A Study on Learning Social Robot Navigation with Multimodal Perception

Bhabaranjan Panigrahi, Amir Hossain Raj, Mohammad Nazeri, and Xuesu Xiao

***Abstract*— Autonomous mobile robots need to perceive the environments with their onboard sensors (e.g., LiDARs and RGB cameras) and then make appropriate navigation decisions. In order to navigate human-inhabited public spaces, such a navigation task becomes more than only obstacle avoidance, but also requires considering surrounding humans and their intentions to somewhat change the navigation behavior in response to the underlying social norms, i.e., being socially compliant. Machine learning methods are shown to be effective in capturing those complex and subtle social interactions in a data-driven manner, without explicitly hand-crafting simplified models or cost functions. Considering multiple available sensor modalities and the efficiency of learning methods, this paper presents a comprehensive study on learning social robot navigation with multimodal perception using a large-scale real-world dataset. The study investigates social robot navigation decision making on both the global and local planning levels and contrasts unimodal and multimodal learning against a set of classical navigation approaches in different social scenarios, while also analyzing the training and generalizability performance from the learning perspective. We also conduct a human study on how learning with multimodal perception affects the perceived social compliance. The results show that multimodal learning has a clear advantage over unimodal learning in both dataset and human studies. We open-source our code for the community's future use to study multimodal perception for learning social robot navigation.**[1]

I. INTRODUCTION

Thanks to decades of robotics research [1], autonomous mobile robots can navigate collision-free in environments like factories and warehouses using onboard sensors such as LiDARs and RGB cameras [2]. However, deploying these robots in human-inhabited public spaces complicates navigation, as they must also consider human interactions and adapt to social norms [3].

Machine learning offers a solution by capturing complex human-robot interactions in a data-driven manner, relieving roboticists from manually designing models [4], [5]. The availability of onboard computation and extensive perception data accelerates the use of learning methods in social navigation. Most modern robots utilize multiple sensors, with LiDARs and RGB cameras being prevalent. While LiDARs provide high-resolution geometric data, RGB cameras offer rich semantic information. Both types of data are crucial for decision-making in social navigation, as geometric structures must be avoided and semantic cues can reveal human intentions [6], [7].

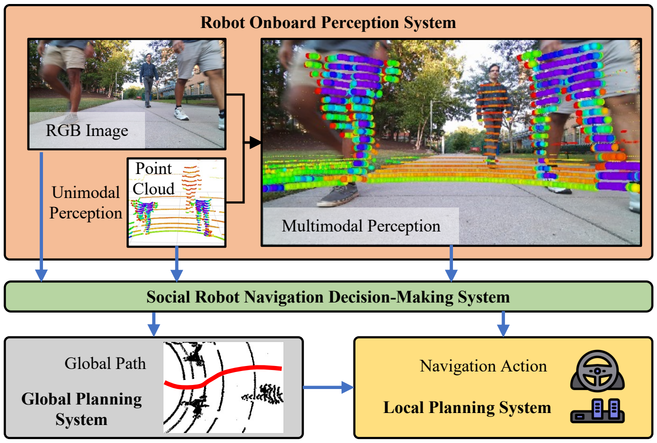

Fig. 1: Social Robot Navigation Decision Making on the Global and Local Level with Multimodal and Unimodal (RGB Image and Point Cloud) Perception Input.

This paper presents a study on using multimodal perception from LiDAR and RGB cameras to enhance robot decision-making during navigation. Conducted on the large-scale Socially Compliant Navigation Dataset (SCAND) [8], we explore the decision-making capabilities of multimodal versus unimodal learning in various social scenarios and assess their training performance. Our findings indicate that multimodal learning is more reliable and robust than unimodal approaches. To our knowledge, this is the first comprehensive study demonstrating the advantages of multimodal perception in learning social robot navigation using real-world data and human interactions.

II. RELATED WORK

We review related work in social robot navigation, machine learning for navigation, and multimodal learning.

