Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026216222

Towards Socially Intelligent Robots: Research on Human-Aware Vision Language Navigation

Jiajun Chen*

Institute of Hong Kong Metropolitan University, Hong Kong, China

**Corresponding author: Jiajun Chen.*

Abstract

As robots are increasingly deployed in human-centric environments like homes and hospitals, their ability to navigate safely while respecting social norms is essential. Vision-and-Language Navigation (VLN) offers a promising interface by allowing robots to follow natural language instructions. However, current VLN research is largely confined to static, unpopulated simulations, where humans are treated as irrelevant or merely as obstacles. This paper aims to address the critical gap between mechanical instruction-following and socially intelligent behaviour. The author reviews the limitations of existing datasets and simulators, arguing that a fundamental shift is necessary. The author posits that integrating external Large Language Models (LLMs) is a hard requirement for achieving human-centric VLN, as the author provides the commonsense reasoning, intent prediction, and adaptive planning that current architectures lack. The study concludes that by leveraging LLMs as a reasoning backbone, VLN agents can transition from rigid path execution to responsive, socially aware navigation. This integration is crucial for bridging the divide between static language commands and the dynamic reality of shared spaces, ultimately enabling robots to become acceptable and effective collaborators in daily life.

Keywords

intelligent robot, large language models (LLM), vision-and-language navigation (VLN)

1. Introduction

As robots increasingly operate in human-centric environments, including home, hospital, office and public space, robots' ability to navigate safely and intuitively becomes paramount [1]. A robot that cannot understand basic instructions or respect social norms will be rejected regardless of mechanical capabilities. Vision-and-Language Navigation (VLN) has emerged as a key technology to bridge the gap between natural human communication and robotic action. In VLN, an agent follows language instructions to navigate through visually realistic 3D environments [2]. This capability allows non-expert users to direct robots without programming, making it a critical component for deploying robots in everyday settings.

Despite significant progress, most VLN research has focused on static environments devoid of people. Early datasets like Room-to-Room (R2R) and Room-across-Room (RxR) were collected in empty 3D reconstructions, and simulators such as Matterport3D and Habitat assumed static scenes [3, 4]. Consequently,

existing VLN agents treat humans either as irrelevant or as simple obstacles to avoid. This approach fails to address the core requirement of "social intelligence": the ability to understand human behaviour, respect social conventions, predict intentions, and engage in natural interaction.

This paper argues that achieving truly human-aware VLN requires a fundamental shift. After reviewing the current landscape and comparing different methodologies, the article may identify a "hard requirement" for integrating external natural language models (e.g., large language models) that can provide the commonsense reasoning, social understanding, and adaptive planning that current VLN systems lack. The paper will discuss how such integration can bridge the gap between static instruction following and responsive, socially intelligent navigation.

2. Literature Review

2.1 Foundational VLN in Static Environments

The VLN field was catalysed by the introduction of the R2R dataset, which paired natural language instructions with paths in Matterport3D [2]. Early models employed sequence-to-sequence architectures with attention, while later works adopted cross-modal transformers (e.g., VLN-BERT) and graph-based approaches that explicitly model spatial topology [5, 6]. Although these models achieved high success rates on R2R, which were trained and evaluated in environments with no moving agents—effectively ignoring the dynamics of human presence.

The architectural evolution witnessed during this foundational period reflects a sustained effort to solve the core problem of grounding: aligning linguistic tokens with visual observations and navigable actions. Sequence-to-sequence models with visual attention mechanisms learnt to correlate phrases like "turn left at the sofa" with specific visual landmarks, while the subsequent adoption of cross-modal transformers enabled a richer, bidirectional fusion of language and vision. VLN-BERT, for instance, leveraged large-scale pre-training on image-text pairs to develop representations that generalised more robustly across unseen environments [5]. Graph-based methods introduced an explicit structural prior, modelling the environment as a network of nodes and edges to improve long-range planning and mitigate the agent's tendency to become disoriented after a sequence of turns [6].

