2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3315-4393-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IROS60139.2025.11247618

# Social-LLaVA: Enhancing Social Robot Navigation through Human-Language Reasoning

Amirreza Payandeh[1], Daeun Song[1], Mohammad Nazeri[1], Jing Liang[2], Praneel Mukherjee, Amir Hossain Raj[1], Yangzhe Kong[1], Dinesh Manocha[2], and Xuesu Xiao[1]

***Abstract*— As mobile robots become increasingly common in human-centric environments, social navigation—adhering to unwritten social norms rather than merely avoiding pedestrians—has drawn growing attention. Existing methods, from hand-crafted techniques to learning-based approaches, often overlook the nuanced context and scene understanding that humans naturally exhibit. Inspired by studies indicating the critical role of language in cognition and reasoning, we propose a new approach to bridge robot perception and socially aware actions through human-like language reasoning. We introduce Social robot Navigation via Explainable Interactions (SNEI), a human-annotated vision-language dataset comprising over 40K Visual Question Answering (VQA) pairs across 2K unique social scenarios, drawn from diverse, unstructured public spaces. SNEI contains perception, prediction, chain-of-thought reasoning, action, and explanation, thereby allowing robots to interpret social contexts in human language. We fine-tune a Vision-Language Model, Social-LLaVA, on SNEI to demonstrate the potential of language-guided reasoning for high-level navigation tasks. Experimental evaluations—both quantitative and qualitative—demonstrate that Social-LLaVA can outperform state-of-the-art models.[†].**

## I. Introduction

As mobile robots become more prevalent in human-centric environments, there is a growing interest in social navigation, augmenting traditional methods by aligning with human social norms and rules rather than merely treating humans as dynamic or static obstacles [1]. An extensive body of work has addressed social robot navigation, ranging from employing various hand-crafted navigation techniques based on geometric and semantic understanding [2]–[4] to learning-based methods using large-scale datasets [5], [6]. While these methods have made progress toward achieving socially compliant behaviors, they often fail to grasp the nuances of the context and scene in the same way humans do.

Several studies suggest that language significantly influences human cognition, reasoning, and comprehension of the world [7]. We posit that if robots' actions are guided by reasoning processes similar to human language-based thought, they could exhibit behavior that more closely resembles human behavior. If robots can perceive a scene, make predictions like humans, engage in reasoning, and generate action descriptions in human language, they are more likely to translate their perception into actions that closely resemble human behaviors Fig. 1.

[1]George Mason University {apayande, dsong26, mnazerir, araj20, ykong7, xiao}@gmu.edu, praneel.mukherjee@gmail.com
[2]University of Maryland, College Park {jingl, dmanocha}@umd.edu
[†]Website: https://cs.gmu.edu/~xiao/Research/SNEI/


Fig. 1: Overview of Social-LLaVA, bridging perception to socially compliant action through Chain-of-Thought reasoning using SNEI, a vision-language dataset for social robot navigation.

To bring human language applications to robot navigation, Vision-Language Models (VLMs) and their reasoning capabilities have received significant attention. There have been several adaptations of VLMs in autonomous vehicles (AVs) [8], [9], propelled by the availability of large-scale datasets developed for autonomous driving in structured road networks governed by clear traffic rules [9], [10]. Trained on general image and language datasets, existing VLMs struggle to understand complex social interactions, particularly in spatiotemporal commonsense reasoning and the high-level decision-making required for social robot navigation [11], [12]. Moreover, AV datasets are not directly applicable to social robot navigation. Unlike road environments, human-populated spaces are inherently unstructured and guided by unwritten, context-dependent social norms that can vary significantly depending on the situation and different human interpretations. To address this, we ask the question: *How can we leverage VLMs to use human language reasoning to navigate social environments?*

We introduce Social robot Navigation via Explainable Interactions (SNEI), a human-annotated vision-language dataset aimed at improving socially aware robot navigation. SNEI contains over 40K manually annotated Visual Ques-

