# Perception–Awareness–Decision: Socially-Aware Robot Navigation and Interaction

Edison Jair Bejarano Sepulveda
edison.bejarano@upc.edu
Institut de Robòtica i Informàtica Industrial
Barcelona, Spain

Valerio Bo
valerio.bo@upc.edu
Institut de Robòtica i Informàtica Industrial
Barcelona, Spain

Alberto Sanfeliu
alberto.sanfeliu@upc.edu
Institut de Robòtica i Informàtica Industrial
Barcelona, Spain

Anaís Garrell
anais.garrell@upc.edu
Institut de Robòtica i Informàtica Industrial
Barcelona, Spain

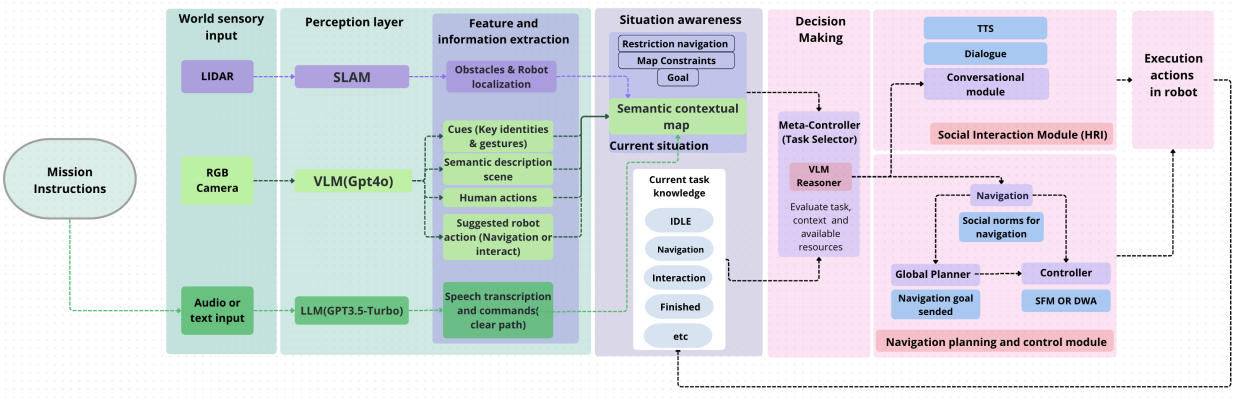


**Figure 1: Perception-awareness-decision (PAD) framework for socially aware navigation.**

## Abstract

Robots working in spaces shared by people need more than geometric mapping: they must recognize people, understand social context, and decide whether to proceed or negotiate passage. Traditional navigation pipelines lack this semantic understanding, often failing when progress depends on human cooperation. We introduce a Perception–Awareness–Decision (PAD) framework that systematically combines Simultaneous Localization and Mapping (SLAM) with Vision–Language Models (VLMs), speech recognition, and Large Language Models (LLMs), rather than simply stacking modules. PAD tries to emulate human perceptual organization by fusing multi-modal cues into a unified situational-awareness map capturing geometry, social context, and linguistic intent. This representation enables the decision layer to choose adaptively between safe replanning and context-appropriate verbal interaction. In a corridor-blocking task, PAD improves task success, increases safety margins, and produces behaviour that participants judged as more socially appropriate than a geometric baseline. These findings offer preliminary evidence that combining VLM-derived semantics with structured situational awareness can support more socially aware robot navigation.

## CCS Concepts

• **Computing methodologies** → *Cognitive robotics.*

## Keywords

Socially Aware Navigation, Dynamic Semantic Mapping, Foundation Models, Perception–Awareness–Decision

**ACM Reference Format:**
Edison Jair Bejarano Sepulveda, Valerio Bo, Alberto Sanfeliu, and Anaís Garrell. 2026. Perception–Awareness–Decision: Socially-Aware Robot Navigation and Interaction. In *Proceedings of 2026 ACM/IEEE International Conference on Human-Robot Interaction (HRI '26).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*HRI '26, Edinburgh, Scotland, UK*
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Robots navigating human shared environments must reach their goals while also respecting social conventions. Classical navigation pipelines, which optimize geometric metrics, often miss the subtler dynamics that shape human behavior. Foundation Models (FMs) promise richer perception and reasoning: Large Language

Models (LLMs) can clarify goals or negotiate with users [1], and Vision–Language Models (VLMs) provide semantic grounding for people, objects, and scene attributes [2]. Yet these models lack physical grounding and do not reliably enforce geometric or temporal constraints, which can lead to unsafe or implausible plans when used in isolation.