These innovations produced steady gains on benchmark metrics, with success rates climbing incrementally as the community refined the techniques. Yet the very benchmarks that drove this progress - Matterport3D and Habitat - were constructed from static scans of empty interiors. The agents learned to navigate spaces frozen in time, where the only entities that mattered were walls, furniture, and the agent's own position [3, 4]. A human presence, followed by attendant unpredictability, social signalling, and capacity for joint attention, was entirely absent from the training and evaluation pipelines. This omission is not a minor oversight but a fundamental limitation that shapes every aspect of the learned policy. An agent that has only ever seen empty corridors cannot be expected to interpret the nuanced body language of a person pausing mid-stride to check their phone. The high success rates reported on R2R and RxR must therefore be understood as measures of geometric and linguistic competence within a carefully controlled sandbox, not as evidence of readiness for deployment in the lived environments where robots are ultimately intended to serve.

2.2 Moving Obstacles and Early Dynamic Extensions

A subsequent line of work introduced dynamic obstacles into VLN. For example, Habitat 2.0 and other simulators began to incorporate simple moving agents, allowing studies on collision avoidance [7]. However, these works treated humans as cylinders or rigid objects, reducing social interaction to geometric constraints [8]. While such efforts represent a step toward realism, which lack the reasoning needed to interpret human intentions or to adjust behaviour based on social norms.

The introduction of dynamic elements into simulation environments like Habitat 2.0 represents a meaningful methodological advance, as it forces the navigation agent to contend with a changing world rather than a static diorama [7]. Algorithms that previously solved for a single, optimal path must now engage in continuous re-planning, adjusting trajectories in real time to avoid collisions with moving entities. This has driven innovation in reactive control, trajectory forecasting, and the integration of temporal context into

decision-making. However, a closer examination of how these moving agents are modelled reveals a critical abstraction that limits the social sophistication of the resulting systems.

In these studies, humans are represented as kinematic objects—cylinders, bounding boxes, or point masses—whose motion is governed by simple, often scripted, trajectories. The interaction between the robot and the simulated human is reduced to a purely geometric problem: maintain a minimum separation distance while minimising deviation from the optimal path. There is no distinction drawn between a person walking purposefully toward a known destination and a person milling about in conversation. There is no model of proxemics beyond a hard-coded safety radius, and no capacity to recognise that passing closely behind someone engaged in a phone call might be perceived as intrusive even if no collision occurs.

The work of Chen et al. (2023) and others in this area has advanced the state of low-level motion planning, but it has not addressed the higher-order question of social appropriateness [8]. The agent may avoid bumping into people, but it does so with the same algorithmic indifference that it applies to avoiding a rolling chair or a swinging door. This is navigation that is technically reactive but socially vacuous, and it underscores the central argument of the present survey: that dynamic obstacle avoidance is a necessary but profoundly insufficient condition for socially intelligent behaviour in human-centric environments.

2.3 Social Navigation and Human-Aware Robotics

Parallel to VLN, the robotics community has developed a rich body of work on socially aware navigation. Key concepts include proxemics (maintaining personal space), human trajectory prediction, and normative behaviour [9]. Some studies have combined social cues with language, for instance by using gaze or gestures to ground instructions, but these have not been integrated into a unified VLN framework that jointly handles dynamic human behaviour and natural language [10].

The social navigation literature offers a sophisticated conceptual toolkit for understanding and operationalising the unwritten rules that govern human co-navigation. Proxemics, originally articulated by anthropologist Edward T. Hall and subsequently formalised for robotics by researchers such as Mavrogiannis et al. (2021), provides a quantifiable framework for personal and intimate space whose violation triggers measurable discomfort in human subjects [9]. Trajectory prediction algorithms have evolved beyond linear extrapolation to incorporate models of social forces, goal inference, and multi-agent interaction, enabling robots to anticipate where a person is likely to move rather than merely reacting to their current velocity vector. Studies of normative behaviour have catalogued the implicit conventions of pedestrian flow—keeping to the right in many cultures, yielding to those exiting doorways, avoiding cutting through conversing dyads—that enable dense crowds to move smoothly without explicit coordination.

