VLM-Social-Nav: Socially Aware Robot Navigation through Scoring using Vision-Language Models*

Daeun Song[1], Jing Liang[2], Amirreza Payandeh[1], Amir Hossain Raj[1], Xuesu Xiao[1], and Dinesh Manocha[2]

***Abstract*— We propose VLM-Social-Nav, a novel Vision-Language Model (VLM) based navigation approach to compute a robot's motion in human-centered environments. Our goal is to make real-time decisions on robot actions that are socially compliant with human expectations. We utilize a perception model to detect important social entities and prompt a VLM to generate guidance for socially compliant robot behavior. VLM-Social-Nav uses a VLM-based scoring module that computes a cost term that ensures socially appropriate and effective robot actions generated by the underlying planner. Our overall approach reduces reliance on large training datasets and enhances adaptability in decision-making. In practice, it results in improved socially compliant navigation in human-shared environments. We demonstrate and evaluate our system in four different real-world social navigation scenarios with a Turtlebot robot. We observe at least 27.38% improvement in the average success rate and 19.05% improvement in the average collision rate in the four social navigation scenarios. Our user study score shows that VLM-Social-Nav generates the most socially compliant navigation behavior.**

I. Introduction

Mobile robots integrated into diverse indoor and outdoor human-centric environments are becoming increasingly prevalent. These robots serve various functions, ranging from package and food delivery [2] to service [3] and home assistance [4]. Overall, these roles necessitate interaction with humans and navigating seamlessly through public spaces with pedestrians. In such dynamic scenarios, it is important for the robots to engage in socially compliant interactions and navigation [5], [6]. This paper focuses on the challenges of social navigation [6].

Humans have various behaviors and the environmental or task contexts cannot be easily categorized [6]. A common strategy to handle the challenge is by learning-based approaches to learn the complicated contexts empirically. Imitation Learning (IL) is a recent emerging paradigm for desired navigation behavior [7], [8]. This approach enables autonomous robots to navigate socially by learning from human demonstrations. Other learning approaches, such as reinforcement learning have also been used to address this problem [9]. While both methods demonstrate promising results in real-world settings, substantial datasets [10]–[12] for training and reward engineering are required for their successful application and it is hard to generalize.

Recent Large Language Models (LLMs) and Vision-Language Models (VLMs) demonstrate a deep understanding of contextual information and have the potential to perform chain-of-thought [13] and common sense reasoning [14].

Fig. 1. The trajectories of VLM-Social-Nav (red), DWA (blue), and BC (yellow) approaches in the frontal encountering scenario (left) and the intersection scenario (right). The resulting trajectories show that VLM-Social-Nav demonstrates more socially compliant behavior because it is instructed by a prompt.

Those processes are inherent to social navigation, especially the challenges of contextual appropriateness and politeness, which require understanding the task/environmental context and the behavior of humans. This capability has also been evaluated across diverse domains of robotics, including human-like driving [15] and autonomous robot navigation [16]. However, using language models for social navigation is not well explored, the language models suffer from high latency for real-time navigation, and the issue impedes the smoothness and efficiency of human-robot interaction.

Main Results: In this paper, we present VLM-Social-Nav, a new approach that uses VLMs to interpret contextual information from robot observation to help autonomous robots improve their navigation abilities in human-centered environments. We leverage a VLM to analyze and reason about the current social interaction and generate an immediate *preferred robot action* to guide an underlying motion planner. We formalize the concept of *social cost* and the problem definition of social robot navigation suitable for language descriptions. Our VLM-based scoring module computes the social cost, which is used for a bottom-level motion planner to output appropriate robot actions. To overcome the limitation of existing VLMs' latency issue, we utilize a state-of-the-art perception model (*i.e.,* YOLO [17]) to detect key entities that are used for social interactions (*e.g.,* humans, gestures, and doors) and query a VLM to generate socially compliant navigation behavior and compute the social cost.

