

SocRATES: Towards Automated Scenario-based Testing of Social Navigation Algorithms

Shashank Rao Marpally¹, Pranav Goyal^{1,2}, and Harold Soh^{1,2}

¹Dept. of Computer Science, National University of Singapore.

²Smart Systems Institute, NUS.

Email: {smarpall@comp.nus.edu.sg, prgoyal@umich.edu, harold@comp.nus.edu.sg}

Abstract— Current social navigation methods and benchmarks primarily focus on proxemics and task efficiency. While these factors are important, qualitative aspects — such as perceptions of a robot’s social competence — are equally crucial for successful adoption and integration into human environments. We propose a more comprehensive evaluation of social navigation through scenario-based testing, where specific human-robot interaction scenarios can reveal key robot behaviors. However, creating such scenarios is often labor-intensive and complex. In this work, we address this challenge by introducing a pipeline that automates the generation of context-, and location-appropriate social navigation scenarios, ready for simulation. Our pipeline transforms simple scenario metadata into detailed textual scenarios, infers pedestrian and robot trajectories, and simulates pedestrian behaviors, which enables more controlled evaluation. We leverage the social reasoning and code-generation capabilities of Large Language Models (LLMs) to streamline scenario generation and translation. Our experiments show that our pipeline produces realistic scenarios and significantly improves scenario translation over naive LLM prompting. Additionally, we present initial feedback from a usability study with social navigation experts and a case-study demonstrating a scenario-based evaluation of three navigation algorithms.

I. INTRODUCTION

In social navigation, robots must navigate through dynamic human environments while adhering to social norms. This presents a dual challenge: ensuring both competent navigation and socially appropriate behavior. Evaluating these aspects is difficult, as social appropriateness is subjective and context-dependent, often requiring human judgment. Current evaluation methods tend to focus on quantifiable metrics like safety and comfort, but these fail to capture the full spectrum of social interactions. If we could automate the generation of socially relevant scenarios, grounded in both location and task, we would significantly enhance the ability to test and refine social navigation algorithms. This would lead to more reliable and robust robots that can function effectively in real-world human environments.

In this work, we propose SocRATES (**Social Robot Assessment Through Scenario Evaluation**) (Fig. 1), a system designed to automate the generation of context-specific scenarios for evaluating social navigation algorithms. The key insight behind SocRATES is the use of large vision-language models (VLMs) to translate high-level, potentially ambiguous inputs into detailed simulation scenarios. By leveraging the commonsense reasoning and context-understanding

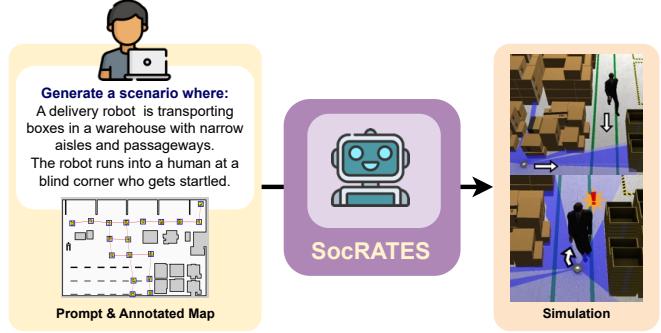


Fig. 1: We propose SocRATES, an automated system that leverages VLMs to generate simulated social navigation scenarios from simple textual and image inputs.

capabilities of these models, SocRATES generates realistic and varied scenarios that capture both the physical layout of the environment and the social interactions that occur within it. We believe this combination of AI-driven scenario generation with simulation makes SocRATES a significant first-step towards testing both the navigation and social appropriateness of robots.

Our work builds on prior research in scenario-based testing for autonomous systems and social navigation. Tools like SEAN 2.0 [1] and HuNavSim [2] have explored human behavior simulation in social settings, but they remain limited by a narrow set of predefined scenarios and/or lack flexibility or control over scenario generation. In contrast, SocRATES allows users to define scenarios based on textual and image-based inputs, providing much greater control over the evaluation process. Additionally, while prior tools focus primarily on the fidelity of human simulation, SocRATES is designed to evaluate the robot’s behavior in relation to both the task and social context. To our knowledge, SocRATES is the first system to integrate LLM-driven scenario generation with simulation for the comprehensive testing of social navigation algorithms.

