Socially Aware Robot Navigation through Scoring Using Vision-Language Models

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

***Abstract*— We propose a novel Vision-Language Model (VLM) based navigation approach to compute a robot's trajectory in human-centered environments. Our goal is to make real-time decisions on robot actions that are socially compliant with human expectations. We utilize the perception model to detect important social entities and prompt a VLM to generate guidance for socially compliant robot behavior. Our approach 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 datasets (for training) 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 36.37% improvement of success rate in four challenging social navigation scenarios comparing with other state-of-the-art methods.**

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 [1], [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]–[7].

Challenges of social navigation: This paper focuses on the challenges of social navigation [7], and encompasses the ability of robots to navigate while adhering to social etiquette, especially, contextual appropriateness, which requires robot to understand the relative importance of the previous objectives and context of the environment, tasks or interpersonal behaviors. However, navigating socially across varying contexts presents distinct challenges [7], [8], including ensuring safety, comfort, politeness and social norms.

Inferring contextually appropriate navigation behaviors is challenging. Humans have various behaviors and the environmental or task contexts are also not easily to be categorized [7]. A common strategy to handle the challenge is by learning-based approaches to empirically learn the complicated contexts. Imitation Learning (IL) is a recent emerging paradigm for desired navigation behavior [9]–[13]. This approach enables autonomous agents to navigate socially by learning from human demonstration. Other learning approaches, such as Reinforcement Learning (RL) have also been used to address this problem [14]. While both methods demonstrate promising results in real-world settings, substantial datasets [15]–[17] for training and reward engineering are required for their successful application.

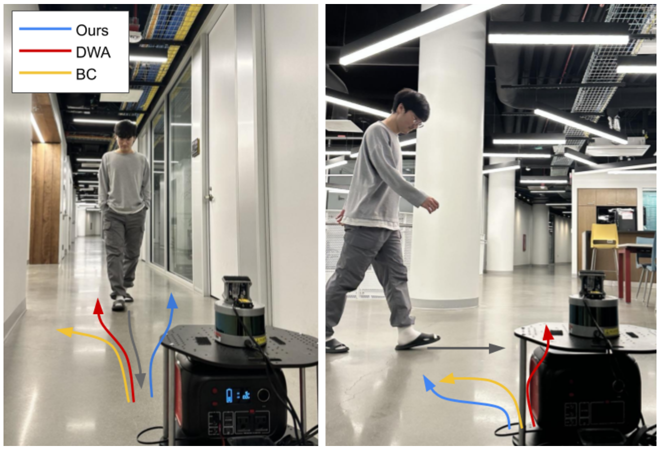


Fig. 1: Our approach (blue) demonstrates more social compliant than DWA (red) and BC (yellow) approaches in the frontal encountering scenario (left) and the intersection scenario (right).

Language models inherently suit to contextual understanding but not well applied in social navigation. Recent advances in 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 [18] and common sense reasoning [19], [20]. It is inherently for 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 scenarios [21], [22] and autonomous robot navigation [23], [24]. However, using language models for social navigation is not well explored and the language models suffer from high latency for real-time navigation, and the issue impedes the smoothness and efficiency of human-robot social interaction.

Main Results: In this paper, we present a new approach that uses VLMs to interpret contextual information from the robot observation to help autonomous agents improve their navigation abilities in human-centered environments. In particular, we leverage a VLM to analyze and reason about the current social interaction, and generate an immediate *preferred robot action* to guide the motion planner. Our VLM-based scoring module computes the *social cost*, which is used

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

for the bottom-level motion planner to output the appropriate robot action. To overcome the limitation of existing VLMs' latency issue, we utilize the state-of-art perception model (*i.e.,* YOLO [25]) to detect key entities that are used for social interactions (*e.g.,* humans, gestures, and doors) and trigger a VLM to compute the social cost. We demonstrate our approach in four different indoor scenarios with human interactions. Unlike previous social navigation approaches, our method can better deal with social navigation scenarios by interpreting the situation based on *common sense*, through observation without any training with a large dataset. Some of our main results include:

- We bridge the gap between VLMs and social navigation and propose a novel approach for social navigation with contextual understanding. By integrating VLMs with optimization-based local motion planners and state-of-the-art perception models, our approach enables robots to detect social entities efficiently and make real-time decisions with respect to socially compliant robot behavior.
- We propose a VLM-based scoring module that translates the current robot observation and the textual instructions into a relevant social cost term. This cost term is used for the bottom-level motion planner to output the sub-optimal robot action.
- We evaluate our approach in four different real-world indoor social navigation scenarios. Our approach achieves at least 36.37% improvement in success rate averagely in the four challenging social navigation scenarios. We measure the success rate, collision rate, and user study score. We also perform a user study and compare the results with a Dynamic-Window approach (DWA) [26] and Behavior Cloning (BC) [27] method trained on a state-of-the-art large Socially CompliAnt Navigation Dataset (SCAND) [16].

II. Related Work

In this section, we give an overview of existing works related to social navigation, including safety requirements of social navigation, different challenges of contextual appropriateness, and Large Foundation Models (LFMs) for robot navigation.

A. Safety Requirement of Social Navigation

For social navigation, keeping safety is the basic requirement for interaction with humans and navigation in dynamic challenging scenarios [26], [28]–[30]. The DWA [26] calculates the collision constraints of the robots' actions and chooses the best feasible action closest to the target as the output. Model Predictive Control approaches [31], [32] are also widely used for collision avoidance tasks, which provide smooth trajectories for the robot to follow; the learning-based MPC [32] enhances the performance of the original MPC method by empirically learning the best actions for navigation. However, this method can be computationally costly. Velocity-Obstacle (VO)-based approaches are more efficient and can be used to simulate the actions of crowds [28], [33], but these approaches don't take into account the uncertainties in the perception output. PRVO [34] and OFVO [29] handle the uncertainties of perception in motion planning, but those approaches require a hard threshold to set the confidence for planning. To deal with this issue, learning-based methods empirically train the policies by demonstrations [30], [35], [36], whereas other methods [30], [35] use reinforcement learning to train the robot in a simulator and implement it in real-world scenarios. However, learning-based approaches require lots of data or even on-policy training with realistic simulators to learn the task.

B. Contextual Appropriateness of Social Navigation

Researchers have proposed diverse methodologies to integrate social awareness into mobile robot navigation systems. The development of a comprehensive social navigation system poses inherent complexity, requiring sophisticated perception and reasoning capabilities to navigate environments shared with humans and other robots [6]. Establishing an accurate definition of social navigation is paramount, given its potential variation across cultures and platforms. Except for the safety requirement, assessing social compliance presents an additional challenge, contingent on the scenario and platform, and should adhere to Contextual Appropriateness, which includes Comfort, Legibility, Social Competency, Politeness, Understanding Other Agents, and Proactivity [8], [37]. Various methodologies are employed to address this challenge, with a significant focus on enhancing learning methods through reinforcement learning, learning from demonstration—particularly by analyzing examples of human trajectories or robots operated by humans—and the utilization of simulated datasets [38]–[42]. Additionally, various datasets have been collected for this purpose [15]–[17]. Despite extensive research employing diverse approaches with machine learning techniques [43], less emphasis has been placed on formulating social navigation using VLMs, which inherently analyze the contextual information of the environment [7], [44], [45].