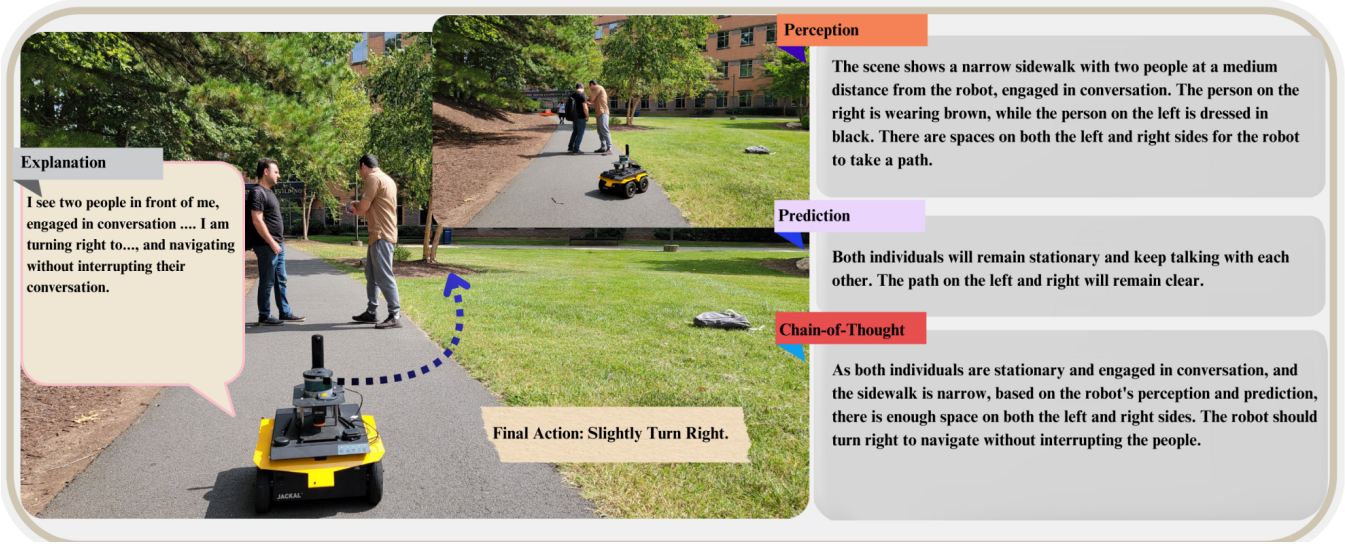


Fig. 2: Proof-of-concept real-world example of Social-LLaVA, enabling the robot to understand context and social cues to navigate, thereby avoiding interrupting people's conversations. There is enough space between the two people for the robot to pass, and it understands that they are talking, so it avoids disruption.

tion Answering (VQA) pairs spanning perception, prediction, chain-of-thought reasoning, action, and explanation, which enhance robot understanding and responses to their surroundings through natural language descriptions of social context, visual cues, and behavioral patterns. SNEI is based on 2K social scenarios, where a robot engages in direct interaction with humans, from SCAND [5], a social robot navigation demonstration dataset collected in diverse, human-crowded public spaces and containing complex human-robot interaction scenarios.

SNEI contains both categorical and free-form annotations, broadening its applicability to various tasks such as spatial reasoning, robot behavior explanation, and scene description. These capabilities enhance other robot navigation tasks such as instruction-following and object-goal navigation tasks. The categorical portion of the SNEI dataset can be used to evaluate Vision-Language Models (VLMs) across a range of tasks, including spatial reasoning, object detection, and human intention recognition.

Moreover, we fine-tune the LLaVA [13] Vision-Language Model, referred to as Social-LLaVA, on SNEI to demonstrate the effectiveness of the SNEI dataset. This adaptation is specifically tailored to our dataset, enabling the generation of high-level navigation action instructions through chain-of-thought reasoning in human language. To demonstrate the effectiveness of the SNEI dataset, we evaluate the performance of Social-LLaVA both quantitatively and qualitatively.

As shown in Table IV, our approach, Social-LLaVA outperforms both GPT-4V [14] and GeminiGemini 1.5 Pro [15] in overall performance. A small amount of hand-crafted human data can significantly improve performance [16], [17], as demonstrated by our quantitative analysis of over 400 VQA samples and qualitative analysis based on the average scores from fifteen different human judges across 50 VQA social navigation tasks.