A. Social Robot Navigation

Collision-free navigation has been a focus for robotics for decades [9], [10], with early examples like the museum tour-guide robots RHINO [11] and MINERVA [12]. Rather than treating humans as mere obstacles [13], researchers have modeled human movement uncertainty [14] and established social norms for navigation [15], developing planners that account for these factors. Physics-based models [16] incorporate behavioral features such as proxemics [17], intentions [18], and social contexts [19]. However, a simple model often fails to capture complex human behaviors, which vary by context, such as rush hours or weekends. These factors must be extracted from raw data from onboard sensors like LiDARs and RGB cameras, posing challenges for

All authors are with the Department of Computer Science, George Mason University {bpanigr, araj20, mnazerir, xiao}@gmu.edu

[1] GitHub: https://github.com/RobotiXX/multimodal-fusion-network/

perception algorithms in human tracking, motion prediction, and intention detection. Consequently, the rise of machine learning has led to data-driven approaches for social robot navigation [20].

B. Machine Learning for Navigation

To address these challenges, machine learning approaches have been used to encode the complexities of human social behaviors in a data-driven way [20], also tackling navigation issues like off-road navigation [21]. These methods include learning representations, costmaps [22], navigation planner parameterizations [23], local planners [24], and end-to-end navigation policies that convert raw perceptions into motor commands [25]. In terms of machine learning, reinforcement learning [26] and imitation learning [27] rely on simulated trial-and-error data and expert demonstrations, respectively. Given the challenges of generating high-fidelity perceptual data and realistic human-robot interactions in simulations, this study employs imitation learning, specifically Behavior Cloning (BC) [28], using a large-scale social robot navigation demonstration dataset.

C. Multimodal Learning

Recent research shows that multimodal learning frameworks can enhance performance in downstream tasks [29]. In autonomous mobile robot navigation, multimodal graph neural networks combining RGB cameras, LiDARs, and odometry have successfully navigated unstructured terrains, demonstrating robustness against occlusion and unreliable data [30]. Studies have also fused laser data, RGB images, point clouds, and distance maps for navigation in time-sensitive scenarios, revealing that multimodal networks outperform those using only RGB images and distance maps [31]. Despite the effectiveness of multimodal perception in navigation, research on its impact on decision-making in social robot navigation is limited, which this study aims to investigate, focusing on learning from multimodal perception inputs rather than multimodal distribution models [29], [32].

*D. Socially Compliant Robot Navigation Dataset (*SCAND*)*

Our study utilizes the open-source, large-scale social robot navigation dataset, SCAND [8], which includes 8.7 hours of data, 138 trajectories, and 40 kilometers of socially compliant, human-teleoperated driving demonstrations. This dataset features multimodal streams, including 3D LiDAR, visual and inertial information, robot odometry, and joystick commands, collected from two different mobile robots—a Boston Dynamics Spot and a Clearpath Jackal—by four human demonstrators in both indoor and outdoor environments. Given its rich social interactions and multimodal perception-to-action navigation decisions, SCAND is ideal for examining social robot navigation with multimodal perception. Specifically, we focus on the impact of 3D LiDAR point cloud data and RGB images—two commonly used perception modalities—since their geometric and semantic information can complement each other, aiding decision-making in human-inhabited public spaces.

III. MULTIMODAL LEARNING FOR SOCIAL ROBOT NAVIGATION

We adopt an imitation learning approach, i.e., BC, to learn socially compliant navigation decisions using multimodal perception from SCAND. Similar to classical navigation systems with a global and a local planning system, we design our multimodal learning framework so that it will produce both global and local plans and study how multimodal and unimodal learning can imitate the navigation decisions made by the human demonstrator on both global and local levels.

A. Problem Formulation

Specifically, at each time step t of each trial in SCAND, the robot receives onboard perceptual input, including a sequence of 3D LiDAR point cloud data L and RGB images I, and a goal G it aims to reach, which is taken as a waypoint 2.5m away from the robot on the future robot odometry. We denote all these inputs necessary to inform the decision-making process during social robot navigation as a navigation input: $\mathcal{I}_t^D = \{L_k^D, I_k^D, G_t^D\}_{k=t-N+1}^{t}$, where N denotes the history length included in the navigation input at t and D denotes that the data is from the SCAND demonstrations.