C. Large Foundation Models for Navigation

Recent advancements in Language Foundation Models (LFMs) [46], encompassing Vision-Language Models (VLMs) and Language-Language Models (LLMs), show significant potential in robotic navigation, despite the challenges like bias and reliability [47] associated with the LFMs. SayCan [48] integrate LLMs for high-level task planning. GPT-Driver [49] evaluates the performance of GPT-3.5 in simulation for autonomous driving, framing motion planning as a language modeling problem. LL3MVN [50] constructs semantic maps of environments and utilizes LLMs to reach long-term goals, while LLaDA [51] enables autonomous vehicles to adapt to diverse traffic rules across regions. LM-Nav [52] utilizes GPT-3 and CLIP [53] to navigate outdoor environments based on natural language instructions, combining language and visual cues for optimal path planning. Xu et al. [54] employ VLMs to generate visual-language maps for navigation tasks. RT-1 [55] and RT-2 [56] models

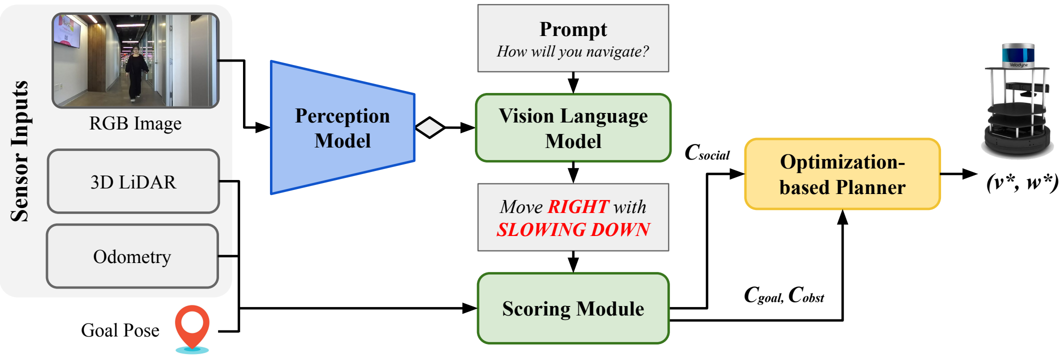


Fig. 2: The overall system architecture of our social navigation algorithm. Our real-world perception (vision) model detects important social entities (*e.g.,* humans, gestures, and doors) in real time and prompts the VLM-based scoring module to compute social cost C_{social}, which is used to generate socially compliant robot action.

leverage multimodal robotics data for navigating complex environments and offer greater generalizability through larger dataset training [52], [54]–[56]. However, for social navigation, language models are not well explored, even though VLMs and LLMs inherently handle the understanding of contextual information. We bridge the gap and propose a novel approach for social navigation by taking advantage of the contextual understanding capability of VLMs.

III. Approach

In this section, we define the social navigation problem and describe our approach in detail.

A. Problem Definition

Navigation is the task of following an efficient collision-free path from an initial location to a goal [7]. In general, the overall system consists of a global planner and a local planner. A global planner is designed to find a collision-free path to reach a goal, while a local planner aims to navigate the robot through its immediate surroundings, making real-time adjustments to deal with dynamic obstacles.

For social robot navigation, humans are no longer perceived only as dynamic obstacles but also as social entities [6], [57]. 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 space consisting of a robot pose, $\mathbf{a} = (v, w) \in \mathcal{A}$ is an action space consisting of a linear and an angular velocity of a robot, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition function characterizing the dynamics of the robot, and $\mathcal{C} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ 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:

$$C(\mathbf{s}, \mathbf{a}) = \alpha \cdot C_{goal} + \beta \cdot C_{obst} + \gamma \cdot C_{social}, \qquad (1)$$

where C_{goal} encourages movement toward the goal, C_{obst} discourages collisions with obstacles, and C_{social} encourages the robot to follow the social norms. α, β, and γ are a non-negative weight for each cost term.

The social cost term C_{social} encompasses various factors that govern human-robot interactions in shared environments. Defining them mathematically poses challenges. For our approach, we define C_{social} as:

$$C_{social} = \|\mathcal{B} - \mathcal{B}_h\|, \qquad (2)$$

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 behavior. While $\mathcal{B}_h$ can be obtained through various methods, including large datasets [15]–[17], we leverage the power of VLM to compute appropriate behavior based on a rich contextual understanding and nuanced interpretations from the image and the given prompts. We elaborate further in Section III-C.

