SocialNav-SUB: Benchmarking VLMs for Scene Understanding in Social Robot Navigation

Michael J. Munje[1], Chen Tang[1], Shuijing Liu[1], Zichao Hu[1], Yifeng Zhu[1], Jiaxun Cui[1],
Garrett Warnell[1,2], Joydeep Biswas[1,3], Peter Stone[1,4]

***Abstract*—Robot navigation in dynamic, human-centered environments requires socially-compliant decisions grounded in robust scene understanding, encompassing spatiotemporal awareness, as well as the ability to interpret human intentions. Recent Vision-Language Models (VLMs) show signs of object recognition, common-sense reasoning, and contextual understanding—capabilities that make them promising for addressing the nuanced requirements of social robot navigation. However, it remains unclear whether VLMs can reliably perform the complex spatiotemporal reasoning and intention inference needed for safe and socially compliant robot navigation. In this paper, we introduce the *Social Navigation Scene Understanding Benchmark (SocialNav-SUB)*, a Visual Question Answering (VQA) dataset and benchmark designed to evaluate VLMs for scene understanding of real-world social robot navigation scenarios. The benchmark provides a unified framework for evaluating VLMs against human and rule-based baselines across VQA tasks requiring spatial, spatiotemporal, and social reasoning in social robot navigation. Through experiments with state-of-the-art VLMs, we find that while the best-performing VLM achieves an encouraging probability of agreeing with human answers, it still lags behind a simpler rule-based approach and human performance, indicating critical gaps in social scene understanding of current VLMs. Our benchmark sets the stage for further research on foundation models for social robot navigation, offering a framework to explore how VLMs can be tailored to meet real-world social robot navigation needs.**

I. Introduction

Social robot navigation, defined as the ability for robots to move effectively and safely within human-populated environments while adhering to social norms, is a fundamental yet challenging task in robotics [20, 9]. As shown in Figure 1, navigating through social navigation scenarios requires robots to interpret human intentions, adhere to social norms, and respond to dynamic environments that demand advanced reasoning capabilities. While promising, existing methods still fall short in handling the complexity and nuance in dynamic real-world social navigation scenarios [20, 27].

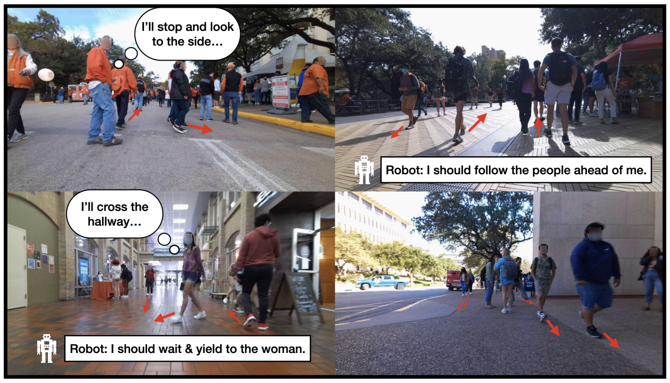
I'll stop and look to the side...
Robot: I should follow the people ahead of me.
I'll cross the hallway...
Robot: I should wait & yield to the woman.

Fig. 1: Examples of social robot navigation scenarios from SCAND [15] where humans in the scene have to be taken into consideration. The ability to determine socially compliant navigation actions requires understanding each dynamic scene by spatiotemporal reasoning (e.g. the movements of people in the scene), social reasoning (inferring the navigation intentions of people in the scene), and complying to implicit social rules.

Recently, the research community has begun to explore whether advances in large Vision-Language Models (VLMs) can be leveraged as part of a solution to social robot navigation, as they have demonstrated strong capabilities in contextual understanding, commonsense reasoning, and chain-of-thought reasoning [17, 24, 34]. Trained in diverse large-scale multimodal datasets that span various real-world scenarios, large VLMs often learn underlying patterns of human behavior that may implicitly encode an understanding of social norms [14]. Recent works like VLM-Social-Nav [32] have shown that using large VLMs for social robot navigation is promising, but their evaluations are limited to a small number of controlled scenarios and offer only preliminary insights. Moreover, studies such as SPACE [29] indicate that state-of-the-art large VLMs still lack robust spatial reasoning, raising questions about whether they can understand complex, realistic social navigation scenarios at all, let alone propose socially compliant actions for them.

In light of these limitations, it remains essential to systematically evaluate whether large VLMs can robustly handle what we consider as three critical dimensions of social robot navigation: **(1)** spatial reasoning, **(2)** spatiotemporal reasoning, and **(3)** the ability to interpret complex social navigation interactions. Existing evaluations have offered only partial assessments [32, 29], often focusing on controlled settings or lacking temporal components, leading to an incomplete picture of how effectively large VLMs can infer human intentions and comply with social norms in realistic, dynamic scenarios. This gap underscores the need for a comprehensive benchmark that rigorously tests these capabilities and may guide the development of VLMs specifically tailored to social robot navigation.

