VLM-Social-Nav: Socially Aware Robot Navigation Through Scoring Using Vision-Language Models

Daeun Song, *Member, IEEE*, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, *Member, IEEE*, and Dinesh Manocha, *Member, IEEE*

***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.**

***Index Terms*—Motion and path planning, task and motion planning, integrated planning and control.**

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.

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] to service [2] and home assistance [3]. 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 [4], [5].

This letter focuses on the challenges of social navigation [5]. It addresses the ability of robots to navigate while adhering to social etiquette, especially contextual appropriateness, which requires robots to understand environment contexts, current tasks, and interpersonal behaviors. Therefore, navigating socially across varying contexts presents distinct challenges [4], [5], [6], including ensuring safety, comfort, and politeness, as well as adhering to social norms.

Inferring contextually appropriate navigation behaviors is challenging: Humans have various behaviors and the environmental or task contexts cannot be easily categorized [5]. 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], [11], [12] for training and reward

Received 6 July 2024; accepted 21 November 2024. Date of publication 4 December 2024; date of current version 11 December 2024. This article was recommended for publication by Associate Editor Harold Soh and Editor Hanna Kurniawati upon evaluation of the reviewers' comments. This work has taken place in the GAMMA Laboratory at the University of Maryland and the RobotiXX Laboratory at George Mason University. GAMMA was supported in part by ARO under Grant W911NF2310046 and Grant W911NF2310352 and in part by the US Army Cooperative Agreement under Grant W911NF2120076. RobotiXX was supported in part by National Science Foundation (NSF) under Grant 2350352, in part by Army Research Laboratory (ARL) under Grant W911NF2220242, Grant W911NF2320004, Grant W911NF2420027 and Grant W911NF2520011, in part by Air Force Research Laboratory (AFRL) and US Air Forces Central (AFCENT) under Grant GS00Q14OADU309, in part by Google DeepMind (GDM), in part by Clearpath Robotics, and in part by Raytheon Technologies (RTX). *(Corresponding author: Daeun Song.)*

Daeun Song, Amirreza Payandeh, Amir Hossain Raj, and Xuesu Xiao are with the Department of Computer Science, George Mason University, Fairfax, VA 22030 USA (e-mail: dsong26@gmu.edu; apayande@gmu.edu; araj20@gmu.edu; xiao@gmu.edu).

Jing Liang and Dinesh Manocha are with the Department of Computer Science, University of Maryland, College Park, MD 20742 USA (e-mail: jingl@umd.edu; dmanocha@umd.edu).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3511409, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3511409

engineering are required for their successful application and it is hard to generalize.

Language models are inherently well-suited for contextual understanding but not well applied in social navigation: 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]. 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 scenarios [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 social interaction.

Main Results: In this letter, 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. The social cost is defined as how well a robot's behavior aligns with socially acceptable norms, i.e., the behavior a human would likely exhibit. 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. We demonstrate VLM-Social-Nav in four different indoor scenarios with human interactions. Unlike previous social navigation approaches, VLM-Social-Nav can better navigate through social scenarios by interpreting the situation based on *common sense* without any dedicated training on a large dataset. Some of our main results include:

- We propose VLM-Social-Nav, a novel approach for social robot navigation, by integrating VLMs with optimization-based or scoring-based motion planners and a state-of-the-art perception model for better VLM efficiency.
- We propose a VLM-based scoring module that translates the current robot observation and textual instructions into a relevant social cost term. This cost term is used for the bottom-level motion planner to output appropriate robot action.
- We evaluate VLM-Social-Nav in four different real-world indoor social navigation scenarios along with a user study and compare the results with a Dynamic-Window Approach (DWA) [18] and Behavior Cloning (BC) [19] method trained on a state-of-the-art large Socially CompliAnt Navigation Dataset (SCAND) [11]. VLM-Social-Nav achieves at least 27.38% improvement in average success rate and 19.05% improvement in average collision rate in four scenarios. The user study score shows that VLM-Social-Nav generates the most socially compliant navigation behavior.