Hybrid techniques such as NaviWM [3] and SocialNav [4] partially address these challenges by combining foundation model reasoning with traditional mapping and planning. However, they lack an explicit intermediate representation that jointly integrates spatial structure, semantic cues, and predictive information. Similarly, cognitive models such as Perception–Intention–Awareness (PIA) [5] emphasize structured processing but do not integrate FM-based reasoning with situational awareness for navigation and interaction. To address these limitations, we propose the *Perception–Awareness–Decision* (PAD) framework, which explicitly separates multimodal perception, situational awareness, and decision making. PAD fuses LiDAR, visual, linguistic, and mission-level inputs into a semantic contextual map that supports task selection across navigation, interaction, and safety behaviors, executed through either an LLM-driven dialogue module or a socially aware controller such as the Social Force Model [6]. By embedding FM reasoning within a structured cognitive pipeline rather than a direct perception-to-action loop, PAD enables more grounded and socially aligned robot behavior, allowing robots to decide when to converse, yield, or proceed in ways that better reflect human expectations.

## 2 Related Works

Research on socially aware robot navigation spans three complementary directions: foundation models for embodied intelligence, socially compliant motion planning, and cognitive frameworks for structured reasoning. Recent advances in Large Language Models and Vision–Language Models have enabled robots to interpret language and semantic scene context [1, 2]. Multimodal systems such as PaLM-E, RT-2, and LM-Nav demonstrate how combining linguistic and visual reasoning supports manipulation and long-horizon navigation. However, these approaches remain weakly grounded, limiting their ability to enforce geometric or temporal constraints. Hybrid methods such as NaviWM [3] and SocialNav [4] partially address this issue through world models or hierarchical planning, but still lack a unified representation that integrates geometric, semantic, and predictive information for downstream decision making.

In parallel, socially aware navigation has traditionally relied on explicit models of human motion and proxemics, most notably the Social Force Model [6, 7]. While recent systems augment these models with semantic cues, perception is often tightly coupled to action, leaving little room for intermediate reasoning or contextual aggregation. Cognitive frameworks such as situation awareness models [8] and the Perception–Intention–Awareness (PIA) cycle [5] emphasize modular processing and hierarchical reasoning, but do not integrate FM-based perception into a situational representation suitable for navigation and interaction.

In contrast, PAD introduces an explicit situational-awareness layer that serves as an interface between multimodal perception and downstream decision making. Rather than proposing a new world model, PAD binds SLAM-based geometry with VLM- and LLM-derived semantics in a compact contextual map that supports task-level switching between navigation and dialogue. Compared with PIA-like cognitive cycles, PAD operationalizes the awareness stage using FM-based perception, enabling robots to decide when to proceed, yield, or engage in interaction within a unified and repeatable pipeline.

## 3 System overview

This section introduces the *Perception–Awareness–Decision* (PAD) framework for socially aware navigation and interaction. As illustrated in Fig. 1, PAD is organized into three tiers Perception, Situation Awareness, and Decision Making connected through a shared contextual map that supports consistent reasoning and task selection.

### 3.1 Perception Layer

The perception layer transforms raw sensory data into structured geometric, semantic, and linguistic descriptors. LiDAR scans are processed through SLAM to estimate robot pose and occupancy for geometric grounding, while VLM-based perception detects humans, objects, gestures, and high-level scene semantics beyond geometry alone. In parallel, LLM-based language perception transcribes and interprets spoken or textual inputs to extract intentions and relevant contextual cues. Together, these components produce a multimodal representation that captures both the physical layout and the social meaning of the environment.

### 3.2 Situation Awareness Layer

This layer fuses geometric, semantic, and linguistic features into a unified *semantic contextual map* that integrates spatial structure, detected agents and actions, semantic scene descriptors, interpreted linguistic intentions, and mission-level metadata.

Following classical situational-awareness theory, the map supports: (1) *perception* of relevant elements, (2) *comprehension* of their affordances and social meaning, and (3) *projection* of future states such as human motion. This intermediate representation enables deliberate reasoning that is absent from direct perception-to-action pipelines.

### 3.3 Decision Making Layer

A layered meta-controller analyzes the contextual map and determines the robot's behavior mode based on mission objectives, social signals, and safety requirements. Control is delegated to:

- **Social Interaction Module:** An LLM-driven conversational system with text-to-speech capabilities for clarification, negotiation, or yielding requests, grounded in situational semantics.
- **Navigation and Control Module:** A global planner and a socially aware controller (e.g., Social Force Model or Dynamic Window Approach (DWA)), augmented with LiDAR-based safety checks to ensure compliant motion.

### 3.4 Information Flow and Coordination

SLAM and automatic speech recognition run continuously, while VLM-based semantic perception updates at a lower frequency due

to inference latency. Their outputs are fused into the semantic contextual map, which serves as the interface between perception and decision making and allows the meta-controller to switch dynamically between navigation and interaction behaviors. This modular flow supports asynchronous sensing, reduces latency through compact representations, and links multimodal perception with socially grounded decision making.