In this paper, we introduce the Social Navigation Scene Understanding Benchmark (SocialNav-SUB), a novel Visual

[1] Department of Computer Science, The University of Texas at Austin, Austin, TX, USA. Correspondence: michaelmunje@utexas.edu

[2] Garrett Warnell is also with DEVCOM Army Research Laboratory.

[3] Joydeep Biswas is also with NVIDIA.

[4] Peter Stone is also with Sony AI.

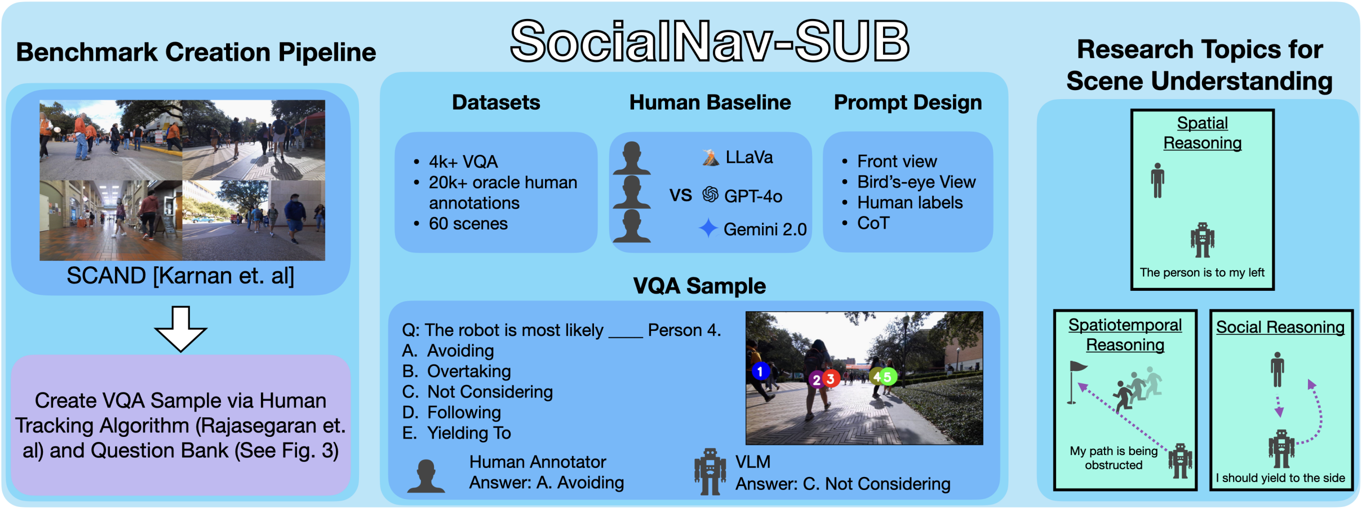

Fig. 2: An overview of SocialNav-SUB, which facilitates the systematic evaluation of VLMs in social robot navigation scenarios. Using SCAND data, human-labeled VQA datasets, and various VLMs, this framework offers the evaluation of VLMs across multiple dimensions of scene understanding for social robot navigation that can enable advancements in prompt designs, social reasoning, and social robot navigation research in general.

Question Answering (VQA) benchmark designed to evaluate VLMs on social robot navigation tasks. Our benchmark (summarized in Figure 2, utilizes data from a human-subject study conducted using social navigation scenarios from the SCAND dataset [15, 16], a robot social navigation dataset of socially compliant navigation demonstrations. We use our comprehensive human-labeled VQA dataset to serve as ground-truth labels for evaluating performance on scene understanding for social robot navigation. This approach enables the systematic evaluation of VLMs' scene understanding in the context of real-world social robot navigation scenarios, which allows us to run experiments on state-of-the-art large VLMs. Our experiments reveal *notable performance gaps between state-of-the-art large VLMs and both human and rule-based baselines, particularly in spatial and spatiotemporal reasoning.* Meanwhile, ablation studies show that chain-of-thought reasoning is crucial for improving model alignment with human judgments, while bird's-eye view prompts improve alignment for some models.

SocialNav-SUB is a first-of-its-kind benchmark that enables roboticists to systematically evaluate and refine VLMs for real-world social robot navigation scenarios. By bridging the gap between VLM capabilities and the challenges of social robot navigation, our work provides a foundation for advancing the use of VLMs for social robot navigation. Our contributions are as follows:

- **Social Navigation Scene Understanding Dataset:** We provide a human-labeled VQA dataset of 4968 unique questions and the accompanied 24840 human responses as a baseline oracle for social robot navigation tasks.
- **Social Navigation VQA Benchmark for VLMs:** We introduce the first VQA benchmark for assessing VLMs' capabilities in social robot navigation scenarios using 60 unique scenarios from SCAND.
- **Experiments using state-of-the-art large VLMs on our benchmark:** We evaluate several large VLMs (e.g., Gemini [34], GPT-4o [24], LLaVa-Next-Video [39]) on our benchmark against human and rule-based baselines. While Gemini outperforms other large VLMs, all models remain behind human and rule-based performance.
- **Insights into VLM Prompt Designs:** We identify and validate effective prompt strategies, such as bird's-eye view prompts and chain-of-thought reasoning, that improve agreement with human answers.