*Extended abstract of the original work published in RA-L [1].

[1]George Mason University, [2]University of Maryland, College Park.

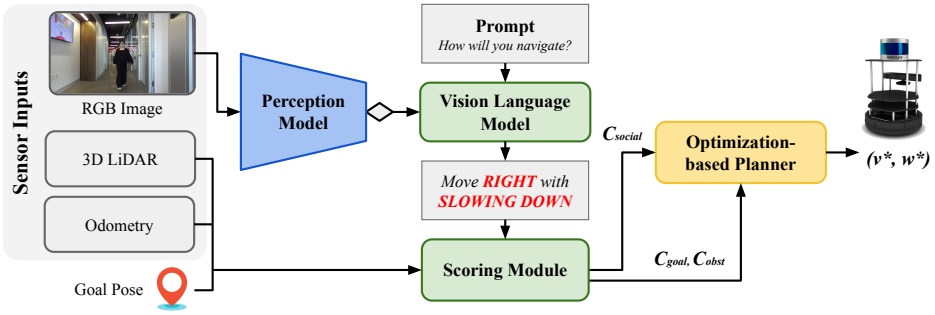

Fig. 2. The overall system architecture of VLM-Social-Nav. Our real-world perception model detects important social entities (*e.g.,* humans, gestures, and doors) in real time and prompts the VLM-based scoring module to compute social cost $\mathcal{C}_{\text{social}}$, which is used to generate socially compliant robot action.

II. Approach

A. Problem Definition

Navigation is the task of generating and following an efficient collision-free path from an initial location to a goal [6]. For social robot navigation, humans are no longer perceived only as dynamic obstacles but also as social entities [5]. It necessitates integrating social norms into robot behaviors. We define the social robot navigation problem as a *Markov Decision Process (MDP)*: $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{C} \rangle$, where $\mathbf{s} = (x, y, \theta) \in \mathcal{S}$ is a state consisting of a robot pose, $\mathbf{a} = (v, w) \in \mathcal{A}$ is an action consisting of a linear and an angular velocity of the robot, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function characterizing the dynamics of the robot, and $\mathcal{C} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a cost function. Given a cost function $\mathcal{C}$, the motion planner finds (v^*, w^*) that minimizes the expected cost. The cost function takes the following form:

$$\mathcal{C}(\mathbf{s}, \mathbf{a}) = \alpha \cdot \mathcal{C}_{\text{goal}} + \beta \cdot \mathcal{C}_{\text{obst}} + \gamma \cdot \mathcal{C}_{\text{social}}, \tag{1}$$

where $\mathcal{C}_{\text{goal}}$ encourages movement toward the goal, $\mathcal{C}_{\text{obst}}$ discourages collisions with obstacles, and $\mathcal{C}_{\text{social}}$ encourages the robot to follow the social norms. α, β, and γ are non-negative weights for each cost term.

The social cost term $\mathcal{C}_{\text{social}}$ encompasses various factors that govern human-robot interactions in shared environments. Defining them mathematically poses challenges. For VLM-Social-Nav, we define $\mathcal{C}_{\text{social}}$ as:

$$\mathcal{C}_{\text{social}} = \|\mathcal{B} - \mathcal{B}_h\|, \tag{2}$$

where $\mathcal{B}$ is a navigation behavior and $\mathcal{B}_h$ is a navigation behavior humans would adopt in accordance with social conventions. Minimizing the deviation between them will encourage the robot to emulate socially acceptable human behaviors. While $\mathcal{B}_h$ can be obtained through various methods, including large datasets [10]–[12], we leverage the power of a VLM to compute appropriate behavior based on its rich contextual understanding and nuanced interpretations from perceived images and given prompts. We elaborate further in Section II-C.

B. VLM-based Social Navigation Architecture

Fig. 2 highlights the overview of VLM-Social-Nav. Our approach integrates a perception layer with an optimization-based motion planner. The motion planner processes sensor inputs and generates a robot action that minimizes the cost function C.