Facing a social navigation input $\mathcal{I}_t^D$, the SCAND demonstrator shows the desired, socially compliant navigation decision $\mathcal{D}_t$ on both global and local levels: P_t is the demonstrated global plan, recorded as the human-driven future robot odometry starting from time t, and takes the form of a sequence of 2D waypoints $P_t^D = \{(x_i^D, y_i^D)\}_{i=t}^{t+M-1}$; A_t is the demonstrated local plan represented as a sequence of joystick action commands $A_t^D = \{(v_i^D, \omega_i^D)\}_{i=t}^{t+K-1}$, where v and ω is the linear and angular velocity respectively. M and K denote the length of the navigation decision on the global and local plan level respectively. The demonstrated navigation decision is therefore defined as $\mathcal{D}_t^D = \{P_t^D, A_t^D\}$.

Producing the navigation decision $\mathcal{D}_t^D$ based on $\mathcal{I}_t^D$ as input, a navigation system is defined as a combination of two functions, $\mathcal{F}^g(\cdot)$ and $\mathcal{F}^l(\cdot)$, responsible of generating the global plan P_t^D and local plan (action) A_t^D:

$$
\begin{aligned}
P_t^D &= \mathcal{F}^g(\mathcal{I}_t^D), \\
A_t^D &= \mathcal{F}^l(\mathcal{I}_t^D, P_t^D).
\end{aligned}
$$

In a data-driven manner, we instantiate both global and local planners by learning $\mathcal{F}_\theta^g(\cdot)$ and $\mathcal{F}_\phi^l(\cdot)$ as deep neural networks with learnable parameters θ and ϕ respectively. In particular, we aim to learn the parameters to minimize a BC loss:

$$
\begin{aligned}
\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmin}} & \sum_{P_t^D, A_t^D, \mathcal{I}_t^D \in \text{SCAND}} \qquad (1) \\
& \left[||P_t^D - \mathcal{F}_\theta^g(\mathcal{I}_t^D)|| + \lambda ||A_t^D - \mathcal{F}_\phi^l(\mathcal{I}_t^D, \mathcal{F}_\theta^g(\mathcal{I}_t^D))|| \right],
\end{aligned}
$$

where the first term is the difference between demonstrated and learned global plan, while the second term is for the local plan, with λ as a weight between them.

In this study, we are interested in studying the effect of including different perception modalities in $\mathcal{I}_t$ on making socially compliant navigation decisions P_t and A_t. We