## II. Related Work

### A. Social Navigation

Extensive research has been conducted in the field of social robot navigation, driven by the need for robots to operate safely and efficiently in human-populated environments [1], [18]–[20]. The complexity of social navigation arises from the necessity to consider a wide range of factors, such as safety, comfort, politeness, and adherence to unwritten social norms that humans instinctively follow [21]. Traditionally, model-based approaches that rely on task-specific, hand-engineered behaviors have been employed in social navigation. One of the earliest models is the Social Force Model (SFM) [3], [22], which simulates human navigation by modeling forces between individuals and obstacles. Another is human-robot proxemics [23], [24], which focuses on the spatial distances humans maintain around robots to ensure comfortable interactions.

To overcome the limitations of hand-engineered features and lack of adaptability, recent research has increasingly turned to learning-based methods, such as Learning from Demonstration (LfD) [25], [26]. These techniques allow robots to learn socially compliant behaviors by observing and replicating human demonstrations. Despite these advances in learning-based methods, simply replicating human trajectory from demonstrations in terms of perception-action pairs without human-like reasoning in between can be overly brittle and may not be sufficient to achieve socially aware navigation in a wide range scenarios.

In this work, we posit that human-like language reasoning and explanation between robot perception and action can facilitate socially compliant navigation behaviors. Therefore, we create SNEI dataset that incorporates human-like comprehension of social contexts in terms of language, including

the ability to perceive the current situation, predict the actions of other agents, and generate socially compliant navigation behaviors through chain-of-thought reasoning.

### B. Visual Instruction Tuning

Visual instruction tuning is a method of fine-tuning on an image-text dataset that trains the model to follow textual instructions with visual inputs and generate the desired outputs, thereby enhancing zero-shot performance on specific tasks. The success of multimodal models, such as InstructBLIP [27] and LLaVA [13], heavily depends on high-quality general visual instruction tuning datasets [28]. Several general-purpose datasets, mostly generated automatically or semi-automatically, have demonstrated improvement on models' performance [13], [29]. LLaVA [13] systematically constructed the LLaVAInstruct-150K dataset by prompting GPT-4 to generate questions and answers using image captions and object bounding boxes from the COCO [30] dataset. InstructBLIP [27] integrated VQA datasets for academic tasks related to visual comprehension. JRDB-Social [31] emphasize multi-person scene understanding and 2D/3D bounding-box annotations for general robotic perception in social contexts. In Autonomous Driving, DriveLM [9] constructed a graph VQA dataset for training VLMs for end-to-end driving. LingoQA [10] proposed a video QA dataset for autonomous vehicle explainability. However, many commonly used instruction-tuning datasets have been found to unexpectedly contain a considerable number of low-quality instances, featuring incorrect or irrelevant responses, potentially due to the (semi-)automatic nature of their collection methods. On the other hand, several works have shown that small, high-quality, human-curated datasets can boost model performance compared to large-scale noisy datasets [16], [17]. Based on these insights, and given that no existing general-purpose datasets adequately address the complexities of human-robot social navigation interactions (refer to our experimental results for details), we propose a novel visual instruction tuning dataset for social robot navigation.

## III. The SNEI Dataset

### A. Motivation

To leverage human-language reasoning, a robot first needs to transform its visual perception and the predictions derived from it into language. Then, using established chain-of-thought reasoning methods [32], it can produce high-level actions through human-like language reasoning [33]. Inspired by DriveLM [9], we hypothesize that mobile robots can leverage a decision-making process that humans implicitly perform, i.e., object-centric perception, prediction, and planning, in the format of language to describe each of these three stages [34].