While LiDAR detects geometric information useful for obstacle avoidance, RGB images provide contextual details of the current environment. They contain rich information crucial for social navigation. To enhance navigation capabilities within social contexts, we propose a VLM-based scoring module. VLMs excel in contextual understanding, interpreting scenes not solely based on visual features but also considering social dynamics [18]. VLMs generate socially appropriate robot actions based on current observations and input instructions. Our VLM-based scoring module then calculates a cost term to be used by the motion planner.

While VLMs can generate navigation behaviors that comply with social norms, continuously querying large VLMs for new responses is prohibitively computationally expensive for real-time navigation. To address this challenge, we incorporate a real-time perception model. This model identifies social entities such as humans, gestures, and doors as the robot navigates its environment. Our VLM-based scoring module activates only when significant social cues are detected, ensuring that the social cost term is integrated only when necessary, *i.e.,* when there is any human interaction involved. This approach reduces the VLM queries and facilitates real-time navigation efficiency for our approach.

C. VLM-based Scoring Module

VLM plays a crucial role in VLM-Social-Nav in inferring immediate socially compatible navigation behavior $\mathcal{B}_h^{t+1}$ based on its pre-trained large internet-scale dataset:

$$\mathcal{B}_h^{t+1} = \text{VLM}(\mathcal{I}^t, \mathcal{P}, \mathbf{a}^t), \tag{3}$$

where $\mathcal{I}^t$ is an RGB image from the robot view at time t, $\mathcal{P}$ is a textual prompt, and $\mathbf{a}^t$ is a current robot action at time t. Inspired by In-Context Learning (ICL), our prompt $\mathcal{P}$ is designed to leverage the VLM's reasoning abilities through zero-shot examples. This approach offers an interpretable interface, mirroring human reasoning and decision-making processes, without extensive training [19].

Our VLM-based scoring module starts from the insight that the action space of a mobile robot can be readily mapped to linguistic terms. For example, the action "move forward at a constant speed" can be linked to a linear velocity of v^t m/s and an angular velocity of 0. The heading direction on the