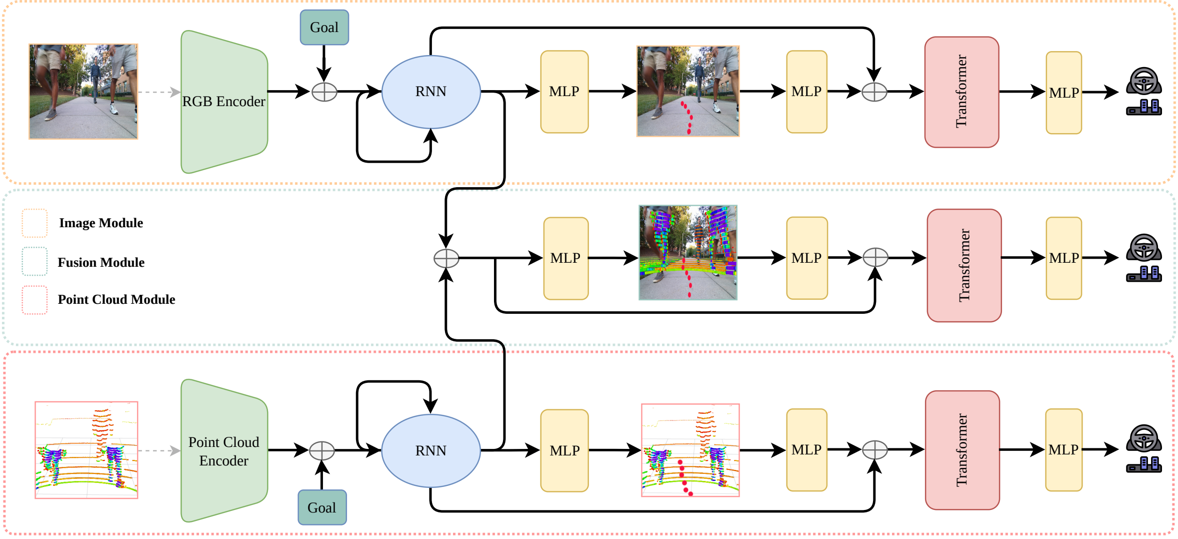

Fig. 2: Image Module, Fusion (Multimodal) Module, and Point Cloud Module Architecture for Social Robot Navigation.

study three scenarios, i.e., multimodal perception $\mathcal{I}_t^{\text{MM}} = \{L_k, I_k, G_t\}_{k=t-N+1}^{t}$, unimodal LiDAR (point cloud) perception $\mathcal{I}_t^{\text{LiDAR}} = \{L_k, G_t\}_{k=t-N+1}^{t}$, and unimodal vision (RGB image) perception $\mathcal{I}_t^{\text{Vision}} = \{I_k, G_t\}_{k=t-N+1}^{t}$. For simplicity and consistency, we keep $N = 1$ for all three cases in this study and leave an investigation into different history lengths as future work.

B. Unimodal Perception

1) Point Cloud Modality: We take points that are within the range of 8 meters in front, 3 meters on either side and within 2.5 meters of height from the robot as perceived by the 3D LiDAR. All points are placed into their respective voxel inside a 3D voxel grid with $5 \times 5 \times 5$cm voxels, resulting in a $160 \times 120 \times 50$ voxel representation for L_k. We use a 3D Convolution Neural Network (CNN) [33] to process the voxel representation to extract meaningful information for our downstream social robot navigation task. The point cloud encoder is shown as the green trapezoid in the red box at the bottom of Fig. 2.

2) RGB Modality: For RGB images, we take a $224 \times 224 \times 3$ image from the camera as input. We use ResNet-18 [34] to extract features for our social robot navigation task. The image encoder is shown as the green trapezoid in the yellow box at the top of Fig. 2.

Both RGB and point cloud inputs have separate decision-making modules, depicted in the upper yellow and lower red boxes of Fig.2. To ensure a fair comparison, both share the same architecture, differing only in input modality. We concatenate embeddings from the input encoders with the local goal (2.5m away) and feed them into a Recurrent Neural Network (RNN) to capture historical information (blue ellipsoids in Fig.2). A Multi-Layer Perceptron (MLP) (yellow boxes) generates a global plan as a sequence of 2D waypoints (red dots), which is then combined with the RNN output, processed by a transformer, and passed through another MLP to produce the local plan, defining actions for linear and angular velocities.

C. Multimodal Fusion

For multimodal fusion, the outputs of the RNNs from the point cloud and image modules are concatenated and passed through the fusion process, shown in Fig. 2 middle. Similar to the unimodal modules, our feature fusion also happens at two different places in our multimodal network. Each fusion caters to different downstream tasks, i.e., producing both global and local plans.

D. Navigation Decisions and Loss Functions

The global navigation decisions are instantiated as a sequence of five future waypoints ahead of the robot, i.e., $P_t^D = \{(x_i^D, y_i^D)\}_{i=t}^{t+4}$ $(M = 5)$, each of which is 0.5m apart taken from the future robot odometry. The local navigation decisions take the form of the current linear and angular velocity commands, i.e., $A_t^D = \{(v_t^D, \omega_t^D)\}$ $(K = 1)$.

For the first and second loss terms in Eqn. 1, we use $L2$-norm of the five future waypoints and $L1$-norm of the current angular and linear velocity. We set $\lambda = 1$.

E. Design Choices

Design choices for neural network hyperparameters and architecture were optimized through extensive trial-and-error for fair comparison across modalities. Detailed hyperparameters are provided in our open-source implementation.

