nd one is the integration of a lower-frequency navigation parameter tuning VLM call to tune speed, inflation, and cost scaling.

And the 3rd is a goal translation call. The user will give verbal commands, and the robot needs to understand what each instruction means and translate it to velocity or navigation goals.

With these implementations, the Tiago Pro will have stronger social compliance. This approach lies between end to end learning based foundation models and handcrafted algorithms, by defining the behaviour but also allowing flexible control from strong foundation models for goal and behaviour tuning.

3.2 Navigation Stack

The default navigation stack is built on ROS 2 Humble and Nav2. The core modification we had to do to optimise latency and hardware compatibility for our purpose was to change the default laser odometry node to a more primitive wheel odometry model. This modification has only been done to the simulation and not the physical robot.

The robot uses an MPPI (Model Predictive Path Integral) Local Planner with a 56-step horizon at 20Hz. For the global planner, the robot uses the 2D SMAC Planner (A variation on A*).

The local costmap is a rolling 5x5 map around the robot populated by laser scans and inflation. The global costmap is generated with the SLAM-Toolbox in ROS 2.

3.3 Latency and System Architecture

Through this project, various latency considerations have been taken into account. Finetuning test has been run on SmolVLA with the SCAND [9] (10 hours of teleoperated socially compliant robot data). The finetuned model could generate trajectories at 2Hz at high-end consumer hardware. This approach would be able to solve social navigation, but at a slower refresh rate, and being end-to-end, it can become challenging to integrate a faster action system to account for the latency.

The chosen approach for this paper incorporates Kahneman's System 1 vs System 2 thinking paradigm, by detecting and integrating humans into the costmaps, making this faster-acting System 1 able to handle faster changes more reliably. For a slower planning System 2, the approach presented here is a more modular integration of VLMs into the tuning and goal-setting parts of the system. By giving a large vision foundation model, the executive decision of translating user input to Nav2 Parameters and goals, the combined stack is able to handle more complex user queries and scenarios.

Experiments

To thoroughly test the combined system presented in this paper, two main tests will be conducted. A simulation-based test and a real-world test.

4.1 Testing Scenarios

The Nav2 baseline stack will be tested against our proposed method in 4 core scenarios adopted from the VLM-Social-Nav paper [5]. The scenarios are as follows:

1. **Frontal Approach**: The robot and a human approach each other on a straight trajectory. The robot is expected to yield or slow down, not obstruct the human path, and keep to the right.
2. **Frontal Approach with Gesture**: Same as the Frontal Approach, but the human will signal the robot to stop. The robot is expected to yield by interpreting the human gesture.
3. **Intersection**: The robot and a human will cross each other on perpendicular trajectories. The robot is expected to drive slowly, approaching the human. Modifying or stopping to go behind the human to not obstruct the path.
4. **Narrow Doorway**: The robot and a human approach each other, moving through a narrow doorway. The robot is expected to wait outside the door and yield to the human.

The scenario approach to testing is chosen as a best practice in this field following recommendations from Francis et al. [1]. These specific scenarios are chosen to allow better comparison with VLM-Social-Nav.

4.2 Simulation and Real-World

To allow for safe testing in the real world, a simulation-based testing environment will be set up. This will also establish ground truth evaluation metrics: *Personal Space Compliance* (PSC), *Time-to-Collision* (TTC), and minimum distance to humans.

The real-world tests will be conducted 5 times for each scenario, each time with the same setup.

4.3 Results

We expect our system to pass the simulation tests, but potentially have unexpected behaviour or other implementation issues in the real-world tests.

Discussion & Conclusion

This part of the paper will dive into a discussion of the impact and results of the testing and approach proposed in this paper.

5.1 Furture Work

Our testing with SmolVLA was promising, but it needs more work to properly integrate into a responsive system. The end-to-end foundation models show a promising path forward, and the integration of these into robotic social navigation will likely show promising results in the future [4].

5.2 Conclusion

Everything is Awsome!

Bibliography

[1] A. Francis, C. Perez-D'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, February 2025, also available as arXiv:2306.16740.

[2] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 137–153, 2015. [Online]. Available: https://link.springer.com/article/10.1007/s12369-014-0251-1

[3] J. Zhang, A. Li, Y. Qi, M. Li, J. Liu, S. Wang, H. Liu, G. Zhou, Y. Wu, X. Li, Y. Fan, W. Li, Z. Chen, F. Gao, Q. Wu, Z. Zhang, and H. Wang, "Navfom: Embodied navigation foundation model," *arXiv preprint arXiv:2509.12129*, 2024.

[4] NVIDIA, :, J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu, "Gr00t n1: An open foundation model for generalist humanoid robots," 2025. [Online]. Available: https://arxiv.org/abs/2503.14734

[5] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 1–8, 2025, also available as arXiv:2404.00210.

[6] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint*, vol. arXiv:2506.01844, 2025, accessed: 3 Nov. 2025. [Online]. Available: https://arxiv.org/pdf/2506.01844

[7] OpenAI, "Gpt-4v(ision) system card," Sep. 2023, accessed: 2025-11-23. [Online]. Available: https://openai.com/index/gpt-4v-system-card/

[8] PAL Robotics, *TIAGo Pro Documentation*, PAL Robotics, Jan. 2025, accessed: 2025-11-23. [Online]. Available: https://docs.pal-robotics.com/25.01/tiagopro

[9] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. W. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *arXiv preprint*, vol. arXiv:2203.15041, 2022, accessed: 3 Nov. 2025. [Online]. Available: https://www.cs.utexas.edu/~xiao/SCAND/SCAND.html