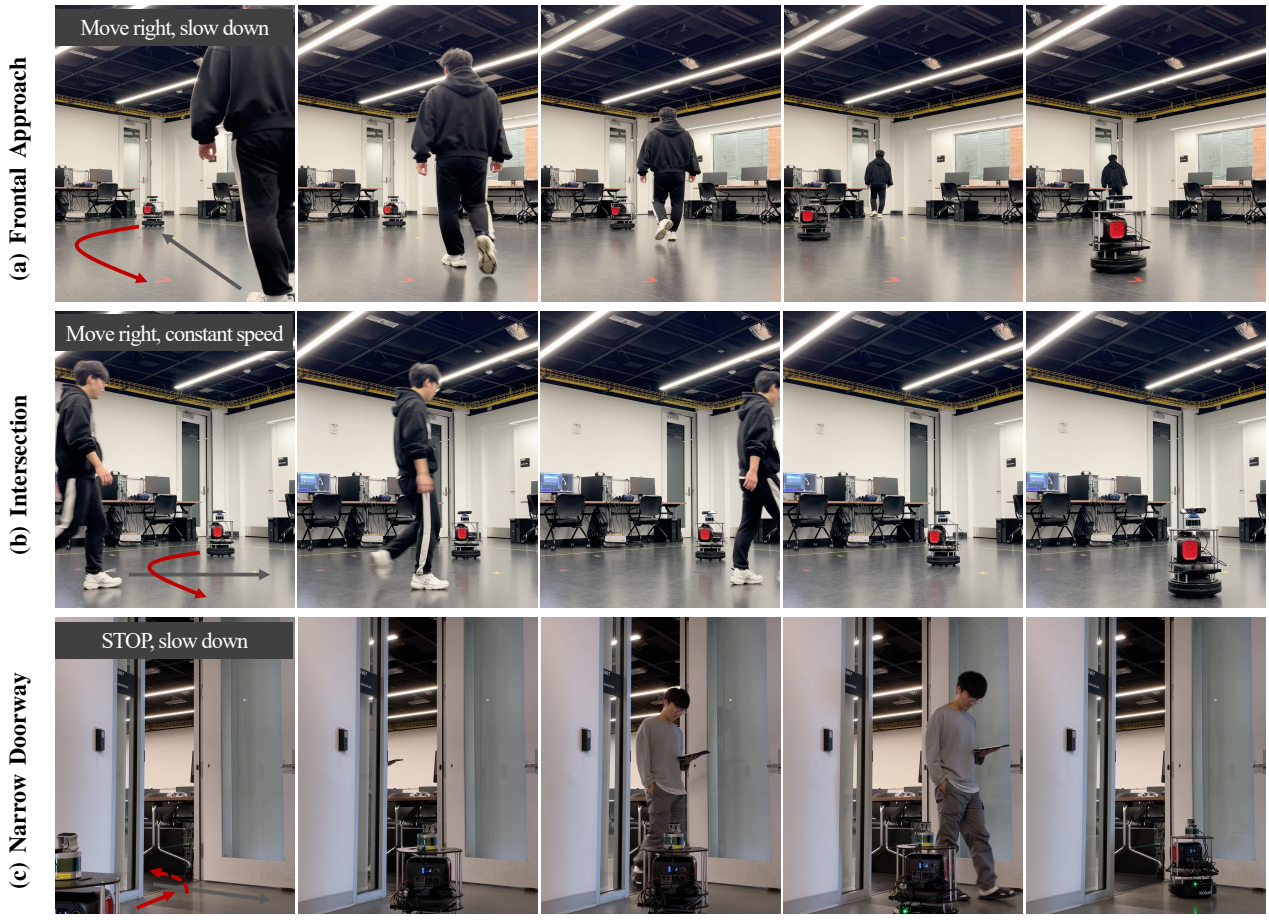

Fig. 3. Qualitative Results: the robot navigation behaviors with VLM-Social-Nav for four social navigation scenarios: (a) Frontal Approach, (b) Intersection, and (v) Narrow Doorway. The solid gray arrow shows the participant's path. The solid red arrow shows the robot's path. The red dashed arrow shows the robot's path after a stop motion. A caption on the top left shows the result from the VLM.

left indicates a positive value of w^t, while the direction on the right indicates a negative value. Leveraging this understanding, we structure the output of the VLM into a linguistic format comprising the heading and the speed. Subsequently, our scoring module extracts $\mathcal{B}_h^{t+1} \mapsto (v_h^{t+1}, w_h^{t+1}) \in \mathcal{A}$ from these tokens; $v_h^{t+1} = v^t + \delta_s$, where δ_s is derived from the response for the speed; $w_h^{t+1} = \delta_d$, where δ_d is derived from the response for the heading. Thus, the social cost term for the next time step can be calculated:

$$\mathcal{C}_{\text{social}}^{t+1} = w_l \cdot \|v - v_h^{t+1}\| + w_a \cdot \|w - w_h^{t+1}\|, \qquad (4)$$

where w_l and w_a are non-negative weights. Given all the cost terms, our low-level optimization-based motion planner finds the robot action (v^*, w^*) that minimizes the cost.

We provide a high-level task description along with an image $\mathcal{I}^t$ captured from the robot's perspective. Furthermore, the current robot action $\mathbf{a}^t = (v^t, w^t) \in \mathcal{A}$ is provided. The angular velocity is mapped into corresponding directional instructions based on predefined categories (*i.e.,* positive values correspond to *left*, values near zero to *straight*, and negative values to *right*). Supplementary instructions regarding walking etiquette are included. Although the VLM demonstrates proficient navigation abilities in the absence of explicit instructions, offering reasoning guidelines enhances its decision-making processes [19].