We tested PointNet [35] and PointNet++ [36] as point cloud encoders, but they performed poorly in the SCAND scenarios due to high variability. Converting point clouds to a voxelized grid and using a 3D CNN significantly improved performance.

For local planning, a simple MLP struggled to adapt to varying velocities based on human proximity. In contrast, a transformer architecture effectively captures these variations through attention mechanisms.

IV. SCAND Study Results

We first present our study results on all the social scenarios in SCAND before presenting our human study results. We divide the SCAND trials into 18 for training and 8 for testing.

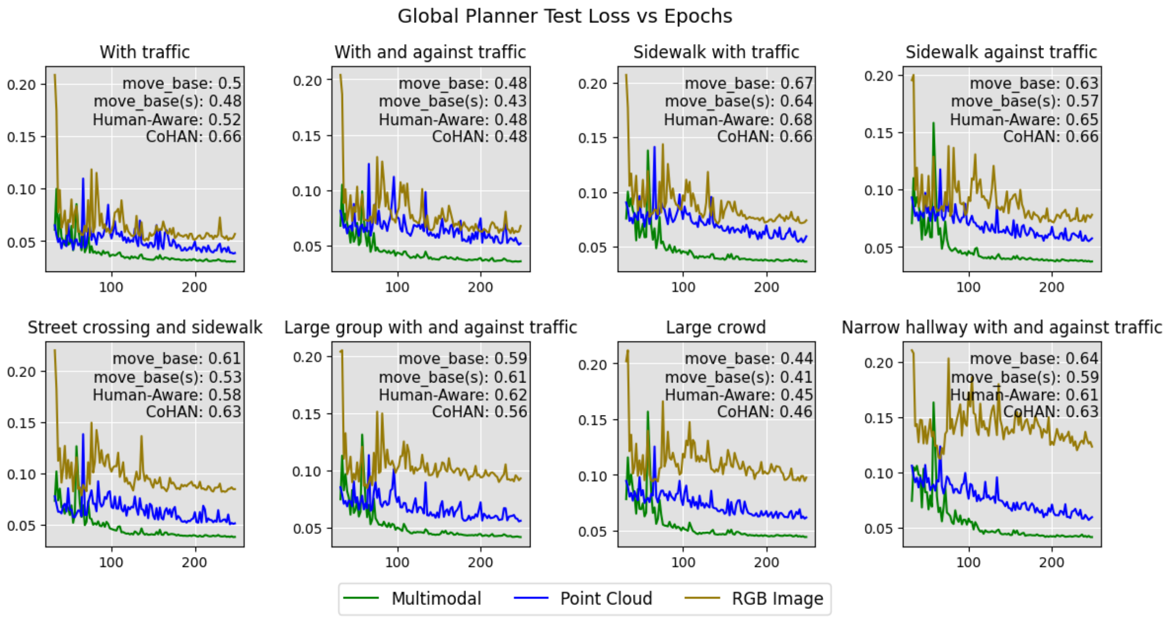

Fig. 3: Test Loss on Eight SCAND ROSBAGS with Multimodal, Point Cloud, and RGB Image Input (Averaged Over Three Training Runs with Negligible Variances Invisible in the Figures). Losses of four Classical Approaches are for Comparison.

We analyze the learning results on the test data from both the machine learning and social robot navigation perspectives. The training loss curves for the global planner in terms of L1 loss on the eight SCAND ROSBAGS are shown in Fig. 3, while the local planner loss in Fig. 4. We also plot the performance of a variety of classical social robot navigation planners using the same loss function between their output and the SCAND demonstration to compare against end-to-end learned policies.

A. *Multimodal Learning Performance*

The results of the eight test SCAND ROSBAGS are arranged to reflect increasing performance discrepancies among modalities, serving as a rough indicator of the ”difficulty” level in social robot navigation. Loss values for most modalities converge faster and to lower points in the earlier ”easy” trials (upper left) compared to the later ”difficult” ones (lower right).