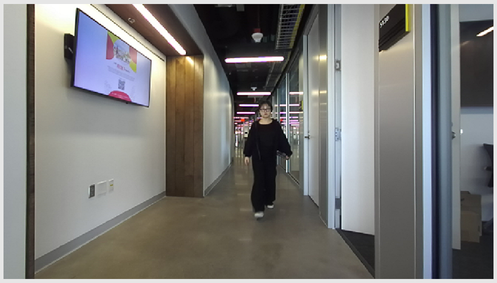
B. VLM-based Social Navigation Architecture

Figure 2 highlights the overview of our approach. Our formulation is built upon an autonomous navigation systen comprising a global path planner and a local motion planner. The global path planner generates a high-level trajectory for the robot to follow, considering obstacles and the goal location. The local planner refines this trajectory by considering the sensor inputs and outputs a robot action that minimizes the cost function C.

To improve navigation abilities in a social-interaction context, we introduce a VLM-based scoring module. As the robot navigates through the environment, our real-world perception module detects social entities, such as humans, gestures, and doors, in real time. When important social cues are detected, our VLM-based scoring module is prompted and outputs the immediate robot action that is socially acceptable based on the current robot observation. It is calculated as a cost term to be used by the motion planner. Our approach is designed to make real-time decisions on robot actions based on the current observations. The social cost term is integrated only when it is necessary, *i.e.,* when there is any human interaction involved.

C. VLM-based Scoring Module

VLM plays a crucial role in our approach to infer immediate socially compatible navigation behavior $\mathcal{B}_h^{t+1}$ based on



Task:
The image depicts your current view while you navigating towards the goal. How will you navigate concerning the person in your view? You will need to follow general walking etiquette.

Ego state:
- heading direction: straight
- linear velocity: 0.28

Remember:
- Move to the right when passing by a person.
- Pass on the left when overtaking a person.
- Do not obstruct others' paths.
- ...

Answer Format:
Move DIRECTION with SPEED
- options for DIRECTION: left, straight, right
- options for SPEED: slow down, speed up, constant, stop

Fig. 3: An example prompt used in our approach. Parameterized inputs are highlighted in blue. Formatted outputs are highlighted in red. The example output of an example image is *Move right with slow down*, to slowly pass by a person on the right side.

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 of the robot view at time t, $\mathcal{P}$ is a textual prompt, $\mathbf{a}^t$ 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 [58].

Figure 3 shows an example prompt $\mathcal{P}$ used in our experiment. 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. Supplementary instructions regarding walking etiquette are also included. Although the VLM demonstrates proficient navigation abilities even in the absence of explicit instructions, offering reasoning guidelines enhances its decision-making processes [58]. These guidelines not only facilitate comprehensive reasoning and judgment within the VLM but also enable the robot to adapt to specific rules more effectively. We format the output of VLM as an immediate action that the robot should take for the next step, specifying the heading direction and the speed.

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, where v^t represents the current linear velocity. The heading direction on the right indicates a positive value, while the direction on the left indicates a negative. Leveraging this understanding, we structure the output of the VLM into a linguistic format comprising the heading, and the speed. Subsequently, our VLM-based 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 SPEED; $w_h^{t+1} = \delta_d$, where δ_d is derived from the response for DIRECTION. Thus, the social cost term for the next time step can be calculated:

$$C_{social}^{t+1} = w_l \cdot \|v - v_h^{t+1}\| + w_a \cdot \|w - w_h^{t+1}\|, \tag{4}$$

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

IV. Experiments

In this section, we detail the implementation of our method and describe our qualitative and quantitative results including comparisons with other methods and the user study.

A. *Implementation Details*

Our approach is tested on a Turtlebot 2 equipped with a Velodyne VLP16 lidar, a MMlove fisheye lens RGB camera, and a laptop with intel i7 CPU, and an Nvidia GeForce RTX 2080 GPU. We use You Only Look Once (YOLO) [25] as our real-world perception model to detect key objects (*e.g.,* humans, door, and gesture) and Generative Pre-trained Transformer 4 with Vision (GPT-4V) [20] as our VLM to comprehend the social dynamics and output the immediate preferred robot action. We combined our method with a low-level motion planner DWA [26]. We compare our method with DWA without social cost C_{social} and BC [27] method trained on a state-of-the-art large set of Socially CompliAnt Navigation Dataset (SCAND) [16].