Our preliminary experiments with off-the-shelf state-of-the-art VLMs reveal significant limitations in spatial reasoning, particularly in tasks critical to social robot navigation, such as determining the relative positions of humans, estimating their intent, and predicting their trajectories. As highlighted by SpatialVLM [11], these shortcomings in spatial reasoning capabilities of state-of-the-art VLMs are attributed more to the limitations of the common datasets used for training than to the models' architectures themselves. This gap can be more pronounced for tasks involving complex, dynamic environments with multiple interacting humans, requiring precise understanding of both spatial relationships and social cues. The majority of available VQA datasets for visual navigation are either general-purpose or task-specific, such as those for autonomous driving. However, due to the distinct nature of these tasks, such datasets are not directly applicable to mobile robot navigation in human-populated public spaces. Furthermore, the (semi-)automated methods used for their collection often result in a significant amount of noise. To this end, we propose a dataset of over 40K VQA instances annotated by humans for mobile robot navigation in unstructured, crowded environments. To the best of our knowledge, this is the first VQA dataset specifically designed for social robot navigation. In summary, SNEI is the first social robot navigation VQA dataset that targets unstructured human-populated environments, jointly focus on robot trajectories and human intentions, and offers a multi-task evaluation.

### B. Data Construction and Analyses

We provide two types of annotations: Categorical labels ensure consistency and structure across key elements like crowd density, agent types, and robot actions, while free-form natural language annotations offer greater expressiveness and nuanced descriptions.

*1) Categorical Labels:* All annotations of this type are selected from predefined categorical options to ensure consistent labeling across the dataset. We use the following categories:

- **Context** includes environment type, crowd density, indoor/outdoor status, and terrain type;
- **Robot** includes robot goals, movement directions, speed levels, and action intentions;
- **Obstacle(s)** includes type (e.g., walls and trash cans), proximity, and position relative to the robot; and
- **Agent(s)** includes type (e.g., individuals, groups, and bicycles), proximity, position relative to the robot, current action, and heading direction.

*2) Free-form natural language:* All annotations of this type take the form of natural language descriptions to cover the following aspects:

- **Perception** describes the robot's visual inputs, focusing on humans, including their clothing color, position, relative distance, action, and surrounding crowd density;
- **Prediction** assesses the potential future movement of agent(s) within the scene;
- **Chain-of-Thought Reasoning** given the perception and prediction, formulates a high-level natural language instruction (e.g., Given the close proximity of the person crossing the robot's path from left to right, the robot should stop, wait for the person to pass, and then continue);

| | Robot Action | Social Context | Agent Intention Recognition | CoT Reasoning | Manual Annotation |
|---|---|---|---|---|---|
| SACSON [26], MuSoHu [6], SCAND [5] | ✓ | ✗ | ✗ | ✗ | ✗ |
| JRDB-Social [31] | ✗ | ✓ | ✓ | ✗ | ✗ |
| **SNEI** | ✓ | ✓ | ✓ | ✓ | ✓ |

TABLE I: Comparison of SNEI with existing social navigation datasets.

- **Final Action** comprises high-level natural language action commands (e.g., Stop and wait for clear path); and
- **Explanation** includes a general explanation of what the robot sees, what the robot does, and why.

We use the SCAND dataset [5], which is collected from various human-crowded public environments and contains intricate human-robot interaction scenarios. We manually select and label 2K scenarios including both direct human-robot interactions and situations where the robot must interpret social cues to navigate in a socially compliant manner [1].

The SNEI dataset is annotated by human following a detailed and structured protocol to capture both categorical and free-form information about social navigation scenarios. Annotators are provided with a front-facing camera image from the robot's perspective and are tasked with filling out spatiotemporal information about the agents and objects in the scene. In the free-form annotations, annotators describe the scene, predict agent trajectories, reason through chain-of-thought (CoT) method, ensuring that nuanced, context-dependent information is preserved. The annotation guidelines emphasize human-awareness in reasoning, ensuring that the robot's decisions prioritize human comfort and safety. For example, if an agent is close, the annotator might label the final action as "Stop" and explain, "I am close to the person with a backpack, so I am stopping to yield and avoid crossing their path." Given that social navigation can be subject to multiple interpretations and variations, ensuring consistency is crucial [1]. To maintain uniformity and reduce variability in interpretation, all annotations are reviewed and verified by the same person, ensuring high-quality and consistent labeling throughout the dataset.

Table II shows the number of images for each type of interaction in the SNEI dataset. As indicated by the numbers, the majority of the dataset consists of images in very crowded human environments.