Notably, multimodal perception significantly outperforms both unimodal modalities in global planning. The green curves representing multimodal loss drop more rapidly, converge sooner, and achieve lower values than the yellow (RGB) and blue (point cloud) curves. Furthermore, the multimodal learning curves are consistent across all eight test SCAND ROSBAGS, highlighting its advantages.

For unimodal types, point cloud perception consistently outperforms RGB images in all trials, although both underperform compared to multimodal learning. In earlier ”easier” trials, point cloud slightly surpasses RGB, but in the later ”difficult” trials, the gap narrows, with point cloud curves approaching those of multimodal learning.

Due to the lack of significant differences in local planning loss across the eight test SCAND ROSBAGS, we combine all curves for each modality into a single figure in Fig. 4. The trends remain similar: multimodal learning achieves slightly better performance than point cloud learning, which in turn outperforms RGB image learning.

B. *Multimodal Social Compliance*

In addition to machine learning statistics, we analyze how each perception modality performs in various social interactions. The performance of RGB images declines significantly from the first to the last test trial, as shown in Fig. 3, resulting in more than double the loss value. In contrast, multimodal and point cloud learning maintain consistent performance. The majority of social scenarios in each test SCAND ROSBAG are noted at the top of each subfigure in Fig. 3.

We find that the increasing ”difficulty” level, particularly for RGB images, correlates with higher human density in confined spaces. While RGB learning performs similarly to point cloud and multimodal learning in the simpler ”with traffic” scenario, performance deteriorates in more complex situations, such as ”against traffic” or ”narrow hallway,” due to increased variance in the RGB input. This degradation likely stems from the lack of direct geometric information in RGB images, making mobile robots less safe in crowded environments compared to point clouds, which provide essential geometric data for ensuring safety.

The gap between multimodal and point cloud learning is notable. Although both perform similarly across all tests, multimodal learning consistently achieves lower loss values and fewer epochs to converge. The additional semantic information from RGB images, combined with the geometric data from point clouds, offers relevant social cues that enhance decision-making in social navigation. This underscores the importance of incorporating semantic information in social robot navigation, which goes beyond simply avoiding obstacles, as seen in traditional mobile robot navigation.

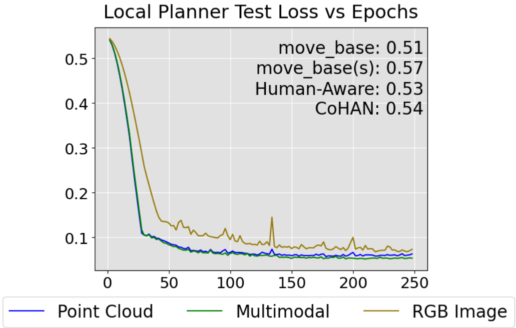

Fig. 4: Average Test Loss on All SCAND ROSBAGS with Multimodal, Point Cloud, and RGB Image Input (Three Training Runs).

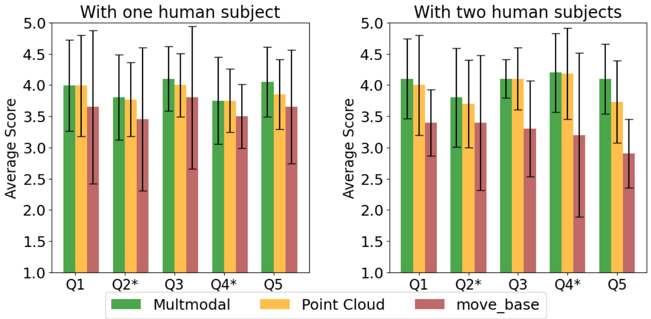

Fig. 5: Human Study Results.

V. HUMAN STUDY RESULTS

We conduct a human study to test whether the findings from our SCAND study can translate to real-world social robot navigation. We use a Clearpath Jackal robot with a Velodyne VLP-16 LiDAR and a ZED2 RGB-D camera for the point cloud and RGB image input respectively. We recruit eight human subjects for our human study.