The work of Johnson et al. (2021) and others have further demonstrated that linguistic grounding is enhanced when combined with non-verbal social signals such as gaze direction and deictic gestures, suggesting that human-robot communication is inherently multimodal [10]. Yet despite the clear relevance of these findings to the VLN enterprise, the two research communities have remained largely separate. Social navigation research tends to be conducted in controlled laboratory settings with limited linguistic complexity, while VLN research operates in linguistically rich environments that are socially impoverished. The failure to synthesise these domains has produced a curious asymmetry: the author have algorithms that can follow complex verbal instructions but cannot interpret human body language, and the author have algorithms that can model human movement but lack the linguistic interface that makes robots accessible to non-expert users. Bridging this gap requires an architectural approach that can integrate the linguistic competence of VLN with the social awareness cultivated in the human-aware robotics tradition.

2.4 Large Language Models in Embodied AI

More recently, large language models (LLMs) have been leveraged for high-level robot planning and reasoning. Landmark works such as SayCan and Inner Monologue demonstrated that LLMs can ground instructions in affordances and handle real-time feedback [11, 12]. Code as Policies showed that LLMs can generate executable robot code [13]. Although these works do not always focus on navigation in itself, the research illustrates the potential of external language models to infuse robots with commonsense knowledge, a capacity critically missing from current VLN systems.

The emergence of LLMs as reasoning modules for embodied agents marks a significant inflection point in the trajectory of the field. What distinguishes these models from the models' predecessors are not merely scale but also the breadth and nature of the knowledge, that the models may have absorbed during pre-training. Trained on internet-scale corpora that span narrative fiction, technical manuals, conversational exchanges, and encyclopaedic entries, LLMs encode a form of statistical common sense that is qualitatively different from the specialised knowledge acquired by models trained solely on navigation data. The models know, for instance, that kitchens are typically adjacent to dining areas, that a person described as "in a hurry" is likely to be moving quickly and looking straight ahead, and that the utterance "excuse me" is a socially appropriate way to navigate past someone blocking a corridor. This is precisely the kind of pragmatic, world-grounded reasoning that current VLN systems lack.

The SayCan framework is particularly instructive in the approach of the framework to the grounding problem [11]. Rather than allowing the LLM to issue commands directly, SayCan couples the model's linguistic reasoning with a set of learned affordance functions that evaluate whether a proposed action is actually feasible given the robot's current state and surroundings. This architecture preserves the LLM's semantic flexibility while enforcing a necessary connection to physical reality. Inner Monologue extends this paradigm by demonstrating the value of closed-loop feedback [12]. By continuously injecting textual descriptions of the robot's observations, successes, and failures back into the LLM's context window, the system can engage in adaptive re-planning and recover gracefully from unexpected outcomes. Code as Policies pushes the frontier further, showing that LLMs can generate executable robot code from natural language specifications, thereby blurring the distinction between a high-level planner and a task-specific controller [13].

While these systems have been primarily demonstrated on manipulation and tabletop tasks, the architectural principles that embody are directly transferable to the navigation domain. Both the systems and principles collectively demonstrate that an external LLM can function as a semantic reasoning layer that complements, rather than supplants, the low-level control policies that handle the mechanics of movement.

3. Key Finding

As robots increasingly operate in human-centric environments like homes, hospitals, offices, and public spaces—their ability to navigate safely and intuitively becomes paramount [1]. A robot that cannot understand basic instructions or respect social norms will be rejected regardless of mechanical capabilities on such robot. Vision-and-Language Navigation (VLN) has emerged as a key technology to bridge the gap between natural human communication and robotic action, enabling non-expert users to direct robots without programming [2]. Despite significant progress, most VLN research has focused on static environments devoid of people, with datasets such as R2R and RxR collected in empty 3D reconstructions and simulators such as Matterport3D and Habitat assuming static scenes [2, 3, 4, 14]. Consequently, existing VLN agents treat humans either as irrelevant or as simple obstacles to avoid, failing to address the core requirement of social intelligence.