Our experimental results demonstrate the effectiveness of SocRATES in generating diverse, contextually relevant scenarios. We conducted a design analysis that shows the system is fast and cost-effective, producing scenarios in under a minute with minimal cost. A user study with researchers in social navigation validated the system’s usability, with all participants preferring SocRATES over manual scenario

generation. Ablation studies further highlighted the importance of structured prompts and error-handling mechanisms in ensuring the quality of generated scenarios. Finally, we established a human study benchmark to compare multiple navigation algorithms, showcasing the practical utility of SocRATES as an evaluation tool.

To summarize, the key contributions of this work are:

- A novel system for automating the generation of social navigation scenarios using LLMs and VLMs. SocRATES uses commonsense reasoning and contextual understanding to create realistic, task-relevant scenarios.
- A flexible framework that allows users to define scenarios based on high-level inputs, providing control over evaluation scenarios.
- Experimental validation through user studies and benchmarks, demonstrating the system's utility and efficiency.

By enabling automated, scalable testing of both navigation and social appropriateness, SocRATES moves the field closer to developing robots that can reliably and effectively navigate human environments.

II. BACKGROUND

A. Evaluating Social Navigation

Evaluating social navigation is complex because it involves two interlinked components: social appropriateness and competent navigation. Comprehensive evaluation requires assessing both, yet the subjective nature of social appropriateness makes it particularly challenging [3]. For example, [4] outlines seven key principles that robots should follow to be considered socially appropriate. Most current evaluation methods primarily test two of these principles: safety (P1) and comfort (P2), often using proxemics-based metrics that measure the robot's and human's relative positions and speeds [5]. However, principles like contextual appropriateness (P8) depend on the robot's social context and task, which are often overlooked due to their qualitative nature.

To address these challenges, recent approaches have emerged in two main directions. The first uses proxy metrics that model specific aspects of a robot's social awareness. For instance, [6] introduces personal and group space intrusion metrics, which are subsequently adopted in other algorithms [7] and evaluation frameworks [2]. The second approach involves using human studies to assess social behaviors. For example, legibility has been modeled [8] and used to improve social navigation. Metrics like Envelope of Readiness, Clarity, and Moments of Confusion [9] rely on continuous observer data to quantify legibility. Additionally, prior studies [10] collect qualitative perceptions of social navigation using validated scales, e.g., PSI [11].

However, collecting scalable and informative human study data remains challenging, and existing benchmarks provide limited coverage of edge cases and special scenarios [4]. Crafting specific simulations for testing edge cases is cumbersome, contributing to the difficulty in gathering useful human evaluation data. These limitations highlight the need for

more comprehensive scenario-based testing, where specific behavioral aspects of the robot can be evaluated in controlled settings. Our work addresses this gap by automating the proposal and generation of context- and location-specific scenarios from simple textual and image-based inputs.

B. Scenario-based Testing for Social Navigation

Automated scenario generation for testing has a long history in Advanced Driver Assistance Systems (ADAS) [12]–[14], with recent efforts extending into Human-Robot Interaction (HRI) policy evaluation [15], [16]. However, most current social navigation evaluation tools, such as HuNavSim [2], Arena [17], and SocNavBench [18], focus primarily on the fidelity of human simulation rather than generating specific scenarios that test a robot's navigation and reactive capabilities.

SEAN 2.0 [1] is one exception that seeks to expose the robot to various social situations. However, SEAN 2.0 models only a limited set of scenarios and environments, offering little control over the scenario itself, which increases the uncertainty of whether a specific user-defined scenario will occur during evaluation. To the best of our knowledge, none of the current tools provide comprehensive simulation of human-robot interactions.

In contrast, SocRATES leverages the rich context and commonsense reasoning capabilities of large language models (LLMs), along with a flexible scenario and map parameterization, to generate a diverse set of scenarios, including human-robot interaction scenarios, in any user-defined location. This approach enables more comprehensive testing by allowing the user to specify detailed parameters, ensuring that specific scenarios of interest can be generated and evaluated.

