Towards Human-Like Instruction Navigation in Real-World

Jiseon Kim[1], Daeun Song[2], Bokeon Suh[1], Hyoseok Ju[1], Yumin Lee[1], Xuesu Xiao[2], and Giseop Kim[1*]

***Abstract*— For robots to successfully accomplish tasks in human daily life, they must possess the capability to navigate by understanding human-like instructions that people often use. However, current Object Goal Navigation (OGN) research primarily relies on detailed instructions that are not typically used in the wild, and thus still lacks the ability to navigate with abstract human guidance. Prior real-world socially compliant navigation work focuses on executing explicit social-norm based instructions, without wayfinding instructions or an explicit goal. We present Human-like INsTruction–grounded Navigation (HINT), a novel task in which an embodied agent must reach a goal solely from human verbal and non-verbal instructions. To support this task, we construct SocialACT, the first real-world dataset that bridges the gap between abstract human instructions and robot behaviors. Unlike traditional OGN tasks that only assess episode-level success or failure, SocialACT enables progress-based evaluation at sub-instruction granularity.**

I. Introduction

For robots to successfully reach their goals in everyday human environments, they must be able to understand the abstract wayfinding instructions that people often provide. Human-to-human navigation routinely relies on concise verbal cues such as "go that way" and non-verbal cues—head turns, gaze, and pointing gestures—grounded in shared spatial perception and commonsense. These cues are concise yet semantically rich, and people integrate them to infer where to go while maintaining safety and social compliance.

However, research in Object Goal Navigation (OGN) [1]–[5] is largely empowered by photo-realistic simulators [2], [3], [6], where agents operate in simulation based on fine-grained instructions [7]. As a result, current agents often lack the ability to act on the concise verbal or non-verbal wayfinding instructions that humans commonly use in everyday settings. However, prior OGN studies mainly assessed episode success as a binary outcome at the episode level, limiting fine-grained evaluation of methods, including how far an agent progressed and where failure occurred.

As illustrated in Fig. 2, research on socially compliant navigation [8]–[12] has primarily focused on socially compliant execution [8], [9], such as forecasting pedestrian trajectories to avoid collisions—under explicit social-norm based instructions, without explicit goals or wayfinding guidance.

However, as shown in Fig. 1, convenient human–robot tasking in real-world deployments cannot rely on explicit social-norm based instructions at every turn. Humans naturally provide brief, human-like wayfinding cues and expect the robot to navigate to the goal while remaining socially compliant. This capability remains underexplored; accordingly, real-world navigation datasets that include natural human wayfinding instructions remain scarce.

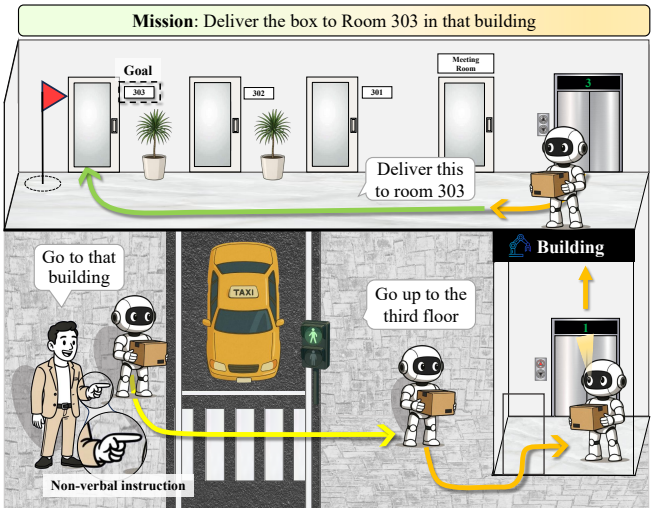

Fig. 1: Overview of HINT and the supervision provided by SocialACT. Humans often give brief instructions or gestures assuming shared common sense. HINT evaluates whether a robot can interpret and act on such human-like instructions. SocialACT pairs linguistic and non-linguistic cues with time-synchronized robot actions.