More related works and discussions can be found in the technical report of our manuscript [20].

II. Related Work

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

A. Safety Requirement of Social Navigation

For social navigation, safety is a basic requirement for interacting with humans and navigating dynamic scenarios [21], [22]. DWA [18] is a well-known collision-free navigation method that calculates collision constraints and selects the best feasible action. Velocity-Obstacle (VO)-based approaches are more efficient and can be used to simulate the actions of crowds [23], but they do not take uncertainties into account. PRVO [24] and OFVO [21] handle the perception uncertainties, but those approaches require a hard threshold for planning. To deal with this issue, learning-based methods empirically train the policies by demonstrations [25] or use reinforcement learning to train the robot in a simulator and implement it in real-world scenarios [22], [26]. However, learning-based approaches require a significant amount of data or realistic simulators to learn the task.

B. Contextual Appropriateness of Social Navigation

Researchers have developed various methods to incorporate social awareness into mobile robot navigation. Creating such systems is complex, requiring advanced perception and reasoning to navigate environments shared with humans and robots [4]. Defining social navigation varies across cultures and platforms. Assessing social compliance, beyond safety, depends on the scenario and requires contextual consideration. 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 [27], [28], [29]. SCAND [11] and MuSoHu [12] are two recent large-scale social human navigation datasets in many natural human-inhabited public spaces for robots to learn similar, human-like, socially compliant navigation behaviors. Although extensive research has explored various machine learning techniques, Vision-Language Models (VLMs) have not yet been applied to the social navigation problem, despite their strong potential for contextual analysis.

C. Large Foundation Models for Navigation

Recent advancements in Language Foundation Models (LFMs) [30], encompassing Vision-Language Models (VLMs) and Large-Language Models (LLMs), show significant potential in robotic navigation. SayCan [31] integrates LLMs for high-level task planning. GPT-Driver [32] evaluates the performance of GPT-3.5in simulation for autonomous driving, framing motion planning as a language modeling problem. L3MVN [33] constructs semantic maps of environments and utilizes LLMs to reach long-term goals, while LLaDA [34] enables autonomous vehicles to adapt to diverse traffic rules across regions. LM-Nav [16] utilizes GPT-3 and CLIP [35] to navigate outdoor environments based on natural language instructions, combining language and visual cues for optimal path planning. Despite their strengths in contextual understanding and commonsense reasoning, language models have not been investigated for social navigation. Our approach proposes a novel method to navigate robots in a socially compliant manner.

III. Approach

In this section, we define the social navigation problem and describe VLM-Social-Nav in detail.

A. Problem Definition

Navigation is the task of generating and following an efficient collision-free path from an initial location to a goal [5]. 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 vehicle dynamics and surrounding obstacles.

For social robot navigation, humans are no longer perceived only as dynamic obstacles but also as social entities [4]. 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], [11], [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 III-C.

B. VLM-Based Social Navigation Architecture

Fig. 2 highlights the overview of VLM-Social-Nav. Our approach is based on an autonomous navigation system that 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[36]. 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. Algorithm 1 summarizes an overview of our VLM-Social-Nav process.

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 [37].

Our VLM-based scoring module starts from the insight that the action space of a mobile robot can be readily mapped to

