From Obstacles to Etiquette: Robot Social Navigation with VLM-Informed Path Selection

Zilin Fang[1], Anxing Xiao[1], David Hsu[1,2], and Gim Hee Lee[1]
[1]School of Computing, [2]Smart Systems Institute, National University of Singapore

Fig. 1: (a) Illustration of robot navigation in a scenario with three geometrically feasible sampled paths, where the robot should reason about social conventions to select the most appropriate path. (b) Navigation performance on four experimental scenarios with our proposed framework. Without any consideration of collision avoidance and social compliance, the goal is set directly ahead of the starting point, marked by a red map pin. For each scenario, we display and highlight one VLM-generated answer to the image query on the left-hand side.

***Abstract*— Navigating socially in human environments requires more than satisfying geometric constraints, as collision-free paths may still interfere with ongoing activities or conflict with social norms. Addressing this challenge calls for analyzing interactions between agents and incorporating common-sense reasoning into planning. This paper presents a social robot navigation framework that integrates geometric planning with contextual social reasoning. The system first extracts obstacles and human dynamics to generate geometrically feasible candidate paths, then leverages a fine-tuned vision-language model (VLM) to evaluate these paths, informed by contextually grounded social expectations, selecting a socially optimized path for the controller. This task-specific VLM distills social reasoning from large foundation models into a smaller and efficient model, allowing the framework to perform real-time adaptation in diverse human–robot interaction contexts. Experiments in four social navigation contexts demonstrate that our method achieves the best overall performance with the lowest personal space violation duration, the minimal pedestrian-facing time, and no social zone intrusions. Project page: path-etiquette.github.io**

I. Introduction

Navigating robots in crowded environments is both challenging and critical for applications such as autonomous delivery and guidance [1]–[3]. Beyond goal reaching under geometric constraints, robots should also adhere to social etiquette in crowds—recognizing cues such as people posing for photographs or a worker on a ladder—and act to minimize disruption and risk [4], as illustrated in Fig. 1. However, generating such contextually adaptive motions remains difficult, as it requires both spatial and semantic awareness to identify socially compliant spaces, yet relevant robot datasets are scarce. In this work, *we propose a social robot navigation framework that leverages Vision-Language Models (VLMs) to evaluate the social conventions associated with feasible geometric paths*, enabling robots to follow social etiquette without explicitly modeling socially compliant spaces.

Prior works employing VLMs for social robot navigation have largely focused on high-level semantic reasoning at the symbolic level, such as predicting explicit social relationships [5], choosing navigation goals projected in images [6], [7], predicting pedestrian intention [8], or scoring predefined actions [9]. However, evaluating purely on symbolic representations fails to capture real-world execution effects, which can lead to intrusions or navigation failures. In addition, frequent queries to large VLMs can result in slow inference, limiting real-time applicability.

To address the above challenges, we present a navigation framework that integrates spatial constraints from obstacles and humans with semantic adaptations derived from human-environment, human-human, and human-robot interactions in real-time. The social robot navigation task is formulated as a multi-objective optimization problem over geometric feasibility and in-context semantics. These objectives are assumed to be decomposable, with the optimal semantic solution space lying within a well-covered subset of the geometric optimum. Accordingly, we adopt a hierarchical architecture with asynchronous modules for geometry-aware path planning, socially compliant path selection, and safe reactive control. At the high level, candidate paths are first sampled under geometric constraints, and then VLMs are employed to select grounded, socially compliant paths from this feasible set. This proceeds in a receding-horizon fashion. For real-time inference, we introduce a pipeline that distills social reasoning capabilities from large models into a smaller model. At the low level, the selected path is fed back to