To address these limitations, we introduce **Human-like INsTruction grounded navigation (HINT)**, a novel task that requires a robot to interpret abstract verbal and non-verbal instructions and navigate to a goal in the real-world. We support HINT with **SocialACT**, a real-world dataset potentially designed for real-world benchmark. We annotate human-like wayfinding instructions that reflect how people naturally direct one another and time-synchronize the low-level robot sensor streams with human-teleoperated navigation executed according to those instructions. SocialACT comprises 100 real-world mission sequences totaling 5 h 14 min, spanning 70 distinct goals and 123 non-verbal sub-missions. In total, we collect 1.32M image frames across panoramic and cubemap images. For each mission, SocialACT enables progress-level evaluation of HINT task, where an agent navigates to a goal from human verbal and non-verbal wayfinding instructions. Our contributions can be summarized as follows.

- We propose HINT, a novel task for real-world settings that requires a robot to reach a goal by interpreting

[1]J. Kim, B. Suh, H. Ju, Y. Lee, and G. Kim are with the Department of Robotics and Mechatronics Engineering, DGIST, Daegu, Republic of Korea [jiseon.kim, bokeon.suh, hyoseok.ju, yumin.lee, gsk]@dgist.ac.kr

[2]D. Song and X. Xiao are with the Department of Computer Science, George Mason University, Fairfax, VA, USA [dsong26, xiao]@gmu.edu

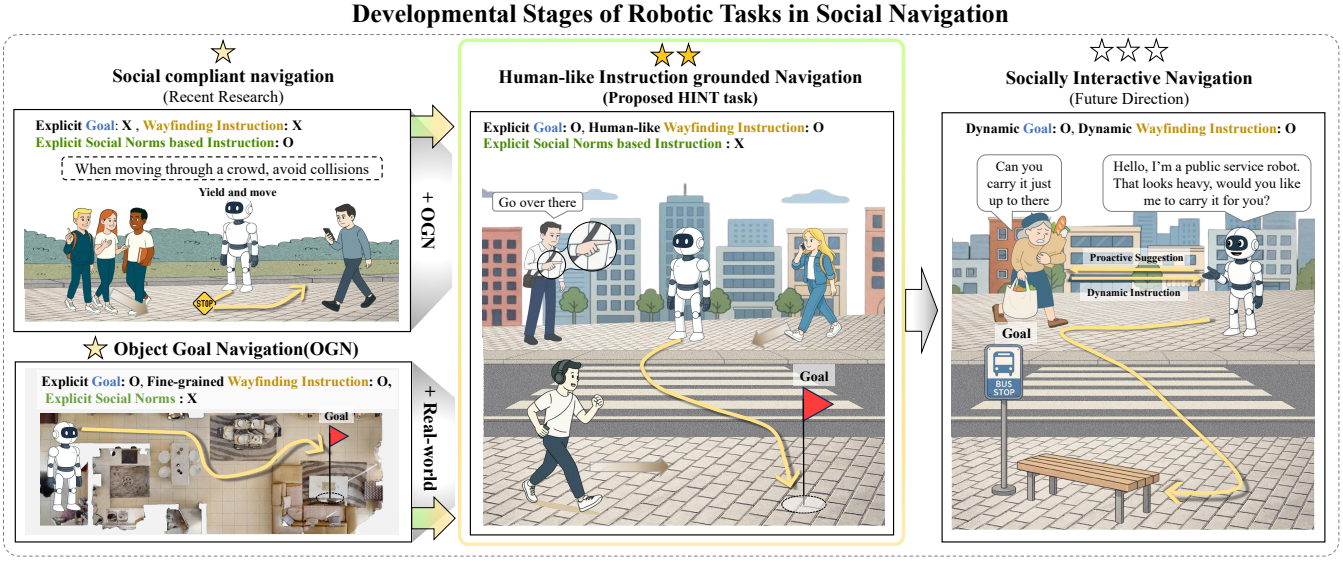

Fig. 2: Positioning of the proposed HINT task within the developmental stages of social navigation. Socially compliant navigation provides robots with explicit norm-based instructions to avoid collisions and to act according to rules such as yielding and waiting. The proposed HINT task requires robots to navigate using only human-like instructions without explicit social norms. Finally, socially interactive navigation represents the future stage, where robots proactively assist humans and adapt their navigation accordingly.

natural human verbal and non-verbal instructions.

- We construct SocialACT, the first real-world navigation dataset pairing natural verbal and non-verbal instruction with timestamp-aligned low-level robot sensor streams, organized into sub-missions.