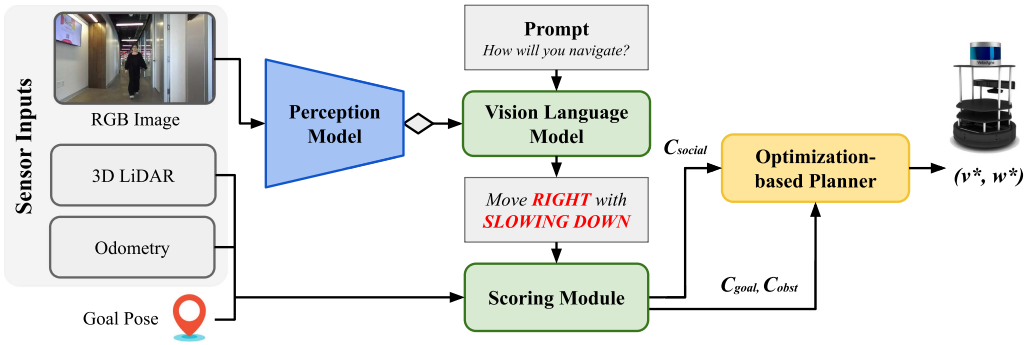

Fig. 2. The overall system architecture of VLM-Social-Nav. Our real-world perception (vision) model detects important social entities (e.g., humans, gestures, and doors) in real time and prompts the VLM-based scoring module to compute social cost $\mathcal{C}_{\text{social}}$, which is used to generate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

Algorithm 1: VLM-Based Social Navigation.

Input : RGB image $\mathcal{I}$, LiDAR point cloud $\mathcal{L}$, prompt $\mathcal{P}$, goal position $\mathbf{p}_g$

Initialize robot position $\mathbf{p}_r$;
while *not at goal position* $\mathbf{p}_r \neq \mathbf{p}_g$ **do**
  $\mathcal{I} \leftarrow$ Read image sensor data;
  $\mathcal{L} \leftarrow$ Read LiDAR sensor data;
  $e \leftarrow$ Perception Model($\mathcal{I}$);
  forall *possible actions* $\mathbf{a}$ **do**
    $\mathcal{C}_{\text{social}} \leftarrow 0$;
    if *social entities detected in* e **then**
      $\mathcal{B}_h = \text{VLM}(\mathcal{I}, \mathcal{P}, \mathbf{a})$;
      $\mathcal{C}_{\text{social}} = \text{VLM-based scoring}(\mathcal{B}_h)$;
    Calculate the total coast
    $\mathcal{C} = \alpha \cdot \mathcal{C}_{\text{goal}} + \beta \cdot \mathcal{C}_{\text{obst}} + \gamma \cdot \mathcal{C}_{\text{social}}$;
  end
  Find action $\mathbf{a}$ with minimal cost $\mathcal{C}$ and execute;
  Update robot position $\mathbf{p}_r$;
end

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 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:

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

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.

Fig. 3 shows an example prompt $\mathcal{P}$ used in our experiment. We provide a high-level task description along with an image

Input Image:

Input Prompt:

Task:
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.
- 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 input image ($\mathcal{I}^t$) and prompt ($\mathcal{P}$) used in VLM-Social-Nav. Parameterized inputs ($\mathbf{a}^t$) are highlighted in blue. Formatted outputs specifying the heading (δ_d) and the speed (δ_s) are highlighted in red. The example input data is one of the frontal approach scenarios from MuSoHu [12].

$\mathcal{I}^t$ 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 [37]. These guidelines not only facilitate comprehensive reasoning and judgment within the VLM but also enable the robot to adapt to specific rules more effectively. For example, in a country where it's customary to walk on the left, we can rephrase the prompt as "Move to the left when passing by another person."

IV. Experiments

In this section, we describe the details of the implementation and qualitative and quantitative experimental results.

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. In our experiment, we focus on key social cues, i.e., humans, doors, and gestures, which are critical considerations when navigating in socially rich environments [38]. 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. This follows a preliminary study that compared its performance with other large and small VLMs. GPT-4V was able to produce reliable results with high consistency, achieving a reasonable average inference time of around three seconds. We combined our approach with a low-level motion planner DWA [18]. We compare VLM-Social-Nav with DWA without social cost $\mathcal{C}_{\text{social}}$ and BC [19] trained on a state-of-the-art, large-scale social navigation dataset, SCAND [11]. The dataset contains various examples of socially compliant navigation behaviors teleoperated by humans including sticking to the right of the road and waiting for a human to pass. We expect that the model can learn to output socially compliant navigation behaviors like the human demonstrations.