II. Related Work

Our work intersects with several key areas of research. We organize our discussion of related work into three main categories: (1) VLMs in Robotics, (2) Social Robot Navigation, and (3) VQA Benchmarks for VLMs.

A. VLMs in Robotics

In robotics, VLMs have demonstrated considerable potential for various tasks such as robotic manipulation [22], task planning [38], and human-robot interaction [1, 2, 8]. The success of VLMs can be attributed to their ability to associate vision and language and generalize to unseen data in a zero-shot manner. For navigation, VLMs have been used for waypoint specification [22, 30], and instruction following [6, 36, 12]. However, these approaches often struggle in complex real-world environments, particularly in dynamic environments, due to limitations in VLMs' spatial reasoning capabilities [29, 3, 33]. This gap highlights the need for specialized evaluations and improvements of VLMs for tasks in dynamic environments, especially social navigation.

B. Social Robot Navigation

Early social robot navigation approaches relied on model-based techniques, such as the Social Force Model (SFM)

[10] and proxemics-based methods [21], which used hand-engineered features to plan paths for robots. Learning-based techniques for social robot navigation, including Learning from Demonstration (LfD) [11, 15] and Reinforcement Learning (RL) [40, 5, 4, 18, 19], have shown promise in enabling robots to acquire and adapt socially compliant behaviors. However, trained on small and specialized data or simulations, these methods often struggle to generalize to complex dynamic scenarios. To address this, datasets for social robot navigation [15] [23] have been developed to provide more diverse and realistic social navigation scenarios, which can lead to improved generalization in social navigation models [13].

More recently, social robot navigation datasets for VLMs have been explored [25], but are limited to qualitative evaluation and single images, when crucial information, such as a person's trajectory, may require a video representation. Fine-tuned VLMs have been explored for social robot navigation [32, 25], but are often evaluated in a limited number of simple, controlled scenarios. These scattered findings suggest that while VLMs *may* enhance social robot navigation, the specific capabilities that drive any observed improvements have yet to be clearly identified. Our work addresses this by introducing a specialized benchmark to systematically evaluate whether VLMs can effectively perform spatial reasoning, spatiotemporal reasoning, and social reasoning for numerous social navigation scenarios.

C. VQA Benchmarks for VLMs

Recent years have seen the development of various VQA benchmarks to evaluate VLMs, assessing capabilities such as spatial reasoning [29], scene understanding for autonomous driving [35], and physical world comprehension [7]. While these benchmarks have advanced our understanding of VLMs' capabilities, they often lack specific focus on social robot navigation; which requires not only spatial reasoning, but also spatiotemporal reasoning, and social reasoning to interpret complex social navigation interactions. Our work addresses this gap by introducing a specialized VQA benchmark for social robot navigation.

III. SocialNav-SUB

To evaluate Vision-Language Models (VLMs) on scene understanding for social robot navigation, we present the **Social Navigation Scene Understanding Benchmark (SocialNav-SUB)**, a VQA benchmark for evaluating VLMs in socially dense navigation scenarios. Following recent works that have demonstrated the effectiveness of visual grounding and object-centric representations [22, 37, 35], we provide numbered labels within visual markers for objects of relevance (in our case, pedestrians) for prompting and object-centric annotations; this provides the benchmarked VLMs clear visual references and contextually rich instructions. SocialNav-SUB is built on top of the SCAND dataset's social navigation scenarios that provide varying levels of crowd density and social navigation interactions and features the following:

(1) *Challenging social navigation scenarios* that capture the complexities of crowded and dynamic human environments;
(2) *Object-centric representations* combining both the robot's visual perspective and a bird's-eye view (BEV) containing pedestrian coordinate tracking for a richer object-centric representation;
(3) *A diverse question set* probing spatial reasoning, temporal understanding, and social reasoning; and
(4) *A robust human baseline*, where multiple annotators provide ground-truth responses for each scenario.

All above features are expanded in the following subsections below.