To validate our approach, we carefully follow the social robot navigation studies [8], [59] that have set up the benchmark scenarios and the metrics for measuring social navigation. We present qualitative, quantitative, and user study results in four different indoor 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 trajectory. The human recognizes the robot and then gestures for it to stop.

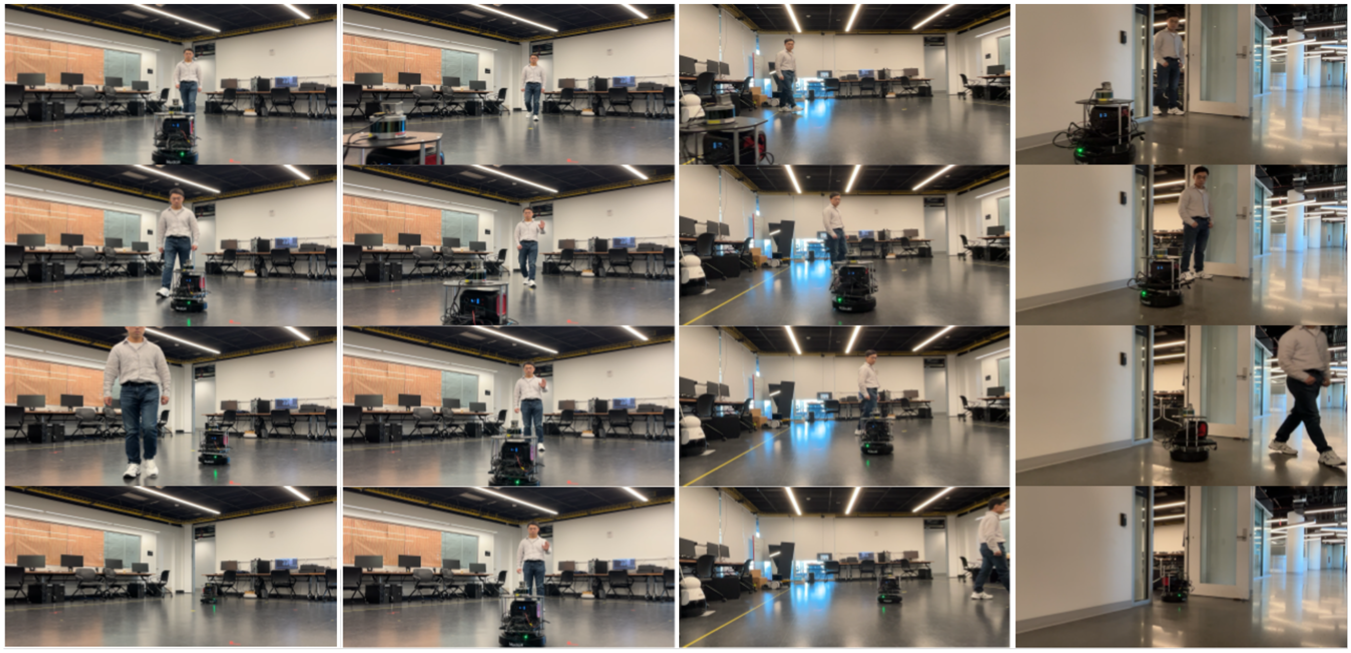


Fig. 4: Results of the robot motion with our approach. The scenario starts from left to right: Frontal Approach, Intersection, Intersection with Gesture, and Narrow Doorway.

TABLE I: Quantitative Reults: we achieve at least 36.37% improvement in success rate averagely in the four challenging social navigation scenarios.