Evaluating the social aspects of social robot navigation is inherently challenging [39]. To validate VLM-Social-Nav, we carefully follow the social robot navigation studies [6], [40], 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 trajectory. 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

Based on the protocols and principles set by other studies [6], [40], 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 that it does not obstruct the human path. Like driving rules in North America, it is conventional to keep on the right.
- **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 VLM-Social-Nav in four different 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. The accompanying supplementary video shows the resulting robot motions.

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. 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 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 IV-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. 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. VLM-Social-Nav improves the

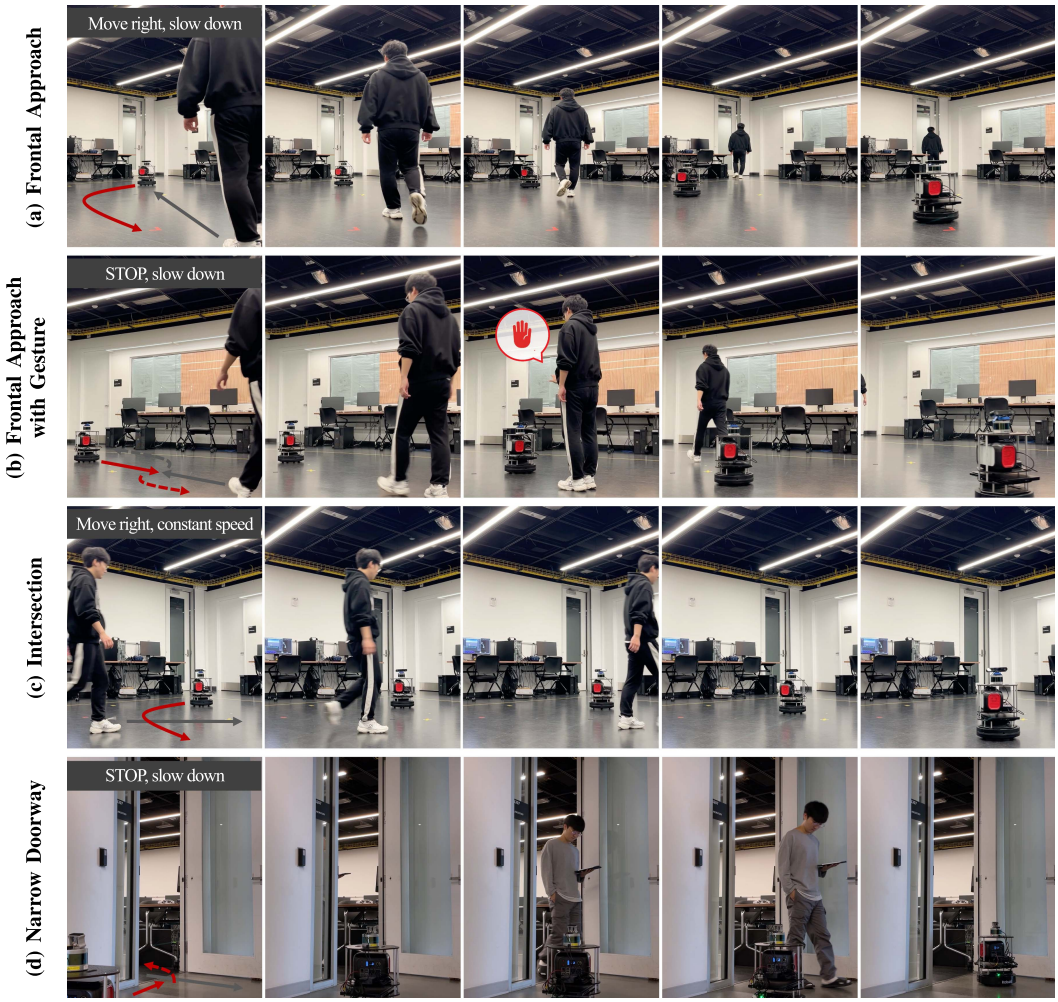

Fig. 4. Qualitative Results: the robot navigation behaviors with VLM-Social-Nav for four social navigation scenarios: (a) Frontal Approach, (b) Frontal Approach with Gesture, (c) Intersection, and (d) Narrow Doorway. The solid gray arrow shows the participant's path. The solid red arrow shows the robot's path. The red and gray dashed arrows show the robot's and participant's paths respectively, after a stop motion. A caption on the top left shows the result from the VLM.

TABLE I
QUANTITATIVE RESULTS: PERFORMANCE COMPARISONS USING BC [19], DWA [18], 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**

The bold values indicate the best-performing metrics for each scenario.