III. Experiments

A. Implementation Details

VLM-Social-Nav is tested on a Turtlebot 2 equipped with a Velodyne VLP16 LiDAR, a Zed 2i camera, and a laptop with an Intel i7 CPU and an Nvidia GeForce RTX 2080 GPU. We use YOLO [17] as our real-world perception model to detect key objects. Generative Pre-trained Transformer 4 with Vision (GPT-4V) [14] is used as our VLM to comprehend the social dynamics and output the immediate preferred robot action. We combined our approach with a low-level motion planner DWA [20]. We compare VLM-Social-Nav with DWA without social cost $\mathcal{C}_{\text{social}}$ and BC [21] trained on a state-of-the-art, large-scale social navigation dataset, SCAND [11].

Evaluating the social aspects of social robot navigation is inherently challenging [22]. To validate VLM-Social-Nav, we carefully follow the social robot navigation studies [23], [24], which set up the benchmark scenarios and the metrics for measuring social compliance. We present qualitative, quantitative, and user study results in four different social navigation scenarios:

- **Frontal Approach:** A robot and a human approach each other from two ends of a straight trajectory.
- **Frontal Approach with Gesture:** A robot and a human approach each other from two ends of a straight trajec-

TABLE I

QUANTITATIVE RESULTS: PERFORMANCE COMPARISONS USING BC [21], DWA [20], AND VLM-SOCIAL-NAV

Metric	Method	Scenario			
		(a) Frontal Approach	(b) Frontal Approach w/ Gesture	(c) Intersection	(d) Narrow Doorway
Success Rate (%) ↑	BC	38.10	0	33.33	42.86
	DWA	**100**	0	90.48	**100**
	VLM-Social-Nav	**100**	**100**	**100**	**100**
Collision Rate (%) ↓	BC	42.86	66.67	28.57	38.10
	DWA	28.57	19.05	19.05	38.10
	VLM-Social-Nav	**14.29**	**0**	**4.76**	**9.52**
User Study Score ↑	BC	2.80 ± 1.45	2.23 ± 1.54	2.80 ± 1.40	2.60 ± 1.33
	DWA	3.99 ± 0.80	3.38 ± 0.64	3.57 ± 0.62	3.59 ± 0.83
	VLM-Social-Nav	**4.31 ± 0.72**	**4.28 ± 0.56**	**4.35 ± 0.70**	**4.04 ± 0.74**

tory. The human recognizes the robot and then gestures for it to stop.

- **Intersection:** A robot and a human cross each other on perpendicular trajectories.
- **Narrow Doorway:** A robot and a human cross each other's paths by moving through a narrow doorway.

B. Qualitative Result

Fig. 3 shows snapshots of the resulting robot motion using VLM-Social-Nav in three selected scenarios. We demonstrate that VLM-Social-Nav follows the social convention and navigates toward its goal as expected. Fig. 1 illustrates the resulting trajectories of VLM-Social-Nav in comparison to those of DWA and BC methods. A notable observation is that, while DWA also effectively avoids collisions with individuals, VLM-Social-Nav generates trajectories that align more closely with social norms. For instance, in the frontal approach scenario, while DWA tends to maneuver around the person either to the right or left, VLM-Social-Nav predominantly bypasses the person on the right side. Similarly, in the intersection scenario, whereas DWA occasionally obstructs the person's path by veering to avoid collision directly in front, VLM-Social-Nav adjusts its trajectory to pass behind the individual, adapting effectively to the human's movement direction. Additionally, BC avoids humans but fails to recover and follow the original path. This leads to many failures in reaching the goal.

C. Quantitative Result

To further validate VLM-Social-Nav, we evaluate the methods using three different metrics. The success rate describes whether the robot reaches the goal. The collision rate describes whether the robot collided with the human or other objects in the environment. We also mark it as in collision when we manually intervene to avoid an imminent collision with the human subject or surroundings. The user study score is an average score we obtained from the user study detailed in Section III-D.