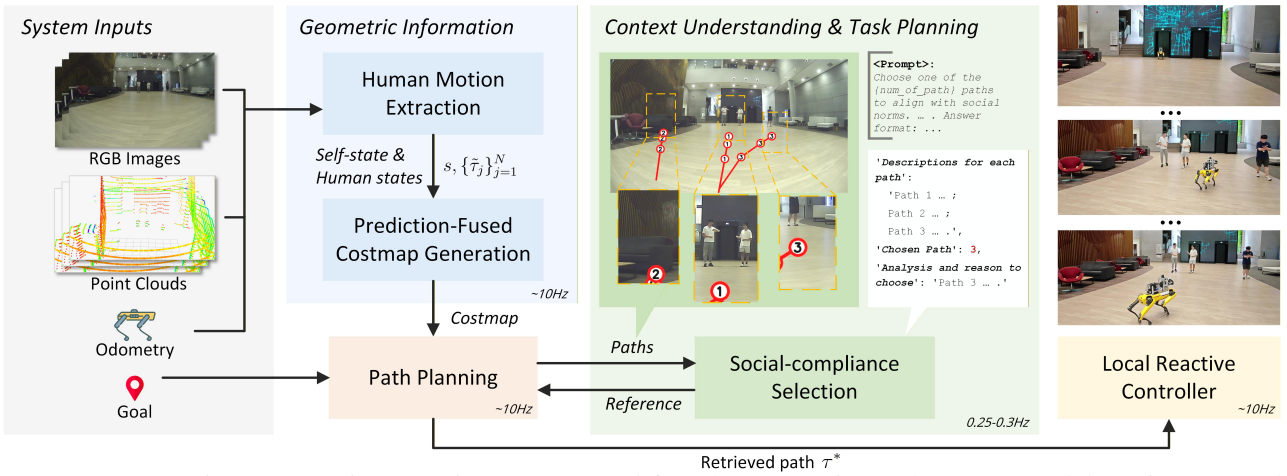

Fig. 2: System overview. Geometric constraints are extracted from human motion and costmap modules using sensor data. Collision-free path candidates are sampled, projected into the image, and evaluated by a fine-tuned VLM. The selection is fed back as reference to retrieve a path for the local controller.

the path planning module as a reference for generating new paths, while a modified ORCA algorithm is employed to ensure pedestrian avoidance. At a conceptual level, our propose–select strategy is close to prior work [10], [11], but it targets a different problem domain with dynamic, interaction-aware reasoning objectives and a different system design.

We evaluate our method through controlled experiments on a Boston Dynamics Spot legged robot in diverse social scenarios involving human activity. These scenarios are motivated by the insight that, while social behavior is inherently multimodal and shaped by individual preferences, many situations follow common patterns that minimize intrusion into social zones. This design reduces reliance on subjective post-hoc surveys and enables social performance to be quantified using interruption-related metrics. Real-world experiments demonstrate that our approach achieves superior social compliance and more efficient goal-reaching compared to representative baselines, including group-based [5], [12], RL-based [13], VLM-based [9] methods, as well as a foundation model for visual navigation [14]. Ablation studies further show that our approach generates collision-free and socially compliant paths, which direct VLM path prediction cannot reliably guarantee.

II. Method

A. Formulation

The social robot navigation task can be conceptually understood as a multi-objective optimization problem, inspired by [4], [15]. Consider a robot r and a planar workspace $\mathcal{W} \subseteq \mathbb{R}^2$ and the traversable subspace is defined as $\mathcal{W}_{obs}^C = \mathcal{W} \setminus \mathcal{W}_{obs}$, where $\mathcal{W}_{obs} \subseteq \mathcal{W}$ is the subspace occupied by static obstacles. Assume that robot r now lies in a configuration $s \in \mathcal{Q} \subseteq SE(2)$, with a mapping M embedding s into the workspace $M(s) \subseteq \mathcal{W}$ such that $q \notin \mathcal{Q}_{obs}\ \forall q \in \mathcal{Q}$ when $M(q) \notin \mathcal{W}_{obs}$. During the normalized time interval $[0, 1]$, the total number of human agents is N, with the individual path of each agent denoted as $\{\tau_j\}_{j=1}^N$, and the estimation of the path represented as $\{\tilde{\tau}_j\}_{j=1}^N$ since it is challenging to acquire the ground truth of the human agent behaviors. During planning, the robot computes a path $\tau \in \mathcal{T}$ toward the goal g formulated as the following optimization problem:

$$\begin{aligned} \tau = \arg\min_{\tau \in \mathcal{T}} \quad & \left(c(\tau), c^{social}(\tau, \{\tilde{\tau}_j\}_{j=1}^N; \rho_{crowd})\right) \\ \text{s.t.} \quad & \tau(t) \in \mathcal{W}_{obs}^C,\ \forall t \in [0,1], \\ & \tau(t) \cap \tilde{\tau}_j(t) = \emptyset,\ \forall j \in \{1, ..., N\}\ \forall t \in [0,1], \\ & \tau(0) = M(s), \tau(1) = M(g) \end{aligned} \tag{1}$$