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 [40] (Table II). * denotes 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 seven participants. We use a five-level Likert scale to ask participants to rate their agreement with these statements.

Fig. 5 and the user study scores in 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

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 like 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	I felt like 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.

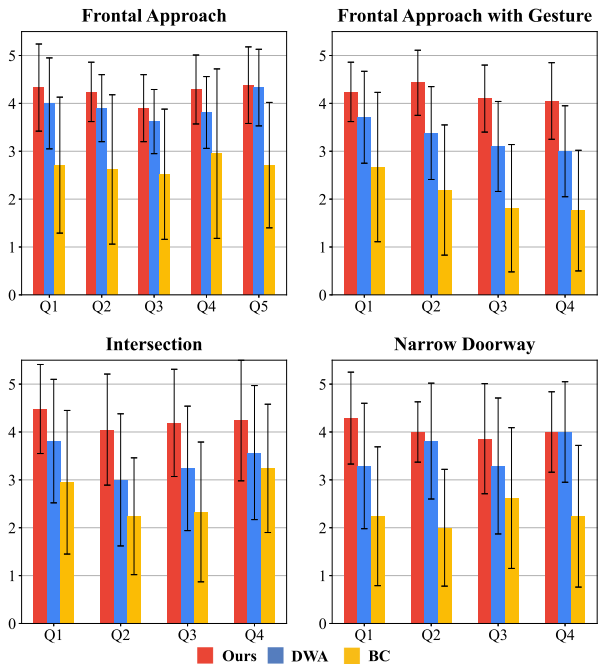

Fig. 5. User Study Average Scores: the per-question average scores for the three methods in each scenario. The results indicate that VLM-Social-Nav earned the highest level of agreement from participants across all questions, highlighting its robust alignment with social norms.

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. The score difference between VLM-Social-Nav and DWA was not large in the narrow doorway scenario. This is because, when attempting to enter the narrow doorway, DWA often failed to find a plan and froze, resembling the result of VLM-Social-Nav, a complete stop in front of the doorway.

E. Discussion

Real-time navigation with VLM: GPT-4 V and similar large VLMs require several seconds to respond to prompts, making continuous querying impractical for real-time navigation tasks. To address this, we optimized VLM-Social-Nav in two ways: first, by formatting prompts and providing predefined choices, which resulted in reduced response times. Second, we minimize queries by using a perception model to detect social cues, allowing for timely VLM queries only when necessary. These choices enable and allow average response times of 2-3 seconds, sufficient for human interaction and navigation. While such a limitation can be problematic in more dynamic scenarios that require frequent interactions, future advancements in fast large language models promise further extensions of our approach.