II. Related Work

As depicted in Fig. 2, HINT advances beyond a simple union of socially compliant navigation and OGN by formulating goal-directed navigation from concise human guidance without hand-specified social-norm rules. Related datasets are summarized in Table I.

A. Social Compliant Navigation in the real-world

Early efforts toward social navigation in the real-world emphasize execution under explicit social norms [11], [13]–[15]. The goal is to move safely among people by avoiding collisions, yielding when necessary, and maintaining interpersonal distance, without an explicit goal [16], [17]. SCAND [8] and MuSoHu [9] are representative demonstration datasets that provide trajectories with social interaction tags. These resources have been valuable for benchmarking compliant motion in crowds and for studying perception signals associated with social behaviors.

Although foundational for early social navigation, these datasets lack human-like wayfinding instructions and explicit goals. Therefore, these datasets were not suitable for studying socially-aware OGN in the wild.

Perception focused datasets such as SNEI [10] and SocialNav-Sub [12] study social scene understanding through Visual Question Answering(VQA) pairs, which broadens the scope of social perception. However, these datasets do not couple perception to real-world action.

Existing research on socially compliant navigation typically relies on separate datasets for perception [10], [12] and execution [8], [9], which impedes integrated learning and evaluation of perception-conditioned action. To bridge this gap, we construct SocialACT, a real-world dataset that pairs concise, human-like wayfinding instructions with time-synchronized, human-teleoperated robot actions at the sub-mission level.

B. Object Goal Navigation in Simulation

OGN [16], [20], [21] have been studied in photorealistic simulators such as MP3D [2], HM3D [3], R2R [19], RxR [22] and REVERIE [4] collected verbal instructions and evaluate episode success. These benchmarks established core protocols for language guided navigation and object grounding. However, these datasets predominantly employ fine-grained wayfinding instructions that do not reflect the brief human guidance typically used in real-world settings. [23].

Recently, Social-MP3D and Social-HM3D [18] have evaluated socially compliant navigation with humanoid avatars. Importantly, humans are treated only as dynamic obstacles, with the focus remaining on collision-free efficiency, while human-like verbal and non-verbal instructions are not addressed.

In parallel, Zhang et al. introduced Uni-NaVid [24] in Habitat 3.0 with humanoid avatars, aiming to handle instructions for Vision-and-Language Navigation(VLN), OGN, Embodied Question Answering(EQA), and Human following. However, despite the diversity of instruction types provided in simulation, non-verbal human wayfinding cues also remain unexplored.

TABLE I: **Comparison of Datasets for Socially-aware Object Goal Navigation**

Task	Domain	Purpose	Dataset	Social compliant behavior	Explicit Goal	Sub–mission Labels (timestamp–level)	Human-like instruction
Social compliant navigation	Real–world	Execution	SCAND [8]	✓			
			MuSoHu [9]	✓			
		Perception	SNEI [10]	✓			
			SocialNav–Sub [12]	✓			
	Simulation	Execution	Social–HM3D, Social–MP3D [18]	✓	✓		
Object Goal Navigation	Simulation	Perception, Execution	Room2Room [19]		✓		Verbal instruction
			REVERIE [4]		✓		Verbal instruction
	Real–world	Perception, Execution	**Ours (SocialACT)**	✓	✓	✓	Verbal instruction, Non–verbal instruction

C. Positioning of Our Work

As illustrated in Fig. 2, the proposed HINT task targets human-level wayfinding capability in real-world settings: given only concise human wayfinding instructions—*without* any explicit social norm based instruction, the agent must navigate to the designated goal in a socially compliant manner. Ultimately, addressing the HINT task paves the way for real-world robots that can understand and act on natural human instructions, enabling seamless and intuitive human–robot interaction.