TABLE II: The number of images for each interaction type in SNEI.

| Interaction Type | Description | # Images |
|---|---|---|
| Against Traffic | Navigating against oncoming traffic | 161 |
| With Traffic | Navigating with oncoming traffic | 358 |
| Street Crossing | Crossing across a street | 75 |
| Overtaking | Overtaking a person or groups of people | 20 |
| Narrow Doorway | Navigating through a doorway where the human opens or waits for others to open the door | 31 |
| Crossing Stationary Queue | Walking across a line of people | 25 |
| Stairs | Walking up and/or down the stairs | 30 |
| Vehicle Interaction | Navigating around a vehicle | 28 |
| Navigating Through Large Crowds | Navigating among large unstructured crowds | 1233 |

## IV. Evaluation

In this section, we show that our dataset can enable human-level language reasoning when facing social navigation interactions in human-crowded spaces.

### A. Social-LLaVA

Using the SNEI dataset, we develop Social-LLaVA, a VLM designed to exhibit human-like reasoning in social navigation scenarios. The primary contribution of this work is the introduction of a dataset tailored for mobile robot navigation in human-crowded environments. We emphasize that our focus lies in dataset development rather than algorithmic innovation. To demonstrate the practical utility of the SNEI dataset, we fine-tune an off-the-shelf, state-of-the-art model as a proof of concept.

To test Social-LLaVA model fine-tuned on SNEI dataset, a large-scale VQA benchmark with reliable performance is required. However, to the best of our knowledge, SNEI is the first human-annotated VQA benchmark designed for robot social navigation, featuring both categorical and free-form labels. We evaluate the effectiveness of SNEI in enhancing performance on the general social navigation task, through both quantitative and qualitative methods. On the test subset of the dataset, we quantitatively assess the categorical labels by calculating the average similarity scores across different categories and comparing them against state-of-the-art models. For the free-form portion of the dataset, which requires human-level reasoning, contextual understanding, and interpretation of social cues, we rely on expert evaluations. Specifically, we ask human experts in the field to rate the model outputs to ensure a fair and comprehensive assessment.

We evaluate the dataset's effectiveness by fine-tuning LLaVA-v1.5-7B [13], a leading VLM, using Low-rank Adaptation (LoRA) [35]. The resulting model, Social-LLaVA, exhibits improved social reasoning capabilities in navigation interactions, as evidenced by both quantitative evaluations and qualitative assessments from human experts.