Socially aware navigation with VLM: We observe that VLMs can analyze and reason about social interactions from single images. Using various single images, including those collected by ourselves and from social robot navigation datasets [11], [12], VLMs accurately describe scenes and suggest socially compliant navigation strategies with reasons. For instance, for the image shown in Fig. 3, GPT-4 V describes the scene as *a person is walking towards the camera along a sidewalk*. To navigate this situation, GPT-4 V advises the robot to *yield the right of way* because *it is generally customary to keep to the right side of the path when encountering someone coming from the opposite direction, similar to driving rules*. However, despite their powerful capabilities, VLMs can still make mistakes. Therefore, relying solely on VLMs for navigation decisions is not safe. Instead, we incorporate their output as a cost term in our overall decision-making process.

More challenging scenarios: Although our robot experiments were conducted only indoors, according to the example in Fig. 3, VLM-Social-Nav can be extended to outdoor scenarios in more complex environments. VLMs successfully retrieve significant environmental information for outdoor social robot navigation, such as sidewalks, zebra crossings, and cars. This will be our immediate focus for future work, *i.e.*, to advance into global outdoor navigation. We also aim to extend our approach to complex scenarios involving multiple individuals. When tested with the scenario with one person, VLM-Social-Nav successfully outputs socially compatible actions, realizing the group of people in the scene. However, when multiple groups are present, simple directions like left or right may not suffice to describe effective robot navigation.

V. 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. However, interpreting social rules and deriving appropriate actions from them based on raw robot perception remains complex. VLMs interpret social situations and determine actions based on these rules, offering a more nuanced approach than simpler rule- or planning-based methods. We demonstrate and evaluate our system in four different real-world social navigation scenarios with a Turtlebot robot.

One immediate future work is to explore visual prompting methods [41], [42] to enhance spatial reasoning in VLMs by marking the images. Another promising future direction is to explore open-source VLMs, such as LLaVA [43]. Their access to lower-level information, such as log probabilities could help detect and address hallucinations. It is also interesting to develop a VLM that generates high-level social navigation instructions through chain-of-thought reasoning, and integrate it into an autonomous navigation system.

ACKNOWLEDGMENT

The majority of the work is conducted as a postdoctoral researcher at the University of Maryland.