Metric	**Method**	**Scenario**			
		Frontal Approach	Frontal Approach w/ Gesture	Intersection	Narrow Doorway
Success Rate (%)	BC	26.66	0	13.33	33.33
	DWA	**100**	**0**	93.33	**100**
	Ours	**100**	**100**	**100**	**100**
Collision Rate (%)	BC	46.66	60.00	33.33	40.00
	DWA	26.66	20.00	13.33	33.33
	Ours	**6.66**	**0**	**0**	**6.66**
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
	Ours	**4.31** ± **0.72**	**4.28** ± **0.56**	**4.35** ± **0.70**	**4.04** ± **0.74**

- **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

Based on the protocols and principles set by other studies [8], [59], the robot is expected to behave in a socially compliant way as follows:

- **Frontal Approach:** The robot is expected to yield or slow down and modify its original trajectory so as not to obstruct the human path. Similar to driving rules, it is conventional to keep on the right side.
- **Frontal Approach with Gesture:** The robot is expected to yield by interpreting the human gesture.
- **Intersection:** The robot is expected to drive slowly when it approaches the human. It may come to a complete stop or modify its original trajectory to go behind the human to not obstruct the path.
- **Narrow Doorway:** The robot is expected to wait outside the door and yield to the human.

Fig. 4 shows snapshots of the resulting robot motion using our approach in four different scenarios. We demonstrate that our approach follows the social convention and navigates toward its goal as expected. Fig. 1 illustrates the resulting trajectories of our approach in comparison to those of DWA and BC methods. A notable observation is that while DWA also effectively avoids collisions with individuals, our approach 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, our approach predominantly bypasses them 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, our approach endeavors to pass behind the individual. Additionally, BC successfully avoids humans but fails to recover and follow the original path.

TABLE II: Social Compliance Questionnaire

Scenario 1: Frontal Approach	
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.*
5	It was clear what the robot wanted to do.
Scenario 2: Frontal Approach with Gesture	
6	The robot maintained a safe and comfortable distance at all times.
7	The robot slowed down and stopped.
8	The robot followed my command
9	I felt the robot paid attention to what I was doing.
Scenario 3: Intersection	
10	The robot let me cross the intersection by maintaining a safe and comfortable distance.
11	The robot changed course to let me pass.
12	The robot paid attention to what I was doing.
13	The robot slowed down and stopped to let me pass.
Scenario 4: Narrow Doorway	
14	The robot got in my way.*
15	The robot moved to avoid me.
16	The robot made room for me to enter or exit.
17	It was clear what the robot wanted to do.

C. *Quantitative Result*

To further validate our approach, we evaluate the methods using three different metrics [8]. The success rate describes whether the robot reaches the goal. For the frontal approach with gesture scenario, we mark it as successful when the robot reacts to the gesture. The collision rate describes whether the robot collided with the human or other objects in the environment. We also mark it as 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 IV-D.

Table I reports the results averaged over 15 runs for each method and scenario. The results demonstrate that our approach, 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. This occurs when the participant walks too quickly and the robot attempts to make an abrupt change in motion. The outcomes of BC varied. At times, when attempting to avoid collisions, it failed to return to its original path. Conversely, there were instances where it didn't attempt collision avoidance at all, resulting in collisions with the participants. For gesture recognition, only our proposed method successfully responded to the participants' gestures. In total, we achieve at least 36.37% improvement in success rate averagely in the four challenging social navigation scenarios.

D. *User Study*

To validate the social compliance of our method, we conduct a user study. We ask the participants to walk along the predefined trajectory and ask them to answer questionnaires about the robot motion [59] (Table II). * denote negatively formulated questions, for which we reverse-code the ratings to ensure comparability with the positively formulated ones. The three methods are randomly shuffled and repeated three times. Each scenario is tested on five different participants. We use a five-level Likert scale to ask participants to rate their agreement toward these statements.

Fig. 5 and the user study score on Table I show the study result. The plot shows the per-question average scores for the three methods in each scenario. Based on the results, it's evident that our method received the highest level of agreement from participants across all questions, indicating its strong adherence to social norms.