where $c : \tau \to \mathbb{R}$ is the cost function for path specifications (e.g., time to goal) and environment constraints (e.g., untraversable obstacles). c^{social} is a joint cost that combines all social considerations, whose form may vary with the density of the crowd ρ_{crowd}. The path $\tilde{\tau}_j$ can be expressed as explicit waypoints in a planar space or implicit feature representations. In this work, we focus on scenarios that involve more analysis of social norms beyond the collision avoidance under low to moderate crowd density, and omit ρ_{crowd} for simplicity in the following context. Since human agents' path estimations should be updated based on information collected up to the current time, the path optimization is performed in a receding horizon manner starting from the current time step, t'. The planning horizon is defined as H with the constraint expressed as $\tau(t) \cap \tilde{\tau}_j(t) = \emptyset,\ \forall j \in \{1, ..., N\},\ \forall t \in [t', t' + H]$. We assume the objective is decomposable into geometric feasibility and in-context semantics, with semantic optima forming a well-covered subset of the geometric-optimal set.

B. Constraint Extraction

The system is built upon the formulated optimization problem. The core idea is to first generate plausible path candidates that inherently satisfy geometric constraints derived from sensor data. A VLM then selects the most socially compliant path. This enables simultaneous semantic scene understanding and task planning. Fig. 2 illustrates the overall navigation pipeline and the interactions between its modules.

Experimentally, we handle the constraints $\tau(t) \in \mathcal{W}_{obs}^C$ and $\tau(t) \cap \tilde{\tau}_j(t) = \emptyset$ using a costmap that fuses real-time ob-

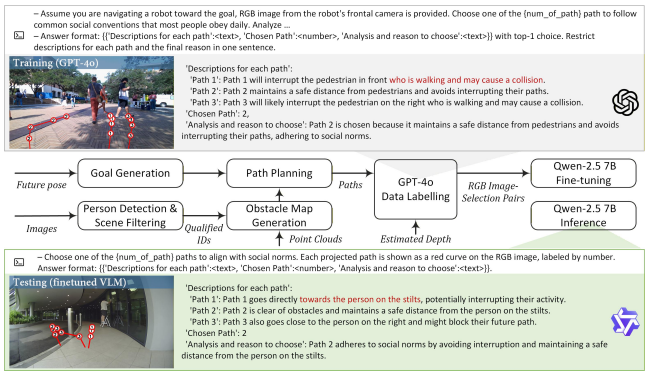

Fig. 3: Fine-tuning data generation and example query-answer pairs.

stacle sensing with short-horizon human motion predictions. Environmental constraints are assumed to be largely captured by the sensed costmap. The path specification part in the cost function $c(\tau)$ is absorbed directly into the sampling and post-processing of candidate paths. Human Motion Extraction and Prediction-Fused Costmap Generation process incoming sensor data to obtain 3D human trajectories and generate future-aware costmaps with trajectory augmentation.

C. Path Planning

Path Planning then samples collision-free path candidates based on this costmap. Since the costmap encodes predicted human motions, distance-related factors in the social cost c^{social} such as personal space and potential future blocking are partially addressed during path generation.

The module builds on the classical A* algorithm to generate multiple feasible paths from the robot to the goal. Since standard A* returns only the shortest path, we introduce anchors to diversify candidates, enabling longer, socially compliant paths that capture yielding or detours. Anchors are sampled within a rectangular region from the ego position toward the goal using a Poisson disk strategy to ensure spatial diversity, which random sampling cannot guarantee. For efficiency, the region is discretized into a grid, and obstacle-free grid corners are selected as anchors based on the cost map. Each candidate path is constructed by concatenating two A* segments: start-to-anchor and anchor-to-goal. Multiple anchors produce diverse path modalities aligned with social navigation behaviors. To reduce redundancy, paths are clustered using a Hausdorff distance [16] threshold, retaining three to six representatives. These candidates are then projected onto the RGB image for input to the VLM. This module continuously replans candidate paths as the robot moves. Due to VLM inference latency, the selected path serves only as a reference and is not directly followed. Upon receiving the VLM output, the system retrieves the closest matching path from the current candidate set, while the local reactive controller handles velocity planning and dynamic obstacle avoidance.