III. METHODOLOGY

Our objective is to develop an easy-to-use and reliable system for automating scenario proposal and generation in social navigation testing. Users provide simple textual and image-based inputs to guide the scenario proposal, which our system then converts into a simulation. Once the simulation is generated, any robot-planner combination based on the ROS2 Navigation stack¹ can be used for evaluation.

Figure 2 provides an overview of our framework. We structure our system by deconstructing a scenario into five main components: (1) the *location* where the scenario takes place (Sec. III-A), (2) a detailed *description* of the scenario (Sec. III-B), (3) the *paths* for pedestrians and the robot (Sec. III-C), (4) the *behavior* of the pedestrians (Sec. III-D), and (5) the *simulation* that realizes these components (Sec. III-E). Each module of SocRATES interactively queries a Vision-Language Model (VLM) to generate elements of the scenario content. The simulation module then executes the scenario using these generated components and the specified navigation algorithm on the robot.

¹<https://docs.nav2.org>

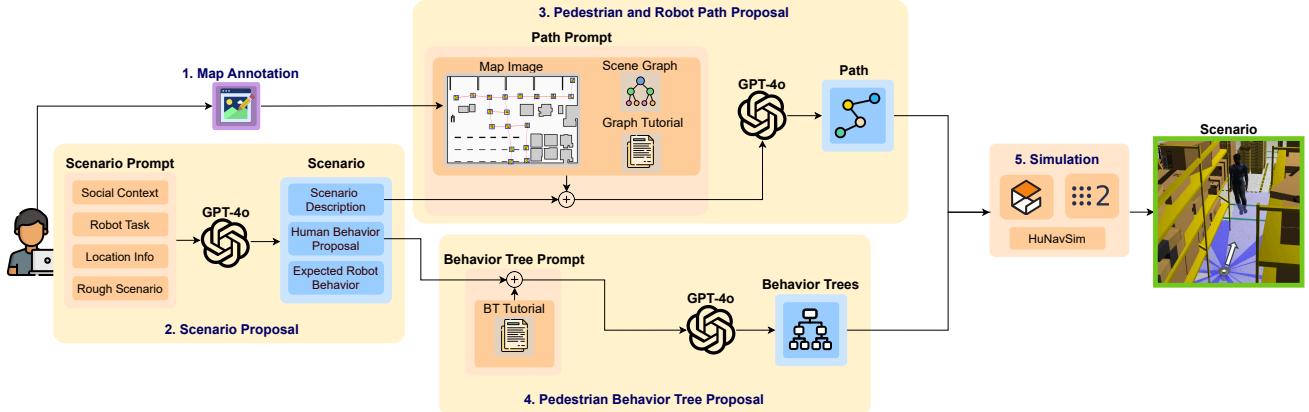


Fig. 2: Overview of our pipeline. We prompt users to annotate the map of their desired location (1) and provide simple textual inputs for their desired scenario. Our pipeline proposes a scenario (2) and then generates the 2 main components of the scenario with structured prompts to an LLM: The paths of the robot and pedestrians (3) and the behavior of the humans (4). Finally these are used by the HuNavSim [2] framework (5) to generate a simulation of the scenario.

A. Map Annotation

SocRATES includes a simple map annotation tool that allows users to provide the necessary contextual information about their desired location. This enables users to generate scenarios in custom simulated environments.

We represent a location as a 2D semantic scene graph, which contains both semantic and spatial information. An example scene graph overlaid on the map image for the Small Amazon Warehouse Gazebo world is shown in Fig. 2. Our annotation tool guides users in annotating a location map by adding nodes and edges. The schema for this graph can be defined by the user, with the requirement being that the node and edge types are self-explanatory. For example, edges are associated with semantic types (e.g., ‘intersection’ and ‘hallway’).

The scene graph serves two main purposes: (a) it provides both semantic and spatial context regarding different areas in the map and how they are connected, and (b) the structure of the scene graph helps to identify and validate paths for pedestrians and the robot, as explained further in Sec. III-C.

B. Scenario Proposal

The scenario proposal module is responsible for translating the user’s potentially ambiguous textual inputs into a detailed and precise characterization of the scenario. Inspired by the concept of scenario-cards [4], we ask users to input *scenario metadata*: (a) the *social context* in which the robot operates (e.g., “A quiet old-age home”), (b) the robot’s intended *task* (e.g., “Deliver coffee from the kitchen to the Hall”), (c) an optional *rough scenario* where the user can describe more scenario details to constrain the system’s outputs, and (d) a description of the *location*.