III. SocialACT Dataset

A. Human-like INsTruction grounded Navigation (HINT)

Humans routinely accomplish long-horizon navigation when another person provides brief guidance, verbally or non-verbally, such as pointing gesture, head turns, and gaze. In our setting, we ask whether an embodied agent can follow concise, human-like wayfinding instructions to reach a goal without any explicit social norm based instruction. Accordingly, *HINT* is a task that requires reaching the goal using only concise human-like verbal and non-verbal wayfinding instructions. The HINT task is defined by the following function. Let $\mathcal{I}_t$ denote the current instruction at time t, $\mathcal{M}_t$ the current visual measurement, which may be a panoramic or cubemap image, and $\mathcal{G}$ a goal image or text provided as a reference. At each decision step t, the agent receives $(\mathcal{I}_t, \mathcal{M}_t, \mathcal{G})$ and must predict the next waypoint action for the subsequent step $t+1$:

$$\mathcal{A}_{t+1} = \text{HINT}(\mathcal{I}_t, \mathcal{M}_t, \mathcal{G}).$$

This formulation evaluates whether the agent can effectively ground human-like instructions into executable actions that lead toward the goal using the available visual measurements.

B. SocialACT Dataset Design

SocialACT is a real-world, socially-aware OGN dataset collected to support the HINT task. Each mission consists of a person guiding the robot to a named goal through concise verbal or non-verbal instructions. The visual measurements includes panoramic and cubemap images captured onboard, together with other synchronized sensors.

To enable progress based evaluation, each mission is decomposed into sub-missions and sub-instructions that partition the full journey into interpretable units, as illustrated in Fig. 3. Annotations specify instruction modality (verbal, non-verbal, or mixed), sub-mission boundaries, goal image, and precise timestamps that synchronize instructions with images and odometry. We provide both panoramic and cubemap representations of the images. Through this design, the HINT task can be evaluated at the progress level in real-world environments.

C. SocialACT Hardware configuration

We set up a comprehensive sensor suite to collect the SocialACT dataset using a Unitree Go2-W platform equipped with an Insta360 X5 360° camera at 10 Hz, a Livox Mid-360 360° LiDAR at 10 Hz, a Gemini 335L stereo RGB-D camera at 10 Hz, and an IMU integrated with the Mid-360 at 200 Hz. LiDAR odometry is computed by KISS-ICP [25] at 10 Hz. All sensor streams are time-synchronized and stored as ROS 2 bag files for each mission. This configuration provides panoramic visual coverage and accurate geometry for socially-aware navigation in the real world.

D. Dataset collection procedure

Teleoperation policy. SocialACT was collected by four human tele-operators, who provided concise human-like wayfinding guidance and simultaneously executed human-like navigation in the real-world, yielding data where both grounding and execution reflect human-like behavior for robots to learn from. Within and across missions, operators alternated roles between teleoperating the robot and providing guidance through concise human-like instructions, including pointing and head or gaze shifts. The alternation of roles results in diverse navigation strategies toward the goal, while maintaining consistent data logging across missions.

Data preprocessing. We mitigate the inherent distortion in panoramic frames collected from the Insta360 camera while improving usability, We additionally project each panorama