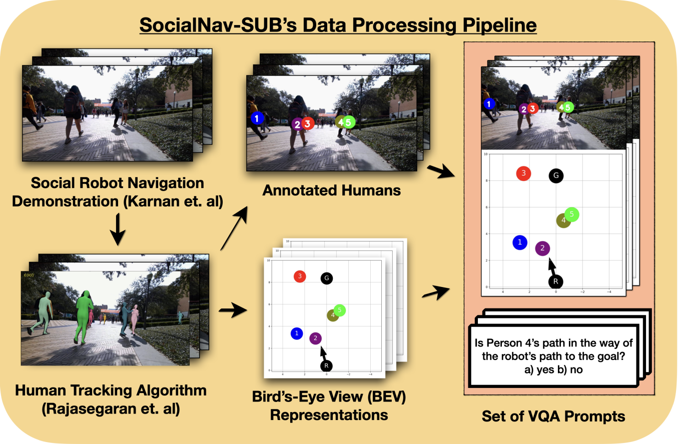
SocialNav-SUB's Data Processing Pipeline
Social Robot Navigation Demonstration (Karnan et. al)
Annotated Humans
Human Tracking Algorithm (Rajasegaran et. al)
Bird's-Eye View (BEV) Representations
Set of VQA Prompts
1
2
3
4
5
G
R
Is Person 4's path in the way of the robot's path to the goal? a) yes b) no

Fig. 3: The data processing pipeline for VQA prompts in SocialNav-SUB. We first mine social robot navigation scenarios from SCAND [15], then use the PHALP algorithm [28] to provide human tracking and estimations of 3D locations, which are used to construct BEV representations of the scene and annotated images. Along with the annotated images and BEV representations, a set of carefully designed questions that evaluate spatial reasoning, spatiotemporal reasoning, and social reasoning are used to provide VQA prompts.

A. Challenging Social Navigation Scenarios

To effectively evaluate VLMs' scene understanding capabilities in practical social robot navigation settings, we leverage the SCAND dataset [15] to construct the SocialNav-SUB benchmark. SCAND features social robot navigation data collected by teleoperated mobile robots navigating in diverse and potentially crowded scenarios. In particular, we extract segments from SCAND that showcase high crowd density, close pedestrian proximity, and dynamically changing human motion. As illustrated in Figure 1, these densely occupied scenarios typically involve pedestrians that obstruct the robot's direct path to its goal. Hence, the teleoperated robots demonstrate complex, socially compliant interactions with the pedestrians, making these samples valuable for evaluating VLMs' scene understanding capabilities in real-world social navigation environments.

B. Rich and Object-Centric Visual Representations

The samples extracted from the SCAND dataset are in the form of RGB image sequences captured by the front-view camera mounted on the robot. While 2D image sequences may suffice for humans to infer the underlying spatial and social relations between the robots and pedestrians, state-of-the-art large VLMs are not necessarily good at extracting spatial or fine-grained object-level information from the same visual queries [29]. To mitigate this issue, some recent studies have shown that augmenting images with additional annotations (e.g., bounding boxes, color-coded labels) using off-the-shelf models can improve VLM performance in VQA tasks [22, 37]

Building on these insights, we augment the original data samples with additional *object-centric* representations leveraging off-the-shelf vision models. Specifically, as shown in Figure 3, we annotate the pedestrians in the raw front-view images with numbered, color-coded circles and generate additional bird's-eye-view (BEV) images illustrating the robots' and pedestrians' locations. The resulting images with combined views preserve the original scene context while providing additional spatial and object-level information—such as distances and obstructed paths—in a clear and structured format.

By querying VLMs with these enriched, object-centric visual inputs, our SocialNav-SUB benchmark could provide practical insights into how to best leverage and complement state-of-the-art large VLMs for practical application in social robot navigation. To ensure fair comparisons between VLMs' outputs and human responses, the same set of visual inputs are provided to human annotators.

Figure 3 illustrates our data processing pipeline for augmenting the raw front-view images from SCAND. We begin by employing the human tracking algorithm, PHALP [28], which tracks pedestrians and provides robust estimations of their 3D poses relative to the camera frame using monocular video input. Using the robot odometry data from SCAND, we transform the relative human poses at future timesteps into global poses relative to the robot pose in the initial frame, and apply Kalman smoothing to smooth the human poses. Afterwards, we use the camera intrinsics and extrinsics provided by SCAND to project the 3D coordinates of pedestrians into both front-view and BEV images. Finally, we annotate human positions and their correspondences in both views with numbered, color-coded circles.

C. Diverse Scene Understanding Questions

Following the aforementioned data processing pipeline, we construct a set of samples consisting of multi-view image sequences with object-centric annotations, each representing a $2.5\,\mathrm{s}$ segment sampled at $4\,\mathrm{Hz}$. To comprehensively evaluate VLMs' scene understanding capabilities in social robot navigation, we design a range of multiple-choice questions (see Table I for more details) that probe across three categories:

- **Spatial reasoning:** Questions about describing the *spatial relations* in a *single frame*.
- **Spatiotemporal reasoning:** Questions about describing the *motion* of the robot and pedestrians *over time*.
- **Social reasoning over time:** Questions that *infer whether* the robot and pedestrians are interacting and *how* they interact.