We generate a detailed scenario description grounded in the user’s provided location by prompting a Large Language Model (LLM) with the scenario metadata, context about the social navigation task, and the capabilities of the pedestrians in the simulator (e.g., *they cannot manipulate objects, only navigate*).

From the Vision-Language Model (VLM), we extract the following outputs: a precise *scenario description* to condition path and behavior generation, a description of the *behaviors* of humans in the scenario, and the *expected behavior of the robot* when navigating the scenario, which serves as ground truth for evaluation. We observe that providing example inputs and corresponding scenarios in the prompt significantly improves the quality of the LLM’s responses. Therefore, we include a set of handcrafted examples in the prompt to guide the model’s output.

C. Pedestrian and Robot Path Generation

This module generates paths for both pedestrians (with group assignments, if applicable) and the robot, orchestrating their movements to align with the scenario. To ensure consistency with the user’s provided location, the path generation is conditioned on the scene graph from Sec. III-A. Human and robot paths are represented as trajectories on this graph. The VLM is queried with a structured prompt, which includes: the scene graph and the scene-graph annotated map image, pedagogical examples on interpreting scene graphs and path generation, and the scenario description (from Sec. III-B).

We extract paths for each pedestrian and the robot simultaneously, along with group assignments. We identify specific nodes where the robot encounters each pedestrian, which is useful for timing pedestrian motion in the simulation to ensure the scenario unfolds as intended. A common error by the VLM is the assignment of discontinuous paths on the scene graph. We detect such errors and re-query the LLM when necessary. Despite structured prompts, the VLM may still generate a valid but unsatisfactory path. To address this, this module is interactive, allowing the user to accept, reject, or edit paths using natural language commands (e.g., “Make the robot’s path longer”), which re-queries the LLM for updates.

D. Pedestrian Behavior Generation

The behavior generation module translates the behavior descriptions from the scenario proposal into encodings compatible with the simulator. We model pedestrians using HuNavSim [2], which supports rich, programmable reactive behaviors through behavior trees (BTs). For each pedestrian, we construct an LLM prompt that includes the *BT node library* (a list of available behavior tree nodes), the required human behaviors (from Sec. III-B), and pedagogical examples on BT syntax and design rules.

The LLM outputs a behavior tree in XML for each pedestrian, which can be directly imported into HuNavSim to control their actions. However, the default behavior trees in HuNavSim do not support interaction between humans and the robot or complex behaviors. To enable interactive scenarios, we implemented additional behaviors allowing humans to gesture towards other agents, as well as recognize gestures. For example, a simulated human can wait until the robot makes a specific gesture or until another pedestrian performs an action.

E. Simulation

This module integrates the generated behaviors and trajectories into a simulation environment. The scene graph paths for each pedestrian are transformed into simulator world frame paths using the map parameters and are used as navigation goals for pedestrians and waypoints for the robot planner. Running the simulation through HuNavSim using ROS2 creates a Gazebo instance where the pedestrians follow the specified paths and act according to the behavior trees.

For interactions (e.g., gesturing), pedestrians and the robot publish integer-coded gestures to specific ROS2 topics. In our experiments, we found that timing the arrival of pedestrians and the robot at their interaction points is critical for orchestrating the scenario. Since the robot’s motion is fully controlled by its onboard navigation algorithm, we introduce a *scenario manager* ROS2 node that synchronizes pedestrian movements with the robot’s location, ensuring their timely arrival at interaction points, thus increasing the likelihood that the desired scenario plays out correctly.

IV. EXPERIMENTS AND RESULTS

We evaluated the effectiveness of SocRATES, by analyzing its reliability, cost, and design choices. We evaluated its utility as a tool through a usability study with academic researchers and separately with a persona-based assessment.