D. Social-compliance Selection

Subsequently, the Social-Compliance Selection module identifies the optimal path through a multiple-choice decision process. Acting as a proxy for the average human, the VLM assesses whether each candidate path adheres to social norms by jointly considering all factors in c^{social}, thereby incorporating social cost into its analysis and solve the minimization problem $\arg\min\left(c(\tau), c^{social}(\tau, \{\tilde{\tau}_j\}_{j=1}^{N})\right)$ by selecting the highest-scoring path. The Path Planning module replans regularly and retrieves the closest path upon receiving a reference path from the VLM output. Finally, the Local Reactive Controller chooses suitable waypoints from the retrieved path as continuously updated sub-goals as the robot moves toward its final goal, and issues corresponding velocity commands while considering human states.

Our work grounds social compliance in selecting a path that minimizes interruptions to ongoing human activities. We leverage VLMs for semantic understanding to provide high-level path reference, inferring likely human motions and interactions from posture and context to reason about social norms along each path, while decoupling path selection from velocity control for asynchronous operation.

1) Visual Prompting: Given an input image with multiple projected paths and depth information, a large VLM such as GPT-4o can interpret the scene, evaluate each path according to social norms, and select the most appropriate option. This evaluation can also be performed by a smaller fine-tuned VLM such as Qwen-2.5 at inference, using a similar constrained prompting strategy. To encourage socially compliant reasoning, prompts include (i) a one-sentence description of each candidate path, (ii) an explicit choice of the selected path, and (iii) a justification of the decision, all within a structured output format as shown in Fig. 3. By constraining the output format, we encourage the model to explicitly generate chain-of-thought reasoning, thereby enhancing complex reasoning performance. GPT-4o receives richer instructions to elicit social-norm analysis, while Qwen-2.5 uses simplified prompts and RGB input only to reduce latency while preserving the output schema.

2) Fine-tuning: While GPT-4o demonstrates strong performance, its high inference latency makes it impractical for real-time decision-making. To address this limitation, our objective is to distill GPT-4o's socially compliant reasoning ability into a smaller and faster VLM, Qwen-2.5 7B [17], which offers significantly lower latency. This enables robots to respond more rapidly to dynamic environmental changes while maintaining social awareness. Specifically, we use the SCAND [18] dataset to generate training data for fine-tuning the Qwen model, as illustrated in Fig. 3. For each frame, a rough goal direction is estimated from the future 6-second trajectory and then perturbed within $\pm 10^{\circ}$ to sample a goal location 13 to 14.5 meters away. Due to the absence of LiDAR–camera extrinsics, we run our path planning algorithm on instantaneous obstacle maps from Velodyne Puck point clouds and project the resulting paths onto the image using approximate calibration. The dataset is downsampled to 1 Hz and frames without humans are filtered out, yielding 4,851 image–selection pairs for training and 574 for validation.

As Qwen-2.5 7B model already possesses strong general visual reasoning abilities, our fine-tuning aims to constrain

TABLE I: Navigation performance comparison on four different scenarios.

(a) Wiping glass-wall

Method	NT↓	PSV↓	TFP↓	SIT↓	Max. SIR↓ (%)
G-MPC	54.55	0.00	0.00	8.41	13.22
AttnGraph-RL	29.11	3.81	1.75	5.65	16.97
ViNT	46.24	0.00	0.00	12.32	16.97
VLM-Social-Nav	32.21	0.00	0.00	0.71	2.86
GSON	30.42	0.00	0.00	8.28	16.97
Ours	**28.48**	**0.00**	**0.00**	**0.00**	**0.00**

(b) Walking-talk

Method	NT↓	PSV↓	TFP↓	SIT↓	Max. SIR↓ (%)
G-MPC	68.13	1.73	5.69	2.63	44.10
AttnGraph-RL	*90.00	0.20	**1.66**	0.00	0.00
ViNT	47.14	1.54	4.01	1.43	42.80
VLM-Social-Nav	*90.00	4.99	7.57	0.00	0.00
GSON	38.25	0.00	6.59	1.35	39.70
Ours	**28.49**	**0.00**	1.80	**0.00**	**0.00**