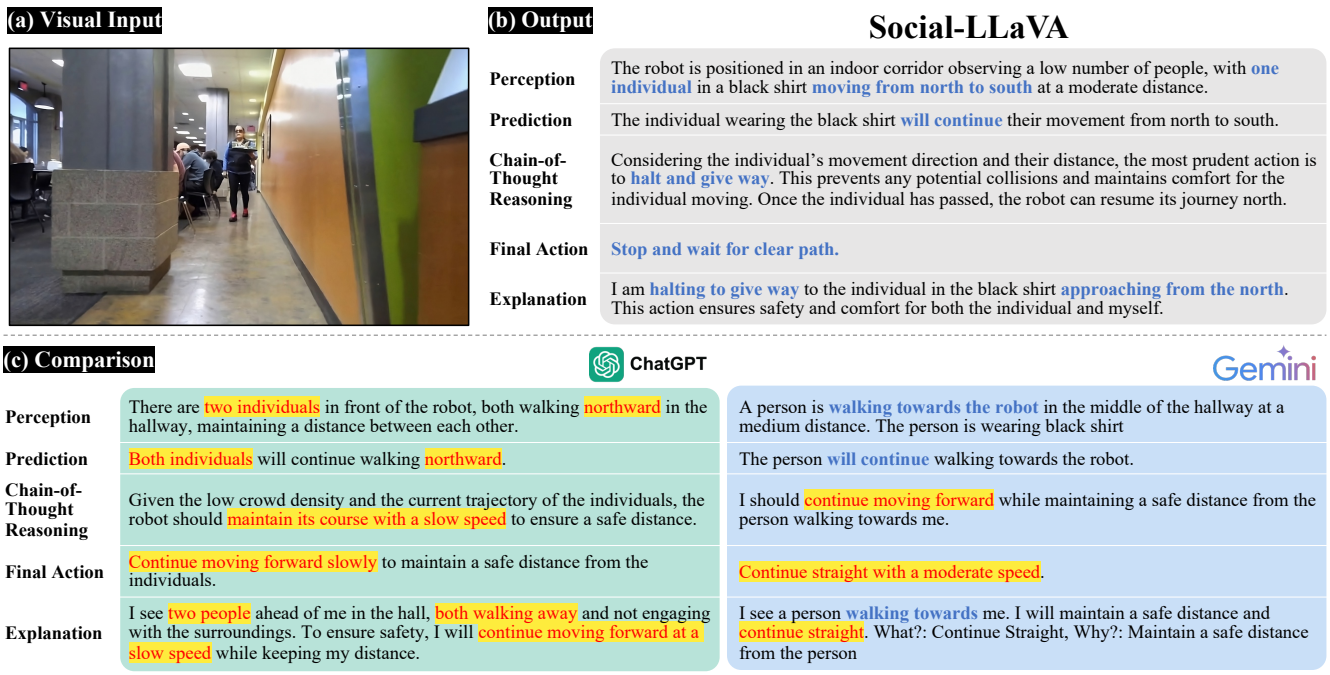


Fig. 3: Qualitative results of our Social-LLaVA model fine-tuned on SNEI dataset compared against GPT4-V [14] and Gemini 1.5 Pro [15]. (a) shows the visual input given to the models. Note that the given scenario involves a robot navigating through a narrow passage partially obstructed by a pillar, while an individual approaches the robot. (b) illustrates the output from Social-LLaVA, while (c) provides comparisons with GPT4-V and Gemini 1.5 Pro. Phrases in **blue** indicate accurate reasoning and socially compliant results, while highlighted phrases mark instances of hallucination.

### *B. Quantitative Evaluation*

To evaluate the categorical output of Social-LLaVA, we assess the predicted agents in the scene (e.g., individuals, groups, and bicycles). Agent-level similarity is computed as the average of per-attribute matches, which include proximity, position relative to the robot, current action (e.g., walking and standing), and heading direction. Specifically, for each predicted agent $j$ and ground truth agent $i$, where $i, j \in \{1, 2, \ldots, N\}$ and $N \leq 5$ (max in dataset), we compare their attributes and calculate an individual similarity score. This process results in a similarity matrix representing the correspondence between all predicted and ground truth agents. To find the optimal assignment between agents, we apply the Hungarian algorithm to minimize the overall cost, defined as the inverse of similarity (cost $= 1 -$ similarity). This ensures each predicted agent is matched with the most appropriate ground truth agent. For each matched pair, we further evaluate two sub-groups of attributes: **relative position to the robot** (degree and clock positions) and **intention** (action type and relative direction), computed as the average similarity of their respective attributes. Finally, for each image, we report three key metrics: (1) **average agent similarity**, calculated from matched pairs; (2) **average relative position similarity**; and (3) **average intention similarity**, providing a comprehensive evaluation of both detection accuracy and attribute-level consistency. Given a similarity matrix $C = [c_{ij}]$ of size $n \times n$, the goal is to find an assignment that maximizes the similarity:

$$\begin{aligned} &1 - \min \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij}, \\ &\text{s.t.} \sum_{j} x_{ij} = 1, \ \sum_{i} x_{ij} = 1, \ x_{ij} \in \{0, 1\}. \end{aligned} \tag{1}$$

Table III compares Social-LLaVA's performance with SoTA VLMs on categorical labels from SNEI, focusing on agent intention prediction and relative position to the robot. Since the original LLaVA 1.5 required fine-tuning to follow prompts, we excluded it from the comparison. Prior research [12] has shown LLaVA underperforms compared to GPT-4V.

| Metric | GPT-4V | Gemini 1.5 Pro | Social-LLaVA |
|---|---|---|---|
| Relative Position | 0.42 | **0.43** | 0.36 |
| Predicted Intention | 0.21 | 0.22 | **0.26** |
| Overall | 0.30 | **0.32** | 0.29 |

TABLE III: A quantitative comparative analysis of the performance of Social-LLaVA, GPT-4V, and Gemini 1.5 Pro on the categorical section of the SNEI. The reported scores represent the average similarity between each model's predictions and the ground truth.