A. Design Analysis

We assessed our system for *reliability, cost* and *speed*. In addition, we conducted an analysis to validate SocRATES structured design:

- **Reliability.** We assessed the reliability of SocRATES for both unguided (no rough scenario) and guided (with rough scenario) generation across two different maps. For unguided generation, we created 20 scenarios in each location, while for guided generation, we generated

5 scenarios, 10 times each, in two locations (100 total). Each scenario was manually evaluated for simulability, contextual appropriateness, and alignment with the input rough scenario. We observed a 55% success rate for unguided generation and a 73% success rate for guided generation. Most failures were due to incorrect behavior trees — often caused by the LLM using incorrect node selection or ordering — and poor trajectories that were too short or inaccurate for the proposed scenario. Note that in this experiment, we did not use interactivity and accepted the first syntactically valid response. It was possible to correct the trajectory-related errors by re-querying the LLM with specific instructions.

- **Cost.** Given that closed-source large models charge based on token usage, we evaluated the total cost of running our system by measuring the combined token length of all prompts, including typical map image inputs (we used the warehouse map). Assuming we accepted the first response for each module, the total input tokens amounted to approximately 15,000 tokens, which costs less than 30 cents per trial using GPT-4o.
- **Speed.** To assess the time required to generate a typical scenario, we created 20 scenarios in one location (without interactivity). We measured the wall-clock time needed to generate a scenario. On average, it took 43 seconds to generate a scenario, which is considerably faster than manually-crafting and coding scenarios.
- **Structured Prompting.** To evaluate the importance of our structured prompts and error-handling mechanisms, we compared our system to a simpler, naive system for path and behavior generation. The naive system lacked our context-rich prompts and directly queried a VLM for path and behavior proposals. As in the reliability test, we again generated and evaluated 100 guided scenarios and 20 unguided scenarios with the naive approach (with the same social context, task and scenario inputs as SocRATES). We observed a 30% overall success rate for guided generation and 10% success rate for unguided generation. This result highlights the critical role of structured prompts and error-handling mechanisms in improving the reliability and quality of scenario generation.

In summary, SocRATES is cost-effective and efficient, generating scenarios in under a minute with minimal cost. The system achieves a 73% success rate for first-cut generation, with improvement through interactive adjustments. The comparative study with the naive system confirms that structured prompts and error-handling are key to improving accuracy.

B. Usability Study with Academic Researchers

We conducted a user study to better understand the usefulness of SocRATES as an evaluation tool for social navigation researchers. We engaged five researchers from the National University of Singapore, Yale University, Miraikan, and