(c) Photography

Method	NT↓	PSV↓	TFP↓	SIT↓	Max. SIR↓ (%)
G-MPC	36.14	3.68	8.07	3.67	27.66
AttnGraph-RL	*90.00	1.71	5.28	0.00	0.00
ViNT	36.82	2.91	9.28	3.43	27.66
VLM-Social-Nav	**27.01**	7.15	13.03	3.89	26.46
GSON	46.54	2.41	5.72	0.68	6.77
Ours	40.15	**0.0**	**4.32**	**0.00**	**0.00**

(d) Queuing

Method	NT↓	PSV↓	TFP↓	SIT↓	Max. SIR↓ (%)
G-MPC	47.11	5.22	21.26	20.26	20.46
AttnGraph-RL	**17.18**	1.25	5.45	2.81	21.03
ViNT	*90.00	1.45	6.25	0.00	0.00
VLM-Social-Nav	*90.00	**0.00**	4.25	0.00	0.00
GSON	29.10	0.38	**2.36**	0.00	0.00
Ours	26.14	0.33	2.70	**0.00**	**0.00**

Numbers reported are the average of three trials.
*If the robot fails to reach the goal within 90 seconds, we note the navigation time (NT) as 90.

its output format and stimulate its inherent social reasoning capabilities beyond the dataset. We perform full-model supervised fine-tuning using two H100 GPUs. At inference, the model is served with vLLM [19] in 4-bit quantization for efficiency: input images are sent to the H100 server, which processes them and returns the output to the robot.

III. Experiments

A. Platform Description

Our implementation is based on a Boston Dynamics Spot legged robot, equipped with an NVIDIA Jetson Orin for onboard navigation processing, excluding the VLM. The fine-tuned Qwen model runs as a network service on an H100 GPU. The sensing inputs include monocular RGB images from an ELP fisheye camera and point clouds from a Hesai JT16 mechanical LiDAR. Odometry data is obtained directly from the robot.

B. Experimental Setup

Experimental scenarios. Drawing on insights from social scenarios and previous works [4], [5], [9], we design four experimental settings with an expected robot behavior, respectively: (a) Wiping glass-wall – The robot is expected to take a wide detour and maintain a safe distance to account for potential fall hazards. (b) Walking-talk – The robot is expected to avoid passing through the gap between the conversing individuals. (c) Photography – The robot is expected to avoid crossing the line of sight between the photographer and the model. (d) Queuing – The robot is expected to not cut in line.

Baselines and Metrics. Social compliance is compared with five different navigation benchmarks, with three trials per scenario: G-MPC [12], Attention-Based Interaction Graph [13] (hereafter referred to as AttnGraph-RL), ViNT [14], VLM-Social-Nav [9], and GSON [5]. These methods span MPC-based control, reinforcement learning, and foundation model–based approaches. We use five metrics to evaluate robot behavior, three of which follow prior studies: (i) Navigation Time (NT); (ii) Personal Space Violation duration (PSV) [20], [21]; (iii) Time Facing Pedestrians (TFP), defined conservatively as the duration within a 30° cone and a range of 3m. Two additional metrics are introduced to assess social-zone disruptions: (iv) Social-zone Interruption Time (SIT); (v) Maximum Social-zone Interruption Ratio (Max. SIR), defined as the largest ratio between the overlapped area of Spot's footprint (0.5×1.1 m) and the area of any social zone. The social zone in scene (a) is defined as a square with a side length of 1.8m based on fall-safety guidelines. In other scenes, it is defined as the area between people engaged in conversation, photography, or queuing.

C. Experiment Results

Fig. 1 (b) visualizes the navigation process of our system, and Table I presents the quantitative comparisons of the five evaluation metrics. Our method delivers consistently strong performance across all scenarios. It achieves collision-free navigation without intruding into social zones and maintains low duration of personal space violation (PSV) and time facing pedestrians (TFP). This performance is attributed to our two-stage strategy: sampling paths that respect both human motion dynamics and obstacle constraints, followed by social context reasoning and task planning through a fine-tuned VLM with a high-frequency low-level controller using the planned path as a reference. The design enables dynamic adaptation to environmental changes, ensures compliance with social norms, and generalizes effectively across diverse contexts. Additionally, predictive future path referencing helps mitigate VLM-induced latency.