### *C. Qualitative VQA Evaluation*

We evaluate the free-form natural language outputs generated by Social-LLaVA, compared against two other state-of-the-art foundation models. We prompt Gemini 1.5 Pro [15]

| | GPT4-V | Gemini 1.5 Pro | Social-LLaVA |
|---|---|---|---|
| Perception | 3.11 | 3.45 | **4.0** |
| Prediction | 3.18 | 3.87 | **4.06** |
| CoT | 3.41 | 3.79 | **4.08** |
| Final Action | 2.77 | 3.46 | **4.19** |
| Explanation | 3.16 | 3.66 | **3.95** |

TABLE IV: A qualitative comparative analysis of the performance of Social-LLaVA, GPT-4V, and Gemini 1.5 Pro on the free-form section of the SNEI, based on average scores per task as evaluated by fifteen human judges. These scores range from 1 to 5.

and GPT4-V [14] with extensive explanations of the task. The human rater chooses a score between 1-5 for each VQA, and finally, we average over each task for each model. Table IV shows the achieved scores from the three models for each task. The results demonstrate that Social-LLaVA achieves higher scores compared to GPT4-V and Gemini. Figure 3 shows an example where a robot navigates a narrow passage partially blocked by a pillar while a person approaches. Social-LLaVA accurately perceives and predicts the scenario, offering socially compliant actions through effective reasoning. In contrast, GPT-4V and Gemini 1.5 Pro mistakenly advise continuing forward, which would block the individual's path. Notably, GPT-4V also produces hallucinated outputs across perception, prediction, reasoning, action, and explanation. This experiment underscores the importance of high-quality VQA data for mobile robot navigation, as it is essential for scene understanding, high-level human trajectory prediction, and chain-of-thought reasoning, which are missing from the pretraining datasets of current state-of-the-art VLMs. We speculate that both aforementioned models would demonstrate significantly improved few-shot performance if fine-tuned on our dataset. However, we cannot validate this speculation, as image fine-tuning is currently unavailable for these models.

### D. Real World Robot Experiment

While SNEI dataset with the preliminary proof of concept, Social-LLaVA model, shows potential for enabling socially compliant robot navigation behaviors through explainable interactions using language, producing concrete robot actions based on language descriptions remains an open problem. In this work, we present a simple proof-of-concept using a hard-coded relationship between high-level language descriptions and low-level robot actions (go straight, turn left, and turn right). In Fig. 2, we showcase example of Social-LLaVA's output after training on our SNEI dataset in a real-world setting, where it understands the context, interprets social cues, and avoid interrupting people's conversations.

## V. Limitations

### A. Ambiguity in Social Navigation

The complexity and unpredictability of human motion, along with inconsistent behaviors across different individuals, further complicate accurate predictions and other downstream stages. Moreover, the absence of a clear definition of social navigation [1] adds to the uncertainty. All these factors make it challenging to generate accurate descriptions and robust predictions for effective decision-making. These challenges also further complicate SNEI data annotation effort, considering that there may be multiple ways or there may not be an agreed-upon way of socially compliant behavior. With these challenges in mind, we aim to generate the most accurate language descriptions possible from a single image to reason based on them and generate high-level actions. While Social-LLaVa serves as a proof-of-concept for using language descriptions as a tool for social robot navigation reasoning, particularly when more than simple obstacle avoidance is required, how to address such ambiguity when annotating data and during training remains open problem. We acknowledge that both categorical and free-form evaluations of Social-LLaVA can be affected by variations in the ground truth label and may differ based on the labeler.

### B. Need for More and Diverse Data

While several research efforts show that with a small amount of high-quality data it is possible to fine-tune VLMs, the data-driven nature of these models should not be overlooked [16]. We diversify the interactions in our SNEI dataset as much as possible. However, it remains limited to the scenarios present in the source dataset, SCAND, which may not capture the full spectrum of human-robot interactions in unstructured environments. Consequently, this limitation could affect the model's performance when deployed in previously unseen environments.