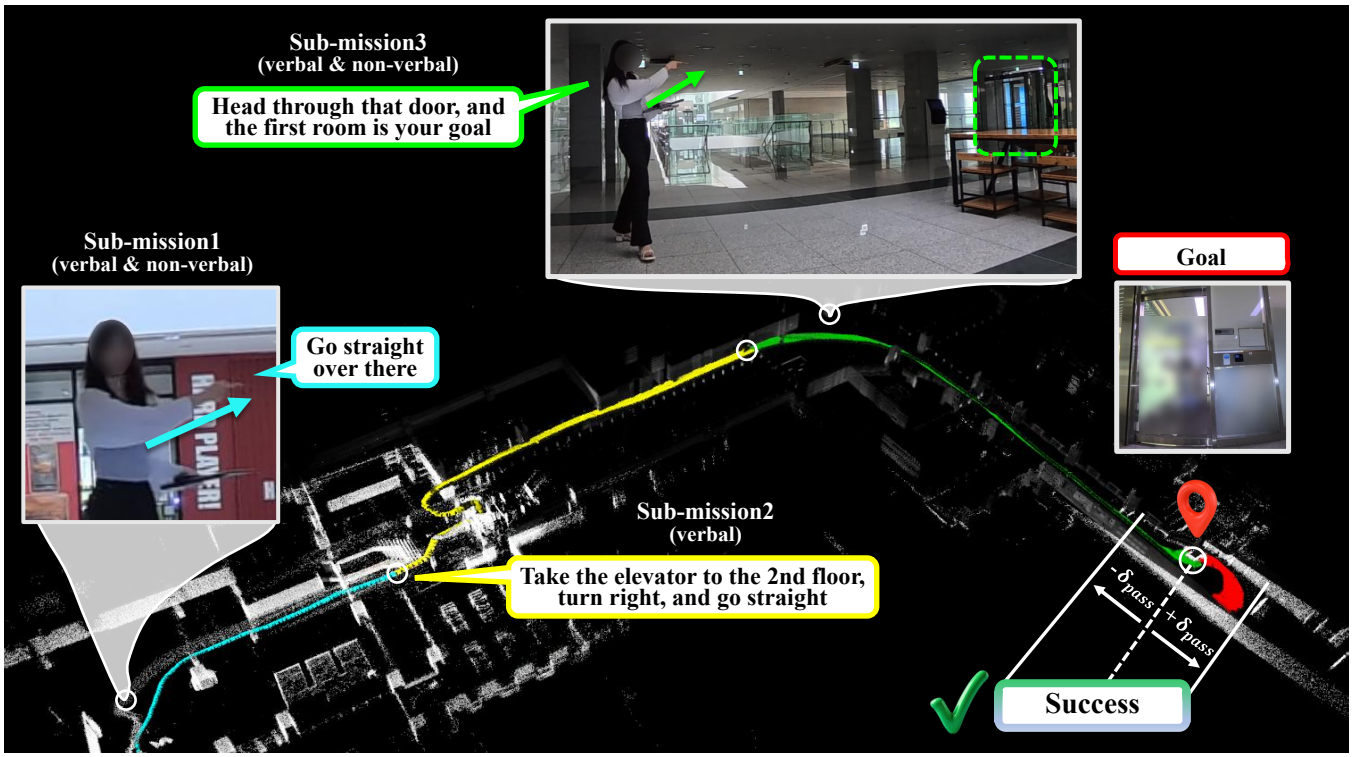

Fig. 3: **HINT task in the SocialACT_030 mission** The mission consists of three sub-missions teleoperated to collect sensor data, visualized on the SLAM map with different colored trajectories for each sub-mission. Each sub-mission ends once the corresponding actions are completed, and the next one begins. The mission is considered successful if the agent outputs the goal_signal within an $\varepsilon_{\text{goal}}$ time window of reaching the destination; otherwise, it is marked as a failure.

into a cubemap frames with six faces (front, right, left, back, bottom, top). All cubemap and panoramic frames are time-synchronized with the raw sensor data on a per-second basis and integrated into a single bag file for each mission to facilitate convenient use of our dataset by robotics researchers. To obtain the ground-truth action sequence for each mission, we employ an state-of-the-art odometry method, KISS-ICP [25], which generates SE(3) odometry.

Data labeling. We label sub-timestamps with one-second precision at sub-mission boundaries and annotate a human-like wayfinding instruction for each sub-mission. Four annotators write the instructions to preserve human-like brevity and clarity while indicating the intended route. A sub-mission ends when the given instruction has been completed, at which point a new sub-mission begins with the next instruction. This procedure yields paired sequences that connect guidance units to the robot's time synchronized observations and actions throughout the mission.

E. Dataset Statistics

We partition the 100 SocialACT missions into three difficulty levels according to the prevalence of non-verbal guidance. The *easy* split contains 40 missions that use only verbal instructions. The *medium* split contains 30 missions, each with exactly one non-verbal sub-mission while the remaining segments are verbal. The hard split consists of the remaining 30 missions, with an average of 3.1 non-verbal sub-missions per mission. This stratification highlights instruction modality and its effect on navigation performance. It also ensures balanced evaluation across difficulty levels, enabling fair comparisons and clear ablations.

IV. Potential Research Topics

To successfully address the HINT task, an agent must be able to leverage commonsense knowledge to infer the missing context from abstract instructions and navigate accordingly. Another promising research direction, as illustrated in Fig. 2, is socially interactive navigation: envisioning robots that recognize when humans may require assistance, proactively suggest the appropriate assistance, and navigate according to human instructions. Realizing this direction will require developing capabilities to interpret diverse social contexts and to infer implicit human needs.

V. Conclusion

We introduce the Human-like INsTruction–grounded Navigation (HINT), task and construct SocialACT, a real-world dataset designed to support and evaluate this task. SocialACT comprises 100 missions totaling 5 h 14 min, with 70 goals and 123 non-verbal sub-missions, and includes 1.32M image frames across panoramic and cubemap views. The dataset aligns verbal and non-verbal instructions with time synchronized low level robot actions at sub-instruction resolution, which enables progress based evaluation beyond episode success through metrics such as progress success rate.