IV. Conclusion

We present a navigation pipeline that enables collision-free, socially compliant, and efficient robot navigation across diverse human-centered contexts. Our method first extracts human motion information from images and point clouds, then fuses motion predictions to guide collision-free path sampling under geometric constraints. A context-aware task planning module, powered by a fine-tuned VLM, interprets social norms and selects the best path accordingly. Extensive real-world experiments demonstrate that our system adapts robustly to various social scenarios, outperforming baseline methods across multiple social metrics.

REFERENCES

[1] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, *et al.*, "Spencer: A socially aware service robot for passenger guidance and help in busy airports," in *Field and Service Robotics: Results of the 10th International Conference*. Springer, 2016, pp. 607–622.

[2] A. Xiao, W. Tong, L. Yang, J. Zeng, Z. Li, and K. Sreenath, "Robotic guide dog: Leading a human with leash-guided hybrid physical interaction," in *IEEE Int. Conf. on Robotics & Automation*, 2021.

[3] S. Cai, A. Ram, Z. Gou, M. A. W. Shaikh, Y.-A. Chen, Y. Wan, K. Hara, S. Zhao, and D. Hsu, "Navigating real-world challenges: A quadruped robot guiding system for visually impaired people in diverse environments," in *ACM CHI Conference on Human Factors in Computing Systems*, 2024.

[4] 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 Trans. on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–65, 2025.

[5] S. Luo, P. Sun, J. Zhu, Y. Deng, C. Yu, A. Xiao, and X. Wang, "Gson: A group-based social navigation framework with large multimodal model," *IEEE Robotics & Automation Letters*, 2025.

[6] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, *et al.*, "Pivot: Iterative visual prompting elicits actionable knowledge for vlms," in *Int. Conf. on Machine Learning*, 2024.

[7] A. J. Sathyamoorthy, K. Weerakoon, M. Elnoor, A. Zore, B. Ichter, F. Xia, J. Tan, W. Yu, and D. Manocha, "Convoi: Context-aware navigation using vision language models in outdoor and indoor environments," in *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, 2024.

[8] F. Munir, S. Azam, T. Mihaylova, V. Kyrki, and T. P. Kucner, "Pedestrian vision language model for intentions prediction," *IEEE Open Journal of Intelligent Transportation Systems*, 2025.

[9] 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 & Automation Letters*, 2024.

[10] D. Song, J. Liang, X. Xiao, and D. Manocha, "Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments," *IEEE Robotics & Automation Letters*, 2025.

[11] J. Liang, K. Weerakoon, D. Song, S. Kirubaharan, X. Xiao, and D. Manocha, "Mosu: Autonomous long-range robot navigation with multi-modal scene understanding," *arXiv preprint arXiv:2507.04686*, 2025.

[12] A. Wang, C. Mavrogiannis, and A. Steinfeld, "Group-based motion prediction for navigation in crowded environments," in *Conference on Robot Learning*, 2022.

[13] S. Liu, P. Chang, Z. Huang, N. Chakraborty, K. Hong, W. Liang, D. L. McPherson, J. Geng, and K. Driggs-Campbell, "Intention aware robot crowd navigation with attention-based interaction graph," in *IEEE Int. Conf. on Robotics & Automation*, 2023.

[14] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "Vint: A foundation model for visual navigation," in *Conference on Robot Learning*, 2023.

[15] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh, "Core challenges of social robot navigation: A survey," *ACM Trans. on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–39, 2023.

[16] A. H. Raj, Z. Hu, H. Karnan, R. Chandra, A. Payandeh, L. Mao, P. Stone, J. Biswas, and X. Xiao, "Rethinking social robot navigation: Leveraging the best of two worlds," in *IEEE Int. Conf. on Robotics & Automation*, 2024.

[17] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[18] 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 & Automation Letters*, 2022.

[19] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[20] S. Narasimhan, A. H. Tan, D. Choi, and G. Nejat, "Olivia-nav: An online lifelong vision language approach for mobile robot social navigation," in *IEEE Int. Conf. on Robotics & Automation*, 2025.

[21] N. Hirose, D. Shah, K. Stachowicz, A. Sridhar, and S. Levine, "Selfi: Autonomous self-improvement with reinforcement learning for social navigation," *arXiv preprint arXiv:2403.00991*, 2024.