REFERENCES

[1] S. Technology, "Starship," 2024. [Online]. Available: https://www.starship.xyz/
[2] D. Robotics, "Dilligent robotics," 2024. [Online]. Available: https://www.diligentrobots.com/
[3] Amazon, "Meet astro, a home robot unlike any other," 2024. [Online]. Available: https://www.aboutamazon.com/news/devices/meet-astro-a-home-robot-unlike-any-other
[4] R. Mirsky, X. Xiao, J. Hart, and P. Stone, "Conflict avoidance in social navigation—a survey," *ACM Trans. Hum.-Robot Interact.*, vol. 13, no. 1, pp. 1–36, 2024.
[5] C. Mavrogiannis et al., "Core challenges of social robot navigation: A survey," *ACM Trans. Hum.-Robot Interact.*, vol. 12, no. 3, pp. 1–39, 2023.
[6] A. Francis et al., "Principles and guidelines for evaluating social robot navigation algorithms," *ACM Trans. Hum.-Robot Interact.*, 2023, *arXiv:2306.16740*.
[7] N. Hirose, D. Shah, A. Sridhar, and S. Levine, "SACSoN: Scalable autonomous control for social navigation," *IEEE Robot. Automat. Lett.*, vol. 9, no. 1, pp. 49–56, Jan. 2024.
[8] A. H. Raj et al., "Rethinking social robot navigation: Leveraging the best of two worlds," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 16330–16337, doi: 10.1109/ICRA57147.2024.10611710.
[9] H. Kretzschmar et al., "Socially compliant mobile robot navigation via inverse reinforcement learning," *Int. J. Robot. Res.*, vol. 35, no. 11, pp. 1289–1307, 2016.
[10] 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 Robot. Automat. Lett.*, vol. 5, no. 2, pp. 676–682, Apr. 2020.
[11] H. Karnan et al., "Socially compliant navigation dataset (SCAND): A large-scale dataset of demonstrations for social navigation," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 11807–11814, Oct. 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 7442–7447.
[13] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 24824–24837.
[14] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
[15] L. Wen et al., "On the road with GPT-4V (ision): Explorations of utilizing visual-language model as autonomous driving agent," 2023, *arXiv:2311.05332*.
[16] D. Shah et al., "LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Proc. Conf. Robot Learn.*, 2023, pp. 492–504.
[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
[18] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robot. Automat. Mag.*, vol. 4, no. 1, pp. 23–33, Mar. 1997.
[19] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1988, pp. 305–313.
[20] D. Song, J. Liang, A. Payandeh, X. Xiao, and D. Manocha, "VLM-social-nav: Socially aware robot navigation through scoring using vision-language models," 2024, *arXiv:2404.00210*.
[21] J. Liang, Y.-L. Qiao, T. Guan, and D. Manocha, "OF-VO: Efficient navigation among pedestrians using commodity sensors," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 6148–6155, Oct. 2021.
[22] 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 *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 4221–4228.
[23] A. Best et al., "DenseSense: Interactive crowd simulation using density-dependent filters," in *Proc. Symp. Comput. Animation*, 2014, pp. 97–102.
[24] B. Gopalakrishnan et al., "PRVO: Probabilistic reciprocal velocity obstacle for multi robot navigation under uncertainty," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 1089–1096.
[25] 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. 69061–69081, 2021.
[26] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 740–759, Feb. 2022.
[27] P. Liu, D. F. Glas, T. Kanda, and H. Ishiguro, "Data-driven HRI: Learning social behaviors by example from human–human interaction," *IEEE Trans. Robot.*, vol. 32, no. 4, pp. 988–1008, Aug. 2016.
[28] M. Li, R. Jiang, S. S. Ge, and T. H. Lee, "Role playing learning for socially concomitant mobile robot navigation," *CAAI Trans. Intell. Technol.*, vol. 3, no. 1, pp. 49–58, 2018.
[29] M. Nazeri, J. Wang, A. Payandeh, and X. Xiao, "VANP: Learning where to see for navigation with self-supervised vision-action pre-training," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024.
[30] R. O. Bommasani, "On the opportunities and risks of foundation models," 2022, *arXiv:2108.07258*.
[31] A. Brohan et al., "Do as i can, not as i say: Grounding language in robotic affordances," in *Proc. Conf. Robot Learn.*, 2023, pp. 287–318.
[32] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT," 2023, *arXiv:2310.01415*.
[33] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging large language models for visual target navigation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 3554–3560.
[34] B. Li et al., "Driving everywhere with large language model policy adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 14948–14957.
[35] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[36] J. Duan et al., "Manipulate-anything: Automating real-world robots using vision-language models," 2024, *arXiv:2406.18915*.
[37] H. Peng, G. Li, Y. Zhao, and Z. Jin et al., "Rethinking positional encoding in tree transformer for code representation?" in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 3204–3214, doi: 10.18653/v1/2022.emnlp-main.210.
[38] T. Kruse, A. K. Pandey, R. Alami, and A. Kirsch, "Human-aware robot navigation: A survey," *Robot. Auton. Syst.*, vol. 61, no. 12, pp. 1726–1743, 2013.
[39] N. Tsoi, J. Romero, and M. Vázquez, "How do robot experts measure the success of social robot navigation?," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, 2024, pp. 1063–1066.
[40] S. Pirk, E. Lee, X. Xiao, L. Takayama, A. Francis, and A. Toshev, "A protocol for validating social navigation policies," 2022, *arXiv:2204.05443*.
[41] S. Nasiriany et al., "PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs," 2024, *arXiv:2402.07872*.
[42] K. Fang, F. Liu, P. Abbeel, and S. Levine, "MOKA: Open-vocabulary robotic manipulation through mark-based visual prompting," *Robot.: Sci. Syst.*, 2024. [Online]. Available: https://roboticsconference.org/2024/program/papers/62/
[43] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.