Two sets of experiments are designed according to a previous protocol to evaluate social robot navigation [37]: frontal approach of the robot with one and two human participants in a public outdoor space (Fig. 6). In the one-human study, participants are instructed to take a natural path towards the robot; Participants in the two-human study are instructed to take three different approaches to initiate social interactions: move directly towards the robot, move forward then diverge, and move towards one side of the robot. After deploying the RGB module, we found that the robot may move dangerously close to the human subjects. Therefore, we exclude the RGB module in the human study.

After each human-robot interaction, we ask the participant to fill in a standard questionnaire [37] with five questions[2]: *1. The robot moved to avoid me*, *2. The robot obstructed my path**, *3. The robot maintained a safe and comfortable distance at all times*, *4. The robot nearly collided with me**, and *5. It was clear what the robot wanted to do*.

The per-question average along with error bars are plotted in Fig. 5 for both the one-person (left) and two-person scenarios (right). For all five questions, the multimodal learning approach is able to consistently achieve higher social compliance scores with smaller variance, compared to move_base, the best classical planner according to the loss values in the SCAND study. Compare the left and right figures, the difference between multimodal learning and move_base increases with more humans, showing multimodal learning's potential to enable socially compliant navigation with higher human density in public spaces, which is consistent with the results we observe in terms of test loss values in the SCAND study (Fig. 3). For our curated human study, we do not observe a significant advantage of multimodal learning in comparison to point cloud only. We posit that it is because our curated social scenarios do not contain sufficiently rich semantic social cues to showcase the necessity of using RGB images.

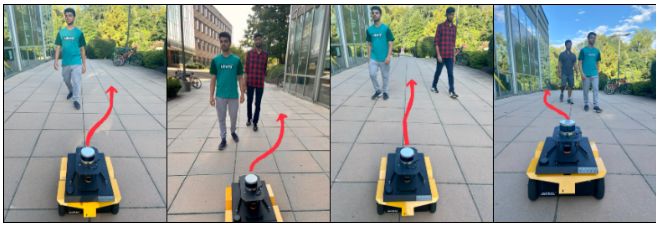

Fig. 6: Human Study with Different Social Scenarios.

VI. CONCLUSIONS

We present a study on learning social robot navigation with multimodal (and unimodal) perception conducted on both a large-scale real-world social robot navigation dataset and in a human study with a physical robot, in comparison to a set of classical approaches. Our study results indicate that multimodal learning has clear advantage over either unimodal counterpart by a large margin in both the dataset and human studies, especially in difficult situations with increasing human density. In terms of unimodal learning, point cloud input is superior compared to RGB input, but it can be improved by utilizing the extra semantic information provided by the camera. Despite the found superiority of multimodal learning, the current study only remains in pre-recorded dataset and curated social scenarios. How multimodal learning will perform in real-world, large-scale, long-term social robot navigation tasks remains unclear and may require extra research and engineering effort.

REFERENCES

[1] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.

[2] X. Xiao, Z. Xu, Z. Wang, Y. Song, G. Warnell, P. Stone, T. Zhang, S. Ravi, G. Wang, H. Karnan *et al.*, "Autonomous ground navigation in highly constrained spaces: Lessons learned from the benchmark autonomous robot navigation challenge at icra 2022 [competitions]," *IEEE Robotics & Automation Magazine*, vol. 29, no. 4, pp. 148–156, 2022.

[3] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[4] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[2] * denotes negatively formulated questions, for which we reverse-code the ratings to make them comparable to the positively formulated ones.

[5] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research: The 14th International Symposium ISRR*. Springer, 2011, pp. 3–19.

[6] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod, "Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior," *Frontiers in psychology*, vol. 4, p. 859, 2013.

[7] J. Hart, R. Mirsky, X. Xiao, S. Tejeda, B. Mahajan, J. Goo, K. Baldauf, S. Owen, and P. Stone, "Using human-inspired signals to disambiguate navigational intentions," in *International Conference on Social Robotics*. Springer, 2020, pp. 320–331.

[8] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, 2022.

[9] S. Quinlan and O. Khatib, "Elastic bands: Connecting path planning and control," in *[1993] Proceedings IEEE International Conference on Robotics and Automation*. IEEE, 1993, pp. 802–807.