References

[1] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4247–4258. Curran Associates, Inc., 2020. 1

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[3] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1, 2

[4] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9979–9988, 2020. 1, 2, 3

[5] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32340–32352. Curran Associates, Inc., 2022. 1

[6] Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[7] Vishnu Sashank Dorbala, Sanjoy Chowdhury, and Dinesh Manocha. Can LLM's generate human-like wayfinding instructions? towards platform-agnostic embodied instruction synthesis. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 258–271, Mexico City, Mexico, June 2024. Association for Computational Linguistics. 1

[8] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022. 1, 2, 3

[9] 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, 2023. 1, 2, 3

[10] Amirreza Payandeh, Daeun Song, Mohammad Nazeri, Jing Liang, Praneel Mukherjee, Amir Hossain Raj, Yangzhe Kong, Dines h Manocha, and Xuesu Xiao. Social-llava: Enhancing robot navigation through human-language reasoning in social spaces, 2024. 1, 2, 3

[11] Anthony Francis, Claudia Pérez-d'Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. Principles and guidelines for evaluating social robot navigation algorithms. *ACM Transactions on Human-Robot Interaction*, 14(2):1–65, 2025. 1, 2

[12] Michael Joseph Munje, Chen Tang, Shuijing Liu, Zichao Hu, Yifeng Zhu, Jiaxun Cui, Garrett Warnell, Joydeep Biswas, and Peter Stone. Socialnav-SUB: Benchmarking VLMs for scene understanding in social robot navigation. In *ICRA 2025 Workshop: Human-Centered Robot Learning in the Era of Big Data and Large Models*, 2025. 1, 2, 3

[13] Christoforos Mavrogiannis, Francesca Baldini, Allan Wang, Dapeng Zhao, Pete Trautman, Aaron Steinfeld, and Jean Oh. Core challenges of social robot navigation: A survey. *ACM Transactions on Human-Robot Interaction*, 12(3):1–39, 2023. 2

[14] Reuth Mirsky, Xuesu Xiao, Justin Hart, and Peter Stone. Conflict avoidance in social navigation—a survey. *ACM Transactions on Human-Robot Interaction*, 13(1):1–36, 2024. 2

[15] Xuesu Xiao, Bo Liu, Garrett Warnell, and Peter Stone. Motion planning and control for mobile robot navigation using machine learning: a survey. *Autonomous Robots*, 46(5):569–597, 2022. 2

[16] 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. *IEEE Robotics and Automation Letters*, 2024. 2

[17] Amir Hossain Raj, Zichao Hu, Haresh Karnan, Rohan Chandra, Amirreza Payandeh, Luisa Mao, Peter Stone, Joydeep Biswas, and Xuesu Xiao. Rethinking social robot navigation: Leveraging the best of two worlds. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16330–16337. IEEE, 2024. 2

[18] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation, 2025. 2, 3

[19] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 2, 3

[20] Daeun Song, Jing Liang, Xuesu Xiao, and Dinesh Manocha. Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments. *IEEE Robotics and Automation Letters*, 2025. 2

[21] Yangzhe Kong, Daeun Song, Jing Liang, Dinesh Manocha, Ziyu Yao, and Xuesu Xiao. Autospatial: Visual-language reasoning for social robot navigation through efficient spatial reasoning learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025. 2

[22] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2020. 2

[23] Justin Hart, Reuth Mirsky, Xuesu Xiao, Stone Tejeda, Bonny Mahajan, Jamin Goo, Kathryn Baldauf, Sydney Owen, and Peter Stone. Using human-inspired signals to disambiguate navigational intentions. In *International Conference on Social Robotics*, pages 320–331. Springer, 2020. 2

[24] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks, 2024. 2

[25] Ignacio Vizzo, Tiziano Guadagnino, Benedikt Mersch, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. KISS-ICP: In Defense of Point-to-Point ICP – Simple, Accurate, and Robust Registration If Done the Right Way. *IEEE Robotics and Automation Letters (RA-L)*, 8(2):1029–1036, 2023. 3, 4