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.

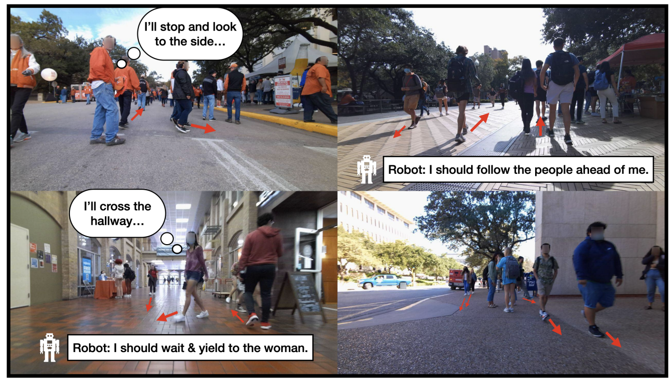

Fig. 1: Examples of social robot navigation scenarios from SCAND [8] where humans in the scene have to be taken into consideration. The ability to determine socially compliant navigation actions requires understanding each dynamic scene by spatiotemporal reasoning (e.g. the movements of people in the scene), social reasoning (inferring the navigation intentions of people in the scene), and complying to implicit social rules.

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 [12]. As shown in Figure 1, navigating through social scenarios requires robots to interpret human intentions, adhere to implicit social rules, and respond to dynamic environments demanding advanced spatial, spatiotemporal, and social reasoning. While traditional methods often relied on model-based techniques like the Social Force Model [5] or proxemics-based methods [13], recent learning-based approaches, including Learning from Demonstration [6, 8] and Reinforcement Learning [28, 3, 11], have shown promise but frequently struggle to generalize effectively in real-world scenarios [12, 18]. To address these limitations, diverse datasets like SCAND [8, 15, 7] have been introduced, providing more realistic social navigation contexts.

Recently, there has been growing interest in applying large Vision-Language Models (VLMs) to robotic tasks due to their demonstrated strengths in contextual understanding, common-sense reasoning, and chain-of-thought reasoning [10, 16, 24]. VLMs have been effectively applied to robotic manipulation [14], task planning [27], and human-robot interaction [1, 4]. Initial explorations, such as VLM-Social-Nav [22], suggest their potential in social robot navigation; however, these evaluations remain limited to controlled scenarios and preliminary assessments. Moreover, state-of-the-art large VLMs still face substantial challenges in robust spatial reasoning, casting doubt on their capabilities to navigate complex, realistic social situations [20, 2, 23].

Given these gaps, a comprehensive evaluation framework is essential to rigorously assess VLM capabilities in three critical dimensions of social robot navigation: spatial reasoning, spatiotemporal reasoning, and social interaction interpretation. Existing evaluations often neglect the complexity of dynamic scenarios or lack temporal components [22, 20]. Addressing this need, we introduce the Social Navigation Scene Understanding Benchmark (SocialNav-SUB), a novel Visual Question Answering (VQA) benchmark utilizing social navigation scenarios from the SCAND dataset [8, 9]. Our benchmark includes robust human-subject evaluations serving as ground truth, enabling systematic assessment of VLM performance on realistic social navigation tasks. Experiments conducted reveal notable performance gaps between state-of-the-art VLMs and both human and rule-based baselines, especially in spatial and spatiotemporal reasoning.

[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.