These three categories map onto what we see as being the key challenges of social navigation: perceiving spatial relations among participants (spatial reasoning), tracking their evolution as people move (spatiotemporal reasoning), and recognizing how humans and robots interact in the context of social navigation (social reasoning over time). By evaluating VLM performance across these dimensions, we gain a fine-grained understanding of where models excel or struggle in parsing and interpreting social navigation scenes.

D. Robust Human Baseline from Human-Subject Study

We conducted human-subject studies to collect human responses as ground-truth labels for these questions under an IRB-approved protocol. Given the subjective nature of many questions, particularly those related to social reasoning, we collected responses from five human participants for each scenario. Participants were recruited via Prolific [26] and were asked to complete a questionnaire containing questions for multiple randomly sampled scenarios. To ensure the quality of the collected responses, we added attention-check questions to the questionnaire and manually inspected the participants' answers to reject low-quality samples.

By gathering this distribution of human responses, we can measure how closely each VLM output aligns with human judgments. Specifically, we compute the agreement between VLM answers and all human answers for a given question, which indicates the extent to which a model's performance approaches human-level responses. Moreover, to establish an *Average Human* performance baseline, we also measure how often one human's response agrees with all other human responses for a given question, capturing the natural variability and consensus levels among human annotators. This baseline thus provides a robust point of comparison for evaluating VLM performance on social robot navigation tasks. Additionally, we also utilize the consensus answers from the human data to construct a *Human Oracle* baseline, serving as an upper bound for performance.

We define two metrics, **Probability of Agreement (PA)** and **Consensus-Weighted PA**, to measure how closely a set of answers (from a VLM, a particular human, or a rule-based baseline) aligns with human responses overall.

Notation and Setup.

- N_Q: total number of questions.
- N_H: number of human respondents per question.
- A_q: the evaluated answer (from a VLM or one human) to question q.
- $A^h_{q,i}$: the i-th human's answer for question q, where $i \in \{1, \ldots, N_H\}$.

We define *Probability of Agreement (PA)* as:

$$\mathrm{PA} \;=\; \frac{1}{N_Q} \sum_{q=1}^{N_Q} \Big(\frac{1}{N_H} \sum_{i=1}^{N_H} \mathbb{I}[\, A_q = A^h_{q,i} \,] \Big), \qquad (1)$$

VLM Capability	Qualitative Description of Question	# of Questions
Spatial Reasoning	**Person Initial Position**: The position of the person at the beginning of the video.	399
	Person Ending Position: The position of the person at the end of the video.	399
	Goal Initial Position: The initial position of the goal with respect to the robot's view.	60
	Goal End Position: The end position of the goal with respect to the robot's view.	60
	Person End Goal Obstruction: Whether the person is obstructing the robot's path towards the goal at the end of the video.	399
Spatiotemporal Reasoning	**Robot Moving Direction**: The direction the robot is moving in the video.	60
	Person Distance Change: The relative distance change of the person to the robot from the beginning of the video to the end.	399
	Person Goal Obstruction: Whether the person is obstructing the robot's path towards the goal during the video.	399
Social Reasoning	**Robot Affected by Person**: Whether the robot's (human operator's) actions are affected by the person.	399
	Robot Action to Person: The high-level relational action of the robot with respect to the person (e.g., the robot avoided person 2).	399
	Person Affected by Robot: Whether the robot's (human operator's) actions are affected by the person.	399
	Person Action to Robot: The high-level relational action of the person with respect to the robot (e.g., person 2 avoided the robot).	399
	Robot Affected by Person at End: Whether the robot's (human operator's) actions are affected by the person at the end of the video.	399
	Robot Action to Person at End: The high-level relational action of the robot with respect to the person at the end of the video.	399
	Person Action to Robot at End: The high-level relational action of the person with respect to the robot at the end of the video.	399

TABLE I: Qualitative descriptions of the text components for questions used in SocialNav-SUB, their pertaining primary reasoning capability, and the number of unique questions through SocialNav-SUB. All questions are multiple choice, with each VQA prompt providing the possible answers. An example of a VQA prompt can be found in Figure 2.

where $\mathbb{I}[\cdot]$ is an indicator function that is 1 if A_q (the evaluated answer) exactly matches the i-th human's response $A^h_{q,i}$, and 0 otherwise for the corresponding multiple-choice question q. Summing over all human responses for each question yields the fraction of total (answer, human answer) pairs that agree. A higher PA indicates that the evaluated answers coincide more frequently with the collected human responses.