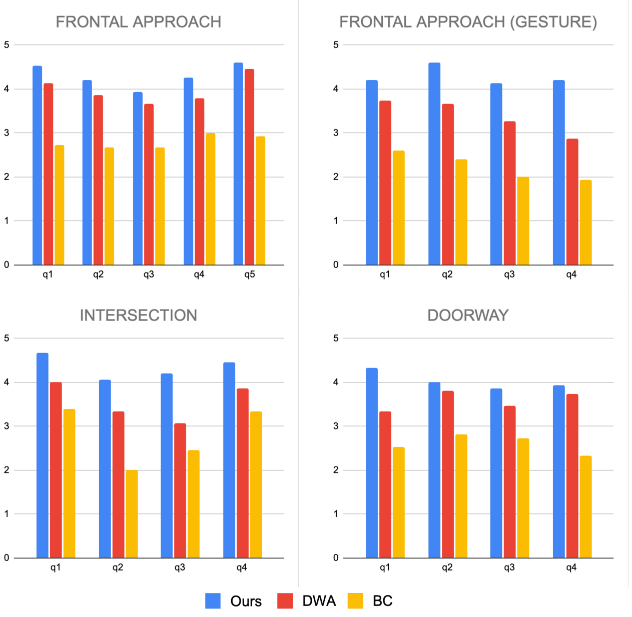


Fig. 5: User Study Average Scores

V. Conclusion

We propose a novel social navigation approach based on Vision-Language Models (VLMs), focusing on real-time, socially compliant desicion-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. Our approach features a VLM-based scoring system that ensures socially appropriate and effective robot actions, minimizing reliance on large training datasets and the need for explicit rules or rewards w.r.t. reinforcement learning-based approaches, thereby enhancing the adaptability and scalability of the decision-making process in autonomous robot navigation. 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.

References

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

[2] Amazon. (2024) Meet scout. [Online]. Available: https://www.aboutamazon.com/news/transportation/meet-scout

[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] 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.

[6] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation–a survey," *ACM Transactions on Human-Robot Interaction*, 2024.

[7] 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.

[8] A. Francis, C. Pérez-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," *arXiv preprint arXiv:2306.16740*, 2023.

[9] N. Hirose, D. Shah, A. Sridhar, and S. Levine, "Sacson: Scalable autonomous control for social navigation," *IEEE Robotics and Automation Letters*, 2023.

[10] X. Xiao, B. Liu, G. Warnell, J. Fink, and P. Stone, "Appld: Adaptive planner parameter learning from demonstration," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4541–4547, 2020.

[11] X. Xiao, T. Zhang, K. M. Choromanski, T.-W. E. Lee, A. Francis, J. Varley, S. Tu, S. Singh, P. Xu, F. Xia, S. M. Persson, L. Takayama, R. Frostig, J. Tan, C. Parada, and V. Sindhwani, "Learning model predictive controllers with real-time attention for real-world navigation," in *Conference on robot learning*. PMLR, 2022.

[12] B. Panigrahi, A. H. Raj, M. Nazeri, and X. Xiao, "A study on learning social robot navigation with multimodal perception," *arXiv preprint arXiv:2309.12568*, 2023.

[13] A. H. Raj, Z. Hu, H. Karnan, R. Chandra, A. Payandeh, L. Mao, P. Stone, J. Biswas, and X. Xiao, "Targeted learning: A hybrid approach to social robot navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[14] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[15] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal, "Thör: Human-robot navigation data collection and accurate motion trajectories dataset," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 676–682, 2020.

[16] 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.

[17] 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 *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7442–7447.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *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.

[19] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, and J. Berant, "Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 346–361, 2021.

[20] OpenAI, "Gpt-4v(ision) system card," 2023.

[21] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. Ma, Y. Li, L. Xu, D. Shang *et al.*, "On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving," *arXiv preprint arXiv:2311.05332*, 2023.