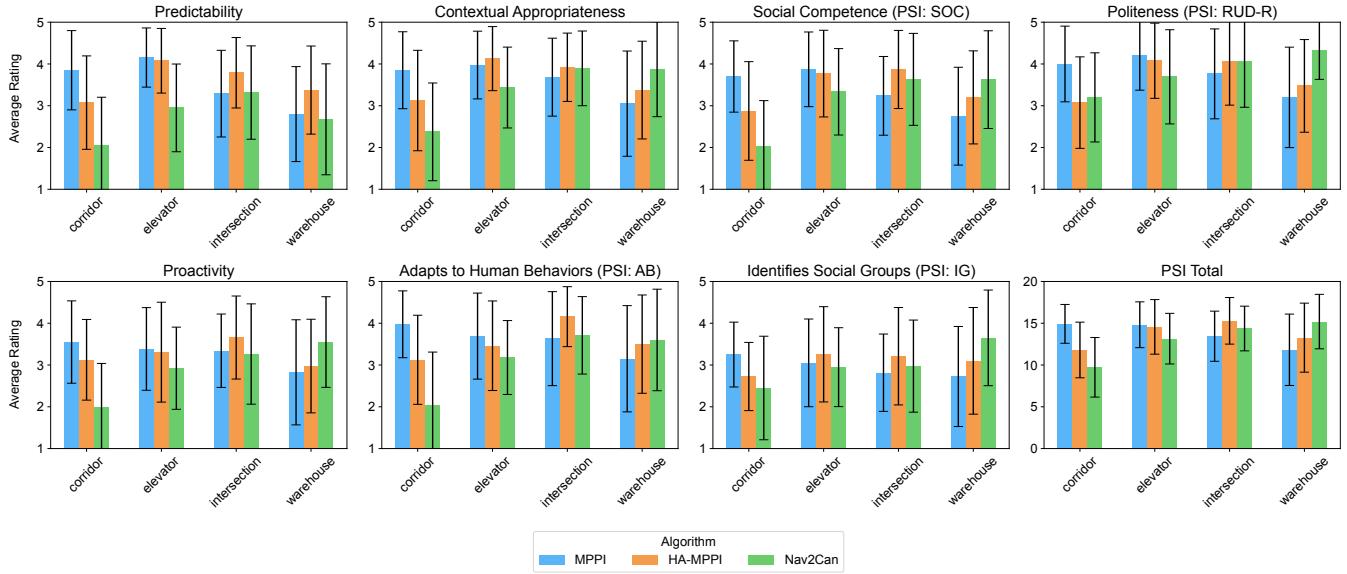


Fig. 3: Participant ratings for the navigation algorithms for the four scenarios across various social dimensions.

LAAS-CNRS. Each participant provided consent², and we demonstrated the map generation process before asking them to input their own scenario metadata. We then showcased the resulting simulation by either teleoperating the robot or running a Nav2-based planner. Below, we summarize the key results:

- **Ease of Use and Reliability:** SocRATES successfully generated the required scenarios for 4 out of 5 users. For one user, the behaviors were semantically correct but syntactically incorrect, an issue that has since been resolved. All participants agreed that the system was easy to use and intuitive. Users expressed a preference for using SocRATES over manually creating scenarios, especially when generating diverse scenarios from ambiguous inputs. However, one user noted that for highly specific scenarios, manual generation may still be preferred.
- **Utility:** Users found various aspects of SocRATES valuable as evidenced by feedback: “*It’s great that I don’t have to program each human in the scenario by hand.*” and “*Being able to generate scenarios with interaction is useful.*”. Users found that “*Proposing scenarios from potentially ambiguous context and task information is very useful.*”, and “*The variability in LLM-generated human behaviors models real-life randomness in human actions.*” Overall, users found SocRATES beneficial for scenario generation and preferred it over manual scenario development.
- **Suggestions:** Users had different use cases and provided a number of suggestions for future improvements. These included integrating more simulators (e.g., Flatland³

²Our user studies and benchmark were approved under the IRB study no. NUS-IRB-2024-590

³<https://flatland-simulator.readthedocs.io/en/latest/>

and Isaac⁴), introducing plug-and-play human models, enhancing the UX design and legibility of the CLI tool, enabling larger crowd simulations, and developing a method to objectively evaluate the fidelity of generated scenarios. These suggestions point to promising directions for future research and development.

Overall, this user study demonstrates that SocRATES is a valuable tool for simplifying scenario generation for social navigation algorithms.

C. Case Study via Persona-based Assessment

In this section, we demonstrate the usability of SocRATES through a case study and aim to identify potential areas for improvement.

We employed a methodology similar to persona-based assessment [19], adopting the role of Amy, an HRI researcher. Amy works at a delivery robot company and is interested in understanding how people perceive the behavior of navigation algorithms in social settings. She has a background in Human-Computer Interaction and is familiar with software development and ROS but does not program robots daily. Amy’s task is to advise her development team on the social capabilities of three different navigation algorithms:

- Model Predictive Path Integral Controller (MPPI) with the NavFn Planner⁵;
- Human-aware NavFn-MPPI (HA-MPPI), which augments the above with costmaps [20] that penalize entering the social zone of nearby humans;
- Nav2Can [21], which uses costmaps to penalize passing through groups and entering the social zone of nearby humans.

⁴<https://developer.nvidia.com/isaac/sim>

⁵<https://docs.nav2.org/configuration/packages/configuring-navfn.html>

Amy's goal is to determine if there is a socially preferred algorithm and how each performs in specific scenarios. Her task involves:

- Using SocRATES to create four distinct scenarios and generate videos of the robot's behavior;
- Collecting feedback from an online pool of participants;
- Analyzing and reporting the results.

We generated four scenarios in which the robot had to:

- 1) Enter an elevator in a multi-floor hospital;
- 2) Navigate a corridor in an office environment⁶ with humans following and obstructing its path;
- 3) Navigate an intersection where humans suddenly appear from a room in an office;
- 4) Navigate past a group of employees in a warehouse.