[10] X. Xiao, Z. Xu, G. Warnell, P. Stone, F. G. Guinjoan, R. T. Rodrigues, H. Bruyninckx, H. Mandala, G. Christmann, J. L. Blanco-Claraco *et al.*, "Autonomous ground navigation in highly constrained spaces: Lessons learned from the 2nd barn challenge at icra 2023," *arXiv preprint arXiv:2308.03205*, 2023.

[11] J. Buhmann, W. Burgard, A. B. Cremers, D. Fox, T. Hofmann, F. E. Schneider, J. Strikos, and S. Thrun, "The mobile robot rhino," *Ai Magazine*, vol. 16, no. 2, pp. 31–31, 1995.

[12] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Haehnel, C. Rosenberg, N. Roy *et al.*, "Probabilistic algorithms and the interactive museum tour-guide robot minerva," *The International Journal of Robotics Research*, vol. 19, no. 11, pp. 972–999, 2000.

[13] A. Nair, F. Jiang, K. Hou, Z. Xu, S. Li, X. Xiao, and P. Stone, "Dynabarn: Benchmarking metric ground navigation in dynamic environments," in *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2022, pp. 347–352.

[14] P. Xu, J.-B. Hayet, and I. Karamouzas, "Socialvae: Human trajectory prediction using timewise latents," in *European Conference on Computer Vision*. Springer, 2022, pp. 511–528.

[15] R. A. Knepper and D. Rus, "Pedestrian-inspired sampling-based multi-robot collision avoidance," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2012, pp. 94–100.

[16] P. Xu and I. Karamouzas, "Human-inspired multi-agent navigation using knowledge distillation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8105–8112.

[17] E. T. Hall, *The hidden dimension*. Garden City, NY: Doubleday, 1966, vol. 609.

[18] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[19] M. Shiomi, F. Zanlungo, K. Hayashi, and T. Kanda, "Towards a socially acceptable collision avoidance for a mobile robot navigating among pedestrians using a pedestrian model," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 443–455, 2014.

[20] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, no. 5, pp. 569–597, 2022.

[21] X. Xiao, J. Biswas, and P. Stone, "Learning inverse kinodynamics for accurate high-speed off-road navigation on unstructured terrain," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6054–6060, 2021.

[22] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 3931–3936.

[23] J. Liang, U. Patel, A. J. Sathyamoorthy, and D. Manocha, "Crowd-steer: Realtime smooth and collision-free robot navigation in densely crowded scenarios trained using high-fidelity simulation," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 4221–4228.

[24] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Toward agile maneuvers in highly constrained spaces: Learning from hallucination," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1503–1510, 2021.

[25] L. Tai, J. Zhang, M. Liu, and W. Burgard, "Socially compliant navigation through raw depth inputs with generative adversarial imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1111–1117.

[26] Z. Xu, G. Dhamankar, A. Nair, X. Xiao, G. Warnell, B. Liu, Z. Wang, and P. Stone, "Applr: Adaptive planner parameter learning from reinforcement," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 6086–6092.

[27] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023.

[28] M. H. Nazeri and M. Bohlouli, "Exploring reflective limitation of behavior cloning in autonomous vehicles," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 1252–1257.

[29] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[30] K. Weerakoon, A. J. Sathyamoorthy, J. Liang, T. Guan, U. Patel, and D. Manocha, "Graspe: Graph based multimodal fusion for robot navigation in unstructured outdoor environments," 2023.

[31] A. Nguyen, N. Nguyen, K. Tran, E. Tjiputra, and Q. D. Tran, "Autonomous navigation in complex environments with deep multimodal fusion network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5824–5830.

[32] K. Li, M. Shan, K. Narula, S. Worrall, and E. Nebot, "Socially aware crowd navigation with multimodal pedestrian trajectory prediction for autonomous vehicles," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[33] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 922–928.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[36] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[37] S. Pirk, E. Lee, X. Xiao, L. Takayama, A. Francis, and A. Toshev, "A protocol for validating social navigation policies," *arXiv preprint arXiv:2204.05443*, 2022.