### C. Grounding Language to Action

Our work primarily focuses on havinge high-level human-language instructions that promote socially compliant behaviors in various scenarios—such as waiting in line or using off-road paths to avoid interrupting conversations—rather than grounding those instructions. We acknowledge that grounding the generated natural-language high-level actions on the robot in real-world scenarios is still an open problem and is not a trivial task.

## VI. Conclusions

In this work, we introduce Social robot Navigation via Explainable Interactions (SNEI), a vision-language dataset specifically designed to bridge the gap between perception and socially compliant actions in crowded public environments through human-like language-based reasoning. We demonstrate that Social-LLaVA, fine-tuned version of LLaVA on SNEI, outperforms state-of-the-art systems such as GPT-4V and Gemini in generating socially compliant navigation descriptions. By leveraging language-based reasoning using SNEI , our preliminary proof of concept demonstration shows that robots can understand social interactions and navigate unstructured social spaces, while observing human social norms. Our results mark a promising step toward

more intuitive and effective social robot navigation in real-world public spaces through explainable human-language reasoning.

## VII. Acknowledgement

This work has taken place in the RobotiXX Laboratory at George Mason University. RobotiXX research is supported by National Science Foundation (NSF, 2350352), Army Research Office (ARO, W911NF2320004, W911NF2420027, W911NF2520011), Air Force Research Laboratory (AFRL), US Air Forces Central (AFCENT), Google DeepMind (GDM), Clearpath Robotics, Raytheon Technologies (RTX), Tangenta, Mason Innovation Exchange (MIX), and Walmart.

## References

[1] 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," *ACM Transactions on Human-Robot Interaction*, 2024.

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

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

[4] G. Ferrer, A. Garrell, and A. Sanfeliu, "Social-aware robot navigation in urban environments," in *2013 European Conference on Mobile Robots*. IEEE, 2013, pp. 331–336.

[5] 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.

[6] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao, "Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 7442–7447.

[7] P. Bloom and F. C. Keil, "Thinking through language." *Mind & language*, vol. 16, no. 4, 2001.

[8] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," 2023. [Online]. Available: https://arxiv.org/abs/2309.17080

[9] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," 2024. [Online]. Available: https://arxiv.org/abs/2312.14150

[10] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton, E. Arani, and O. Sinavski, "Lingoqa: Video question answering for autonomous driving," 2024. [Online]. Available: https://arxiv.org/abs/2312.14115

[11] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 14 455–14 465.

[12] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.01584

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

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

[15] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[16] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "Lima: Less is more for alignment," 2023. [Online]. Available: https://arxiv.org/abs/2305.11206

[17] L. Wei, Z. Jiang, W. Huang, and L. Sun, "Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4," 2023. [Online]. Available: https://arxiv.org/abs/2308.12067

[18] 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.

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

[20] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, vol. 10, no. 1, pp. 508–515, 2025.

[21] 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.

[22] G. Ferrer, A. Garrell, and A. Sanfeliu, "Robot companion: A social-force based approach with human awareness-navigation in crowded environments," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1688–1694.

[23] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 331–338.

[24] K. Charalampous, I. Kostavelis, and A. Gasteratos, "Robot navigation in large-scale social maps: An action recognition approach," *Expert Systems with Applications*, vol. 66, pp. 261–273, 2016.

[25] 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.

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

[27] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2305.06500

[28] B. Zhao, B. Wu, M. He, and T. Huang, "Svit: Scaling up visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2307.04087

[29] J. Huang, J. Zhang, K. Jiang, H. Qiu, and S. Lu, "Visual instruction tuning towards general-purpose multimodal model: A survey," *arXiv preprint arXiv:2312.16602*, 2023.

[30] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[31] S. Jahangard, Z. Cai, S. Wen, and H. Rezatofighi, "Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups," 2024. [Online]. Available: https://arxiv.org/abs/2404.04458

[32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[33] J. A. Groeger, *Understanding driving: Applying cognitive psychology to a complex everyday task*. Routledge, 2013.

[34] D. Marr, *Vision : a computational investigation into the human representation and processing of visual information*. Cambridge, Mass: The MIT Press, Jul. 2010.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685