Adopting the Amy persona, we found scenario generation using SocRATES to be straightforward. For example, to generate the elevator scenario, Amy used the prompt: *The Robot approaches an elevator to go to a different floor to deliver supplies. The elevator opens and has a few people in it. 2 people leave the elevator while one of them is startled by the robot and keeps looking at it and doesn't leave the elevator.* The complete prompts used are available in our online repository⁷. On average, each scenario took 10 minutes to complete. The most difficult scenario to craft was the group scenario in the warehouse, where the generated trajectories were suboptimal in that the humans were placed in an unrealistic group formation and specific behaviors of the humans were triggered too early. As such, it was necessary to modify the trajectories and behaviors manually (though, this was easier than creating the trajectories from scratch). Annotating the hospital map also required finding a balance between very specific trajectories and giving the navigation planner freedom to choose alternate paths. These observations highlighted areas for improvement in future versions of SocRATES.

After generating and visualizing the scenarios in Gazebo, Amy easily recorded videos for use in a questionnaire. Snapshots of the videos for selected scenarios are shown in Fig. 4. She selected questions from the PSI [11] and added questions about social navigation principles [4], focusing on predictability, contextual appropriateness, and proactivity.

Amy surveyed 40 participants per scenario (adults with no vision disabilities) through Prolific⁸ — note that we actually conducted this study as part of our persona-based assessment. The results revealed significant differences between the algorithms in the scenarios. For example, as shown in Fig. 3, the basic MPPI algorithm received higher Perceived Social Intelligence (PSI) scores than Nav2Can in the corridor scenario, while Nav2Can performed better in the warehouse scenario. This was reinforced by participants comments: “*I think it [MPPI] was very effective as it got through the*



(a) Elevator Scenario



(b) Intersection Scenario

Fig. 4: Two of the scenarios generated in the Persona-based Assessment

corridor at a good speed without being too much of an issue for people passing through.”, “[Nav2Can] moved out of the way to go around the group of people. [MPPI] and [HA-MPPI] drove right into the group of people, which could cause accidents in the warehouse.” This led Amy to suggest that different algorithms might be more suitable for different social contexts, or further development was needed on general social navigation.

To summarize, this study affirmed the findings from our design validation and usability studies. Without SocRATES, we estimate that Amy would have spent approximately a month conducting this study manually, compared to 4 days with SocRATES. The study also revealed areas for improvement, such as in trajectory generation and modification.

V. CONCLUSION

In this work, we introduced SocRATES, an automated system for generating diverse, contextually rich social navigation scenarios in simulated environments. Through usability studies with researchers and a persona-based assessment, we demonstrated the system’s effectiveness and ease of use. SocRATES marks an important step towards scalable, scenario-based testing and benchmarking for social navigation algorithms. By incorporating human evaluations alongside scenario-based testing, we aim to complement existing proxemics-focused benchmarks and enable the evaluation of subjective, hard-to-define social metrics.

Building on the feedback from our user study, we plan several improvements to enhance the system’s robustness and user experience. First, LLMs do make reasoning errors and we plan to explore fine-tuning models specifically for scenario generation to improve reliability. Additionally, the current requirement for users to manually annotate maps can be time-consuming, so we plan to automate this process and include procedurally generated environments based solely on social context and location descriptions. Finally, we aim to develop a validation module to ensure that the generated scenarios accurately reflect the user’s input. With these enhancements, SocRATES has the potential to significantly advance the testing and evaluation of social navigation algorithms, providing a scalable, flexible platform for assessing both technical performance and social behavior.

⁶https://github.com/Arena-Rosnav/arena-simulation-setup/tree/master/worlds/arena_nus_com1_building

⁷Repository available after review.