We empirically found that it was not uncommon for humans to disagree on answers, indicating there is a degree of judgement involved for particular questions. This motivates a metric that can be more forgiving for subjective questions that humans disagree on and emphasize questions that have a strong consensus, to which we establish *Consensus-Weighted Probability of Agreement (CW PA)*. We start by defining

$$\mathrm{HA}_q = \max_{\alpha}\left\{\frac{\#(\text{humans who answered } \alpha \text{ for question } q)}{N_H}\right\},$$

i.e., HA_q is the fraction of human respondents that chose the most common answer α for question q. We then define:

$$\text{CW PA} = \frac{1}{N_Q}\sum_{q=1}^{N_Q}\Big(\frac{1}{N_H\,\mathrm{HA}_q}\sum_{i=1}^{N_H}\mathbb{I}[\,A_q = A^h_{q,i}\,]\Big). \qquad (2)$$

In this formulation, each agreement with a human response for question q is scaled by $1/\mathrm{HA}_q$. Consequently, questions on which humans mostly concur (i.e., high HA_q) impose a greater penalty for incorrect answers, while questions where humans are more divided have a lower penalty. This weighting ensures that VLM (or human) answers are held to a higher standard on ”easier” questions where strong human agreement exists.

IV. Experiments

A. Research Question

Our central research question examines *how well state-of-the-art large VLMs that support image sequences capture spatial reasoning, scene understanding, and social reasoning in social robot navigation scenarios*. By focusing on this question, we aim to rigorously assess the capabilities and limitations of large VLMs for understanding complex social robot navigation environments.

B. Experiment Process

Our experiment process begins by presenting survey prompts alongside their visual and BEV representations to the VLM, using the data processing pipeline previously shown in Figure 3. The format given to the VLMs closely resembles the same visual and text format that was received by human participants, ensuring fair comparison. Furthermore, we use chain-of-thought reasoning as a prompting technique to carry out our experiments, since this is highly similar to the sequential manner in which humans provided answer labels, allowing for fair comparison. Specifically, our usage of chain-of-thought provides the previous answers of the VLM for future questions which may help it deduce the answer to question; for example, the pedestrian is at the left in the beginning and the end and the goal is on the right, so the pedestrian is likely not obstructing the path to the goal. The responses generated by the VLM are then compared against human responses from the human dataset using the previously defined Probability of Agreement (PA) metric.

Humans can naturally infer the underlying spatial and social relations between the robots and pedestrian, making them

Category	Model	All	Spatial Reasoning	Spatiotemporal Reasoning	Social Reasoning
Baseline	Human Oracle	0.74 ± 0.00	0.71 ± 0.01	0.73 ± 0.01	0.76 ± 0.01
	Rule-Based	0.64 ± 0.00	0.57 ± 0.01	0.62 ± 0.01	0.71 ± 0.00
VLM	Gemini 2.0	0.58 ± 0.00	0.55 ± 0.01	0.46 ± 0.01	0.63 ± 0.01
	GPT-4o	0.50 ± 0.00	**0.56 ± 0.01**	0.51 ± 0.01	0.47 ± 0.01
	LLaVa-Next-Video	0.46 ± 0.01	0.35 ± 0.01	0.58 ± 0.01	0.48 ± 0.01

TABLE II: Average Performance Across Question Categories. We compute the Probability of Agreement (PA) for all questions and for each question category, along with standard error across the unique questions.

excellent references for comparing VLM performance to. On the other hand, are large VLMs truly necessary for analyzing these social robot navigation scenarios, or can a simpler, rule-based system suffice? To address both of these, our baselines are as follows:

(1) *Human Oracle Baseline*: Selects the most common answer for each question from the human distribution. This baseline serves as an upper bound for performance when models may only provide one answer.
(2) *Rule-Based Baseline*: Uses the position data of pedestrians in the scene (extracted using the Optical Character Recognition (OCR) algorithm Tesseract [31]) and uses a set of hand-crafted rules to generate answers to VQA prompts.

C. Results

We run our experiments by querying each VLM model once per unique question using default hyperparameters for each VLM. The average results over all questions and question categories is shown in Table II, which indicate that average human performance serves as a reliable baseline. Among the large VLMs evaluated, Gemini achieves the highest overall performance, but still has a considerable gap compared to the human oracle and Rule-Based baselines. This performance gap suggests that state-of-the-art large VLMs are not yet fully ready for the challenges of scene understanding for social robot navigation.

When examining performance across the three question categories, models consistently lag behind the human oracle and the Rule-Based baseline, though the extent of the gap varies by category. In spatial reasoning, the consensus among humans (human oracle) far exceeds that of the best models, indicating that current large VLMs struggle to accurately interpret static spatial relationships compared to human observers. A similar finding is observed in spatiotemporal reasoning, where models show even greater difficulty at capturing dynamic changes and interactions over time. In contrast, in social reasoning tasks, models perform relatively closer to human consensus levels, suggesting that large VLMs are somewhat more adept at interpreting social cues and interactions than they are at understanding spatial relationships, although there remains a noticeable gap. Empirically, we found many cases of VLMs failing on questions with high human consensus in all three reasoning categories, especially in cases of high crowd densities.