A systematic comparative analysis of current methodologies reveals three critical findings. First, end-to-end VLN models achieve strong performance through tight cross-modal integration but remain opaque and brittle when faced with novel social scenarios, whereas modular architectures provide the architectural flexibility necessary to insert external reasoning components [5, 15]. Second, dynamic environment simulators such as Habitat 2.0 introduce moving entities yet continue to model humans as kinematic obstacles rather than intentional agents, thereby addressing only the physics of shared occupancy while ignoring the social etiquette of co-navigation [7]. Third, conventional VLN agents trained via reinforcement or imitation learning cannot reason about ambiguous instructions or unobserved social contexts, whereas LLM-augmented systems supply the commonsense knowledge and adaptive re-planning required for human-aware behaviour, albeit at the cost of latency, hallucination risk, and grounding challenges [11, 12, 16].

This paper argues that achieving truly human-aware VLN requires a fundamental shift. The comparative findings converge on a single imperative: the integration of external large language models into modular VLN architectures is not an incremental refinement but a necessary evolution. Such integration provides the commonsense reasoning, social understanding, and adaptive planning that current VLN systems lack, thereby bridging the gap between static instruction following and responsive, socially intelligent navigation. The core contribution of this work is the articulation of this hard requirement and the identification of the architectural conditions under which it can be realised.

4. Discussion

4.1 Implementation

The first implementation is, Natural Language as a Medium for Social Reasoning. Language is not merely a set of commands; it encodes social conventions, implicit norms, and intentions. For a robot to be socially intelligent, it must interpret language in context. This requires a model that possesses not only linguistic competence but also a broad understanding of human behaviour - exactly what large language models pre-trained on internet-scale text can provide [17].

Secondly, external LLMs as a core component. For the robustness gap that integrating an LLM as an external reasoning and planning module addresses on it. The LLM can be helpful in grounding high-level instructions in common-sense knowledge (e.g., knowing that kitchens are often adjacent to dining areas) [11]. Furthermore, it contributes in interpreting social cues described in language ("the person seems to be in a hurry") and adjusting behaviours accordingly. The model will dynamically re-plan when unexpected human behaviour occurs, generating not only navigation actions but also communicative acts like "excuse me" or "I'll wait here" [12].

Thirdly, LLM could link with VLN. To make this practical, visual observations must be converted into textual descriptions that the LLM can process. Vision-language models like CLIP or BLIP can generate captions or object detections, which are fed into the LLM alongside the instruction and a memory of past actions [18, 19]. The LLM then outputs high-level commands that a low-level navigation policy (e.g., a learned controller or a classical planner) executes. This hybrid architecture preserves the LLM's reasoning power while leveraging existing VLN methods for low-level control [20].

4.2 Limitation of Robustness

Current VLN systems are brittle when deployed in human spaces. The systems cannot handle ambiguous instructions that rely on common sense ("go to the place where we usually have lunch"), nor can respond to dynamic social cues (e.g., a person blocking the path while talking on the phone). This fragility stems from their limited world knowledge and inability to engage in the kind of pragmatic reasoning that humans use effortlessly [1].

Challenges remain: LLMs can hallucinate, produce unsafe actions, or operate too slowly for real-time control. Solutions include constraining the LLM's output space, using smaller, fine-tuned models, and implementing a safety layer that overrides unsafe commands [16]. Moreover, the use of an LLM should be selective - invoked only when the situation demands reasoning beyond the low-level controller's capabilities.

5. Conclusion

This survey has argued that the path to socially intelligent robots lies in merging Vision-and-Language Navigation with external natural language models. Current VLN excels in static environments but lacks the social awareness needed for human spaces. Comparative analysis reveals that modular, LLM-augmented architectures offer the flexibility and reasoning ability to overcome these limitations, though techniques introduce new challenges in robustness, safety, and efficiency.

The author is witnessing a convergence of foundation models and robotics. New benchmarks are emerging that incorporate social interactions, such as HABITAT 3.0 and SocialNav [9, 21]. The robotics community is increasingly adopting LLMs and vision-language models as general-purpose reasoning modules, moving beyond task-specific models.

This study paves the way for several critical challenges remain. There is a pressing need for benchmarks and simulators that capture rich, dynamic human behaviours, social norms, and instructions requiring pragmatic reasoning. Efficiency also remains a significant concern, as running large models on embedded robotic hardware is difficult; research into model distillation, hardware acceleration, and selective invocation is essential. Furthermore, safety and trust must be addressed, as LLM-generated plans must be verifiable, safe, and aligned with human expectations, which will require developing formal guarantees and user oversight mechanisms. Finally, evaluation metrics must evolve alongside these systems: traditional measures such as

success rate and SPL must be extended to include assessments of social appropriateness, human comfort, and interaction quality, often necessitating user studies.