[22] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[23] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 10 608–10 615.

[24] D. Shah, B. Osiński, b. ichter, and S. Levine, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 492–504. [Online]. Available: https://proceedings.mlr.press/v205/shah23b.html

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[26] 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.

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

[28] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Optimal reciprocal collision avoidance for multi-agent navigation," in *Proc. of the IEEE International Conference on Robotics and Automation, Anchorage (AK), USA*, 2010.

[29] J. Liang, Y.-L. Qiao, T. Guan, and D. Manocha, "Of-vo: Efficient navigation among pedestrians using commodity sensors," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6148–6155, 2021.

[30] 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.

[31] S. H. Arul, J. J. Park, and D. Manocha, "Ds-mpepc: Safe and deadlock-avoiding robot navigation in cluttered dynamic scenes," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 2256–2263.

[32] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic mpc for model-based reinforcement learning," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1714–1721.

[33] A. Best, S. Narang, S. Curtis, and D. Manocha, "Densesense: Interactive crowd simulation using density-dependent filters." in *Symposium on Computer Animation*, 2014, pp. 97–102.

[34] B. Gopalakrishnan, A. K. Singh, M. Kaushik, K. M. Krishna, and D. Manocha, "Prvo: Probabilistic reciprocal velocity obstacle for multi robot navigation under uncertainty," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1089–1096.

[35] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 740–759, 2020.

[36] H. Sun, W. Zhang, R. Yu, and Y. Zhang, "Motion planning for mobile robots—focusing on deep reinforcement learning: A systematic review," *IEEE Access*, vol. 9, pp. 69 061–69 081, 2021.

[37] Y. Gao and C.-M. Huang, "Evaluation of socially-aware robot navigation," *Frontiers in Robotics and AI*, vol. 8, p. 721317, 2022.

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

[39] 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*, 2009, pp. 3931–3936.

[40] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Data-driven hri: Learning social behaviors by example from human–human interaction," *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 988–1008, 2016.

[41] M. Li, R. Jiang, S. S. Ge, and T. H. Lee, "Role playing learning for socially concomitant mobile robot navigation," 2017.

[42] M. Nazeri, J. Wang, A. Payandeh, and X. Xiao, "Vanp: Learning where to see for navigation with self-supervised vision-action pre-training," 2024.

[43] A. Payandeh, K. T. Baghaei, P. Fayyazsanavi, S. B. Ramezani, Z. Chen, and S. Rahimi, "Deep representation learning: Fundamentals,

technologies, applications, and open challenges," *IEEE Access*, vol. 11, pp. 137 621–137 659, 2023.

[44] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1726–1743, 2013.

[45] K. Charalampous, I. Kostavelis, and A. Gasteratos, "Recent trends in social aware robot navigation: A survey," *Robotics and Autonomous Systems*, vol. 93, pp. 85–104, 2017.

[46] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2022.

[47] A. Payandeh, D. Pluth, J. Hosier, X. Xiao, and V. K. Gurbani, "How susceptible are llms to logical fallacies?" 2023.

[48] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," 2022.

[49] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," 2023.

[50] B. Yu, H. Kasaei, and M. Cao, "L3mvn: Leveraging large language models for visual target navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Oct. 2023. [Online]. Available: http://dx.doi.org/10.1109/IROS55552.2023.10342512

[51] B. Li, Y. Wang, J. Mao, B. Ivanovic, S. Veer, K. Leung, and M. Pavone, "Driving everywhere with large language model policy adaptation," 2024.

[52] D. Shah, B. Osinski, B. Ichter, and S. Levine, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," 2022.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[54] C. Xu, H. T. Nguyen, C. Amato, and L. L. S. Wong, "Vision and language navigation in the real world via online visual language mapping," 2023.

[55] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[56] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[57] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *International Journal of Social Robotics*, vol. 7, pp. 137–153, 2015.

[58] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" *arXiv preprint arXiv:2202.12837*, 2022.

[59] 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.