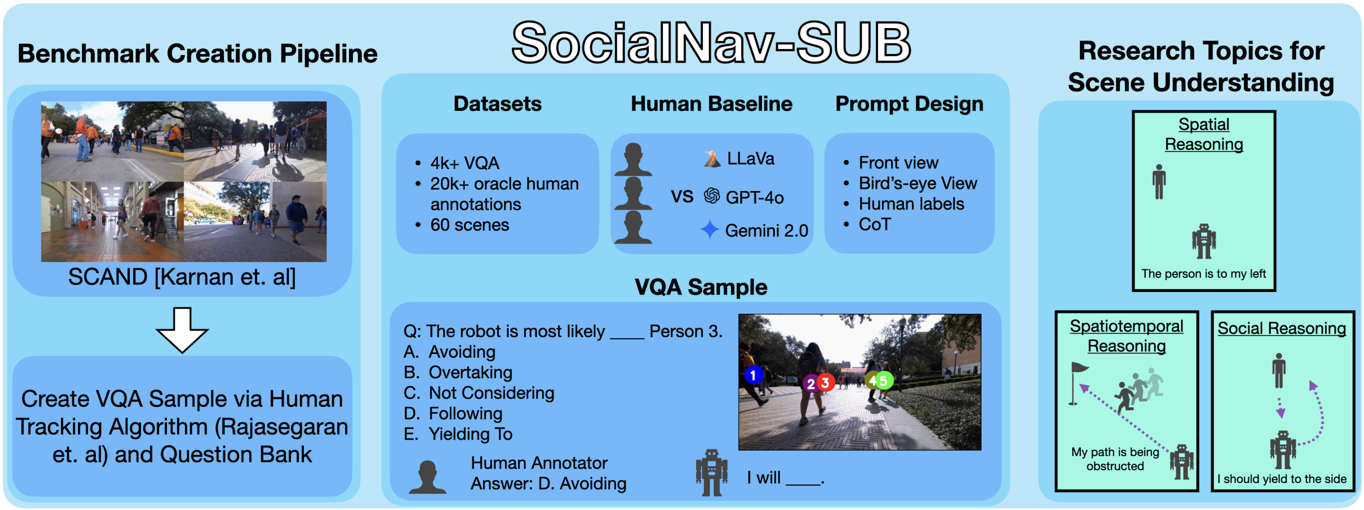

Fig. 2: An overview of SOCIALNAV-SUB, which facilitates the systematic evaluation of VLMs in social robot navigation scenarios. Using SCAND data, human-labeled VQA datasets, and various VLMs, this framework offers the evaluation of VLMs across multiple dimensions of scene understanding for social robot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

SOCIALNAV-SUB is a first-of-its-kind benchmark specifically designed to systematically evaluate and refine VLMs for real-world social robot navigation, providing a clear framework for comparing models and identifying key strengths and weaknesses. SOCIALNAV-SUB establishes a foundation for continuous and iterative advancements, guiding the robotics community toward developing more socially intelligent robotic systems capable of effectively navigating complex, dynamic human-centered environments. By bridging the gap between VLM capabilities and the challenges of social robot navigation, our work provides a medium to advance the use of VLMs for social robot navigation.

II. 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 [14, 26, 25], 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.

A. Challenging Social Navigation Scenarios

To effectively evaluate VLMs' scene understanding capabilities in practical social robot navigation settings, we leverage the SCAND dataset [8] 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 from SCAND are RGB image sequences captured from a robot-mounted front-view camera; however, large VLMs often struggle to infer spatial and fine-grained object-level relationships from these visual queries alone [20]. To address this, we augment images with object-centric annotations using off-the-shelf vision models, a strategy shown to enhance VLM performance in VQA tasks [14, 26]. Specifically, pedestrians in the front-view images are annotated with numbered, color-coded circles, and additional BEV images illustrating pedestrian and robot positions are generated, preserving scene context while clearly representing spatial relationships such as distances and obstructed paths. Our data processing pipeline employs PHALP [19] to estimate pedestrian 3D poses from monocular videos, transforms these poses into global coordinates using robot odometry data, applies Kalman smoothing, and finally projects pedestrian positions onto front-view and BEV images using SCAND's camera parameters. Querying VLMs with these enriched visual inputs allows SOCIALNAV-SUB to yield practical insights into effectively leveraging VLMs for social robot navigation, while ensuring fairness by providing identical inputs to human annotators.

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 s segment sampled at 4 Hz. To comprehensively evaluate VLMs' scene understanding capabilities in social robot navigation, we design a range of multiple-choice questions.

- **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 [17] 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. We define a metric, **Probability of Agreement (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 $i \in \{1, \dots, N_H\}$.

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

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

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.

III. PRELIMINARY EXPERIMENT

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 **??**. 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	0.83 ± 0.00	0.80 ± 0.01	0.82 ± 0.01	0.85 ± 0.01
	Rule-Based	0.69 ± 0.00	0.61 ± 0.01	0.67 ± 0.01	0.73 ± 0.01
VLM	Gemini 2.0	**0.62 ± 0.01**	0.58 ± 0.01	0.46 ± 0.01	**0.68 ± 0.01**
	GPT-4o	0.51 ± 0.01	**0.58 ± 0.01**	0.51 ± 0.02	0.48 ± 0.01
	LLaVa-Next-Video	0.48 ± 0.01	0.34 ± 0.01	**0.62 ± 0.01**	0.50 ± 0.01

TABLE I: 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. We separate rows into two broad categories (Baseline and VLM). All VLMs use chain-of-thought reasoning, since each human provided answers sequentially.

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 [21]) 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 I, 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.

IV. Conclusion

This paper introduced the Social Navigation Scene Understanding Benchmark (SocialNav-SUB), a novel VQA benchmark leveraging densely populated, dynamic environments from SCAND to provide object centric visual inputs, including augmented front view images and BEV prompts, paired with diverse questions targeting spatial, spatiotemporal, and social reasoning. Grounded in robust human subject evaluations, SocialNav-SUB offers quantifiable metrics reflective of human-level understanding in social robot navigation contexts. By highlighting current VLMs' strengths and weaknesses, SocialNav-SUB facilitates systematic model comparisons, enables exploration of prompt design choices, fosters method development, and guides iterative improvements toward more socially aware and reliable robotic systems for social robot navigation.

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] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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