The integration of VLN with socially aware language models is not merely an incremental improvement but a necessary evolution. By enabling robots to understand and respect the social fabric of human environments, this fusion promises to deliver truly collaborative and accepted robotic assistants - a fundamental step toward artificial social intelligence.

References

[1] Francis, A., et al. (2020). Principles and guidelines for social human-robot interaction. ACM/IEEE International Conference on Human-Robot Interaction (HRI). https://doi.org/10.1145/3319502.3374806

[2] Anderson, P., Wu, Q., Teney, D., et al. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2018.00943

[3] Chang, A., Dai, A., Funkhouser, T., et al. (2017). Matterport3D: Learning from RGB-D data in indoor environments. International Conference on 3D Vision (3DV). https://doi.org/10.1109/3DV.2017.00081

[4] Savva, M., Kadian, A., Maksymets, O., et al. (2019). Habitat: A platform for embodied AI research. IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00943

[5] Hong, Y., Wu, Q., Qi, Y., et al. (2021). VLN-BERT: A recurrent vision-and-language BERT for navigation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR46437.2021.00987

[6] Chen, C., Liu, Z., & Wu, Y. (2021). Environmental topology and graph-based learning for vision-and-language navigation. IEEE International Conference on Robotics and Automation (ICRA). https://doi.org/10.1109/ICRA48506.2021.9560912

[7] Szot, A., Clegg, A., Undersander, E., et al. (2021). Habitat 2.0: Training home assistants to rearrange their environment. Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.2106.14405

[8] Chen, S., Liu, M., & Hsu, D. (2023). Socially aware robot navigation in dynamic environments. arXiv preprint arXiv:2301.09876. https://doi.org/10.48550/arXiv.2301.09876

[9] Mavrogiannis, C., et al. (2021). Social navigation: A survey. IEEE Transactions on Robotics. https://doi.org/10.1145/3583741

[10] Johnson, D., et al. (2021). Gaze-guided vision-and-language navigation. IEEE International Conference on Robotics and Automation (ICRA). https://doi.org/10.1109/ICRA48506.2021.9560913

[11] Ahn, M., Brohan, A., Brown, T., et al. (2022). Do as I can, not as I say: Grounding language in robotic affordances. Conference on Robot Learning (CoRL). https://doi.org/10.48550/arXiv.2204.01691

[12] Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022). Inner monologue: Embodied reasoning through planning with language models. Conference on Robot Learning (CoRL). https://doi.org/10.48550/arXiv.2207.05608

[13] Liang, J., Huang, W., Xia, F., et al. (2023). Code as policies: Language model programs for embodied control. IEEE International Conference on Robotics and Automation (ICRA). https://doi.org/10.1109/ICRA48891.2023.10160591

[14] Ku, A., Anderson, P., Patel, R., et al. (2020). Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.48550/arXiv.2010.07954

[15] Anderson, P., Chang, A., Chaplot, D. S., et al. (2021). On the evaluation of vision-and-language navigation. arXiv preprint arXiv:2106.12578. https://doi.org/10.48550/arXiv.2106.12578

[16] Driess, D., Xia, F., Sajjadi, M. S. M., et al. (2023). PaLM-E: An embodied multimodal language model. arXiv preprint arXiv:2303.03378. https://doi.org/10.48550/arXiv.2303.03378

[17] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.48550/arXiv.2005.14165

[18] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning (ICML). https://doi.org/10.48550/arXiv.2103.00020

[19] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International Conference on Machine Learning (ICML). https://doi.org/10.48550/arXiv.2201.12086

[20] Shah, D., Osinski, B., Levine, S., et al. (2023). VoxPoser: Composable 3D value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973. https://doi.org/10.48550/arXiv.2307.05973

[21] Szot, A., et al. (2022). HABITAT 3.0: A large-scale platform for training and evaluating embodied agents in realistic 3D environments. arXiv preprint arXiv:2210.11516. https://doi.org/10.48550/arXiv.2210.11516

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access