Table I reports the results averaged over 21 runs for each method and scenario. The results demonstrate that VLM-Social-Nav, DWA with social cost, outperforms other methods in every metric. DWA excels at following a path smoothly, yet it faces challenges in collision avoidance as it relies solely on the LiDAR sensor and does not consider social compliance. Most of the collisions occurred when DWA navigated in a way that interfered with a person's path, for example, going in front of the person when intersecting. We also observe that the outcomes of BC varied. At times, when attempting to avoid collisions, it failed to return to its original path and failed to reach the goal. VLM-Social-Nav improves the average success rate by 27.38% and reduces the average collision rate by 19.05% across four social navigation scenarios.

D. User Study

To validate the social compliance of VLM-Social-Nav, we conduct a user study. We ask the participants to walk along the predefined trajectory and then to answer questionnaires about the robot motion [24].The three methods are randomly shuffled and repeated three times. Each scenario is tested on seven participants. We use a five-level Likert scale to ask participants to rate their agreement with these statements.

The user study scores in Table I show the study result. Based on the results, it's evident that VLM-Social-Nav receives the highest level of agreement from participants across all questions, indicating its strong adherence to social norms. The standard error of the BC method was large, indicating that the performance of the BC method was not consistent.

IV. CONCLUSION

We propose a novel social navigation approach based on VLMs, focusing on real-time, socially compliant decision-making in human-centric environments. We utilize the perception model to detect important social entities and prompt a VLM to generate guidance for socially compliant behavior. VLM-Social-Nav features a VLM-based scoring that ensures socially appropriate and effective robot actions. This minimizes the dependence on extensive training datasets and eliminates the necessity for explicit rules or hand-tuned parameters typically associated with imitation learning approaches. By furnishing textual instructions to VLM, we can instruct the robot to adhere to specific navigation rules, such as navigating on the right or left according to cultural norms.

REFERENCES

[1] 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. 508–515, 2025.

[2] S. Technology. (2024) Starship. [Online]. Available: https://www.starship.xyz/

[3] D. Robotics. (2024) Dilligent robotics. [Online]. Available: https://www.diligentrobots.com/

[4] Amazon. (2024) Meet astro, a home robot unlike any other. [Online]. Available: https://www.aboutamazon.com/news/devices/meet-astro-a-home-robot-unlike-any-other

[5] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation—a survey," *ACM Transactions on Human-Robot Interaction*, vol. 13, no. 1, pp. 1–36, 2024.

[6] C. Mavrogiannis *et al.*, "Core challenges of social robot navigation: A survey," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[7] N. Hirose *et al.*, "Sacson: Scalable autonomous control for social navigation," *IEEE Robotics and Automation Letters*, 2023.

[8] A. H. Raj *et al.*, "Rethinking social robot navigation: Leveraging the best of two worlds," in *IEEE International Conference on Robotics and Automation*, 2024.

[9] H. Kretzschmar *et al.*, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[10] A. Rudenko *et al.*, "Thör: Human-robot navigation data collection and accurate motion trajectories dataset," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 676–682, 2020.

[11] H. Karnan *et al.*, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, 2022.

[12] 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 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 7442–7447.

[13] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] L. Wen *et al.*, "On the road with gpt-4v (ision): Explorations of utilizing visual-language model as autonomous driving agent," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[16] D. Shah *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," 2022.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[18] J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, and R. Krishna, "Manipulate-anything: Automating real-world robots using vision-language models," *arXiv preprint arXiv:2406.18915*, 2024.

[19] S. Min *et al.*, "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[20] 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.

[21] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[22] N. Tsoi, J. Romero, and M. Vázquez, "How do robot experts measure the success of social robot navigation?" in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 1063–1066.

[23] A. Francis *et al.*, "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Transactions on Human-Robot Interaction*, 2024.

[24] 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.