Overall, our evaluation reveals that while state-of-the-art large VLMs like Gemini show promising advances, they still fall short of human and rule-based performance across key reasoning tasks. Although models come closer to human oracle performance in social reasoning tasks, the results suggest that significant improvements to large VLM architectures or refining querying strategies are needed before these large VLMs can reliably support complex, real-world social robot navigation.

V. Limitations and Future Work

While SocialNav-SUB advances the evaluation of VLMs for social robot navigation, it has two major limitations. First, the benchmark currently relies on scenarios from the SCAND dataset, which is limited to social navigation in a university campus setting. Second, while initial experiments provide valuable insights, they are based on a limited set of models and scenarios; further exploration with a broader range of large VLMs, datasets, and refined methodologies is necessary to overcome these challenges and enhance the benchmark's applicability.

Looking ahead, several promising avenues can further enhance and leverage the capabilities of SocialNav-SUB. First, expanding the dataset to include additional social robot navigation datasets could expand its diversity and robustness, offering a more comprehensive evaluation of model capabilities. Additionally, fine-tuning VLMs on the human dataset provided in SocialNav-SUB may lead to VLMs that are more capable of social robot navigation. Another promising avenue is expanding upon the VLM models evaluated; some VLMs of interest include VLMs fine-tuned for spatial reasoning and VLMs fine-tuned for social robot navigation. Lastly, an interesting future direction is evaluating hybrid approaches that utilize VLMs in specific ways (such as social reasoning) while having dedicated modules to cover their weaknesses. By offering a targeted evaluation framework across multiple reasoning categories, SocialNav-SUB can not only systematically evaluate VLM performance and highlight weaknesses but also guide future improvements in VLMs for both scene understanding and socially compliant navigation, enabling the development of more reliable real-world robotics systems.

VI. Conclusion

This paper introduced the Social Navigation Scene Understanding Benchmark (SocialNav-SUB), a novel VQA benchmark designed to evaluate VLMs within complex social robot navigation scenarios. Drawing on densely populated and dynamic environments from the SCAND dataset, SocialNav-SUB provides object-centric visual representations, including augmented front-view images and BEV prompts paired with a diverse set of questions targeting spatial, spatiotemporal, and social reasoning. By grounding these evaluations with a robust human-subject study, the benchmark offers clear, quantifiable metrics that reflect human-like understanding and decision-making in social navigation contexts.

SocialNav-SUB advances the state of the art by highlighting specific strengths and weaknesses of current VLMs in handling intricate social scenes, thereby setting a clear agenda for future research. It enables researchers to systematically compare models, refine prompting strategies, and develop new methods to bridge the gap between machine and human understanding of social robot navigation interactions. The benchmark's comprehensive design supports the iterative improvement of VLMs in real-world applications, ultimately guiding the development of more socially aware and reliable robotic systems.

ACKNOWLEDGMENT

This work has taken place in the Learning Agents Research Group (LARG) and the Autonomous Mobile Robotics Laboratory (AMRL) at The University of Texas at Austin. LARG research is supported in part by the National Science Foundation (FAIN-2019844, NRT-2125858), the Office of Naval Research (N00014-18-2243), Army Research Office (W911NF-23-2-0004, W911NF-17-2-0181), Lockheed Martin, and Good Systems, a research grand challenge at the University of Texas at Austin. AMRL research is supported in part by NSF (CAREER-2046955, OIA-2219236, DGE-2125858, CCF-2319471), ARO (W911NF-23-2-0004), Amazon, and JP Morgan. The views and conclusions contained in this document are those of the authors alone. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

[1] Peixin Chang, Shuijing Liu, and Katherine Driggs-Campbell. Learning visual-audio representations for voice-controlled robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[2] Peixin Chang, Shuijing Liu, Tianchen Ji, Neeloy Chakraborty, Kaiwen Hong, and Katherine Rose Driggs-Campbell. A data-efficient visual-audio representation with intuitive fine-tuning for voice-controlled robots. In *Conference on Robot Learning (CoRL)*, 2023.

[3] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

[4] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6015–6022, 2019.

[5] Yu Fan Chen, Michael Everett, Miao Liu, and Jonathan P How. Socially aware motion planning with deep reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1343–1350, 2017.

[6] Hao-Tien Lewis Chiang, Zhuo Xu, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, Fei Xia, Jasmine Hsu, Jonathan Hoech, Pete Florence, Sean Kirmani, Sumeet Singh, Vikas Sindhwani, Carolina Parada, Chelsea Finn, Peng Xu, Sergey Levine, and Jie Tan. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs, 2024. URL https://arxiv.org/abs/2407.07775.

[7] Wei Chow*, Jiageng Mao*, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025.

[8] Zichao Dong, Weikun Zhang, Xufeng Huang, Hang Ji, Xin Zhan, and Junbo Chen. Hubo-vlm: Unified vision-language model designed for human robot interaction tasks. *arXiv preprint arXiv:2308.12537*, 2023.