⁸<https://www.prolific.com/>

REFERENCES

- [1] N. Tsoi, A. Xiang, P. Yu, S. S. Sohn, G. Schwartz, S. Ramesh, M. Hussein, A. W. Gupta, M. Kapadia, and M. Vázquez, “Sean 2.0: Formalizing and generating social situations for robot navigation,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 047–11 054, 2022. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9851501>
- [2] N. Pérez-Higueras, R. Otero, F. Caballero, and L. Merino, “Hunavsim: A ros 2 human navigation simulator for benchmarking human-aware robot navigation,” *IEEE Robotics and Automation Letters*, 2023. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10252030/>
- [3] 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.
- [4] A. Francis, C. Pérez-D’Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, H.-T. L. Chiang, M. Everett, S. Ha, J. W. Hart, J. P. How, H. Karnan, T.-W. E. Lee, L. J. Manso, R. Mirsky, S. Pirk, P. T. Singamaneni, P. Stone, A. V. Taylor, P. Trautman, N. Tsoi, M. Vázquez, X. Xiao, P. Xu, N. Yokoyama, A. Toshev, R. M.-M. L. Robotics, Nvidia, Stanford, Google, Epfl, Purdue, Cmu, U. Austin, Mit, Northeastern, G. Tech, Aston, B. Ilan, Adobe, Laas-Cnrs, U. de Toulouse, A. Sony, Honda, Yale, Gmu, and Apple, “Principles and guidelines for evaluating social robot navigation algorithms,” *ArXiv*, vol. abs/2306.16740, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259287246>
- [5] E. T. Hall, “The hidden dimension,” *Garden City*, 1966.
- [6] B. Okal and K. O. Arras, “Formalizing normative robot behavior,” in *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*. Springer, 2016, pp. 62–71.
- [7] N. Pérez-Higueras, F. Caballero, and L. Merino, “Teaching robot navigation behaviors to optimal rrt planners,” *International Journal of Social Robotics*, vol. 10, pp. 235–249, 2018.
- [8] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, “Legibility and predictability of robot motion,” in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.
- [9] A. V. Taylor, E. Mamantov, and H. Admoni, “Observer-aware legibility for social navigation,” in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1115–1122.
- [10] A. Honour, S. B. Banisetty, and D. Feil-Seifer, “Perceived social intelligence as evaluation of socially navigation,” in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 519–523.
- [11] K. A. Barchard, L. Lapping-Carr, R. S. Westfall, A. Fink-Armold, S. B. Banisetty, and D. Feil-Seifer, “Measuring the perceived social intelligence of robots,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 4, pp. 1–29, 2020.
- [12] J. Ma, X. Che, Y. Li, and E. M.-K. Lai, “Traffic scenarios for automated vehicle testing: A review of description languages and systems,” *Machines*, vol. 9, no. 12, p. 342, 2021.
- [13] M. Wen, J. Park, and K. Cho, “A scenario generation pipeline for autonomous vehicle simulators,” *Human-centric Computing and Information Sciences*, vol. 10, no. 1, p. 24, 2020.
- [14] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, “Language-guided traffic simulation via scene-level diffusion,” in *Conference on Robot Learning*. PMLR, 2023, pp. 144–177.
- [15] M. C. Fontaine and S. Nikolaidis, “Evaluating human–robot interaction algorithms in shared autonomy via quality diversity scenario generation,” *J. Hum.-Robot Interact.*, vol. 11, no. 3, sep 2022. [Online]. Available: <https://doi.org/10.1145/3476412>
- [16] V. Bhatt, H. Nemlekar, M. C. Fontaine, B. Tjanaka, H. Zhang, Y.-C. Hsu, and S. Nikolaidis, “Surrogate assisted generation of human–robot interaction scenarios,” *arXiv preprint arXiv:2304.13787*, 2023.
- [17] L. Kästner, V. Shcherbyna, H. Zeng, T. A. Le, M. H.-K. Schreff, H. Osmaev, N. T. Tran, D. Diaz, J. Golebiowski, H. Soh, and J. Lambrecht, “Arena 3.0: Advancing social navigation in collaborative and highly dynamic environments,” 2024. [Online]. Available: <https://www.arxiv.org/abs/2406.00837>
- [18] A. Biswas, A. Wang, G. Silvera, A. Steinfeld, and H. Admoni, “Socnavbench: A grounded simulation testing framework for evaluating social navigation,” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 3, pp. 1–24, 2022.
- [19] S. Faily and I. Flechais, “Persona cases: a technique for grounding personas,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 2267–2270.
- [20] D. V. Lu, D. Hershberger, and W. D. Smart, “Layered costmaps for context-sensitive navigation,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 709–715.
- [21] T. Schwörer, J. E. Schmidt, and D. Chrysostomou, “Nav2can: Achieving context aware navigation in ros2 using nav2 and rgb-d sensing,” in *2023 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2023, pp. 1–6.