[9] Anthony Francis, Claudia Pérez-D'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, Hao-Tien Lewis Chiang, Michael Everett, Sehoon Ha, Justin Hart, Jonathan P. How, Haresh Karnan, Tsang-Wei Edward Lee, Luis J. Manso, Reuth Mirksy, Sören Pirk, Phani Teja Singamaneni, Peter Stone, Ada V. Taylor, Peter Trautman, Nathan Tsoi, Marynel Vázquez, Xuesu Xiao, Peng Xu, Naoki Yokoyama, Alexander Toshev, and Roberto Martín-Martín. Principles and guidelines for evaluating social robot navigation algorithms, 2023. URL

https://arxiv.org/abs/2306.16740.

[10] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[11] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 2023.

[12] Noriaki Hirose, Catherine Glossop, Ajay Sridhar, Dhruv Shah, Oier Mees, and Sergey Levine. Lelan: Learning a language-conditioned navigation policy from in-the-wild videos, 2024. URL https://arxiv.org/abs/2410.03603.

[13] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1): 49–56, 2024. doi: 10.1109/LRA.2023.3329626.

[14] Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. Viva: A benchmark for vision-grounded decision-making with human values, 2024. URL https://arxiv.org/abs/2407.03000.

[15] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Soeren Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation, 2022. URL https://arxiv.org/abs/2203.15041.

[16] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Soeren Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially Compliant Navigation Dataset (SCAND), 2022. URL https://doi.org/10.18738/T8/0PRYRH.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

[18] Shuijing Liu, Peixin Chang, Zhe Huang, Neeloy Chakraborty, Kaiwen Hong, Weihang Liang, D Livingston McPherson, Junyi Geng, and Katherine Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 12015–12021, 2023.

[19] Shuijing Liu, Haochen Xia, Fatemeh Cheraghi Pouria, Kaiwen Hong, Neeloy Chakraborty, and Katherine Driggs-Campbell. Height: Heterogeneous interaction graph transformer for robot navigation in crowded and constrained environments. *arXiv preprint arXiv:2411.12150*, 2024.

[20] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *J. Hum.-Robot Interact.*, 12(3), April 2023. doi: 10.1145/3583741. URL https://doi.org/10.1145/3583741.

[21] Jonathan Mumm and Bilge Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 331–338, 2011. doi: 10.1145/1957656.1957786.

[22] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024. URL https://arxiv.org/abs/2402.07872.

[23] Duc M Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7442–7447. IEEE, 2023.

[24] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark et. al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[25] Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dinesh Manocha, and Xuesu Xiao. Social-llava: Enhancing robot navigation through human-language reasoning in social spaces. *arXiv preprint arXiv:2501.09024*, 2024.

[26] Prolific. Prolific. https://www.prolific.com, 2014. *Accessed on [date accessed].*

[27] Amir Hossain Raj, Zichao Hu, Haresh Karnan, Rohan Chandra, Amirreza Payandeh, Luisa Mao, Peter Stone, Joydeep Biswas, and Xuesu Xiao. Targeted learning: A hybrid approach to social robot navigation. *arXiv preprint arXiv:2309.13466*, 2023.

[28] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2740–2749, 2022.

[29] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models?, 2024. URL https://arxiv.org/abs/2410.06468.

[30] Adarsh Jagan Sathyamoorthy, Kasun Weerakoon, Mohamed Elnoor, Anuj Zore, Brian Ichter, Fei Xia, Jie Tan, Wenhao Yu, and Dinesh Manocha. Convoi: Context-aware navigation using vision language models in outdoor and indoor environments, 2024. URL https://arxiv.org/abs/2403.15637.

[31] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[32] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models, 2024. URL https://arxiv.org/abs/2404.00210.

[33] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang,

Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning, 2024. URL https://arxiv.org/abs/2410.16162.

[34] Gemini Team and Petko Georgiev et. al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[35] Weizhen Wang, Chenda Duan, Zhenghao Peng, Yuxin Liu, and Bolei Zhou. Embodied scene understanding for vision language models via metavqa. *arXiv preprint arXiv:2501.09167*, 2025.

[36] Kasun Weerakoon, Mohamed Elnoor, Gershom Seneviratne, Vignesh Rajagopal, Senthil Hariharan Arul, Jing Liang, Mohamed Khalid M Jaffar, and Dinesh Manocha. Behav: Behavioral rule guided autonomy using vlms for robot navigation in outdoor scenes, 2024. URL https://arxiv.org/abs/2409.16484.

[37] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[38] Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Guiding long-horizon task and motion planning with vision language models. *arXiv preprint arXiv:2410.02193*, 2024.

[39] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[40] Kai Zhu and Tao Zhang. Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, 26(5):674–691, 2021.