Although our experiment focuses on navigation and verbal interaction, the same decision-layer structure can naturally scale to additional behaviors such as yielding, waiting, multi-party negotiation, or more complex task switching.

## 4 Experimental Setup

We evaluate PAD in a corridor-blocking scenario designed to test whether the robot can resolve a navigation conflict that requires human cooperation. The study was conducted on a TIAGo++ (IVO) robot operating in an indoor corridor environment (Fig. 2). Each trial begins with the robot navigating toward a fixed goal until a participant blocks the corridor, after which the robot either replans (P0) or uses PAD to initiate context-grounded dialogue (P1/P2) before resuming navigation. All participants provided informed consent prior to the study, and data were collected and analyzed in anonymized form in accordance with institutional guidelines for human-subjects research.

### 4.1 Hypotheses

We evaluate the PAD framework through the following hypotheses:

- **H1:** PAD-enabled robots, integrating multimodal perception and structured situational awareness, resolve social navigation conflicts more reliably than purely geometric navigation strategies, as reflected in higher task-completion rates and fewer proxemic violations.
- **H2:** Dialogue conditioned on the situational-awareness map leads to higher human cooperation than strategies without contextual communication, highlighting the benefit of combining FM-based reasoning with explicit context representation.
- **H3:** In this corridor-blocking scenario, PAD-based behaviors (P1, P2) tend to be perceived as more natural and appropriate than SLAM-only navigation (P0), as reflected by higher subjective ratings and preference counts in our small participant sample.

### 4.2 Experimental Conditions

We evaluated three behavioral modes reflecting different levels of contextual reasoning and interaction ability. All conditions use identical sensing hardware and differ only in the extent to which they leverage the PAD architecture:

- **P0 – SLAM-only baseline.** The robot relies exclusively on LiDAR-based perception and classical navigation. It performs no semantic interpretation or dialogue and repeatedly attempts local replanning until failure.
- **P1 – Context-aware conservative interaction.** The PAD perception and situational-awareness layers are active. Upon detecting sustained blockage, the meta-controller triggers concise, cooperative dialogue grounded in the contextual map, prioritizing user comfort and cautious re-verification.
- **P2 – Context-aware assertive interaction.** As in P1, the robot uses full PAD perception and awareness but adopts a more goal-oriented interaction strategy incorporating justification or light persuasion, with escalation across turns when necessary.

### 4.3 Procedure and Measures

Eight participants completed one trial per condition in counterbalanced order. In each trial, the robot navigated toward the goal until blocked by the participant. For P1 and P2, the robot initiated dialogue to request clearance; for P0, it relied solely on replanning until success or failure. Interaction in P1/P2 was capped at $N_{\text{turns}} = 8$, after which the trial was considered unsuccessful.

We logged task success, trial duration, number of replanning attempts, and robot–human stopping distance. For interactive conditions, we also recorded dialogue turns and average robot tokens per turn as proxies for verbosity. After each trial, participants rated perceived clarity and naturalness and indicated their preferred strategy.

## 5 Results and Discussion

Both P1 and P2 achieved perfect task completion and were rated as providing clearer explanations (Fig. 3(a)). These results indicate

We evaluated robot performance across the three behavioral modes defined in Section 4: **P0** (navigation only), **P1** (context-grounded conservative interaction), and **P2** (context-grounded assertive interaction). Performance was assessed using task-level metrics—success rate, re-planning attempts, stopping distance, trial duration, number of dialogue turns, and average robot tokens per turn—summarized in Table 1. Tokens per turn are reported as a descriptive proxy for verbosity.

**Table 1: Performance metrics across behavioral modes. P0: navigation only, P1: conservative interaction, P2: assertive interaction. Stopping distance and trial duration are reported in meters and seconds; all other metrics are dimensionless.**

| Metric | P0 | P1 | P2 |
|---|---|---|---|
| Success rate | 2/8 | **8/8** | **8/8** |
| Re-planning attempts (avg.) | 7.25 | 4.71 | **4.29** |
| Stopping distance (m, avg.) | 0.70 | **0.83** | 0.79 |
| Trial duration (s, avg.) | 105.94 | **99.95** | 162.21 |
| Dialogue turns (avg.) | – | **3.4** | 6.2 |
| Tokens per turn (avg.) | – | 35.45 | **21.36** |

### 5.1 Overall Performance Trends

The SLAM-only baseline (P0) showed the lowest task success and required the most re-planning attempts. Lacking semantic cues or interaction, the robot repeatedly searched for alternative paths and often approached the participant too closely, highlighting the limitations of purely geometric navigation in socially constrained settings.

In contrast, both PAD-enabled behaviors (P1 and P2) achieved perfect success (8/8). Fewer re-planning attempts indicate that

**Figure 2: Example experimental trial. The robot detects the blocking human, initiates dialogue, verifies that clearance has been granted, and continues navigation toward the goal.**

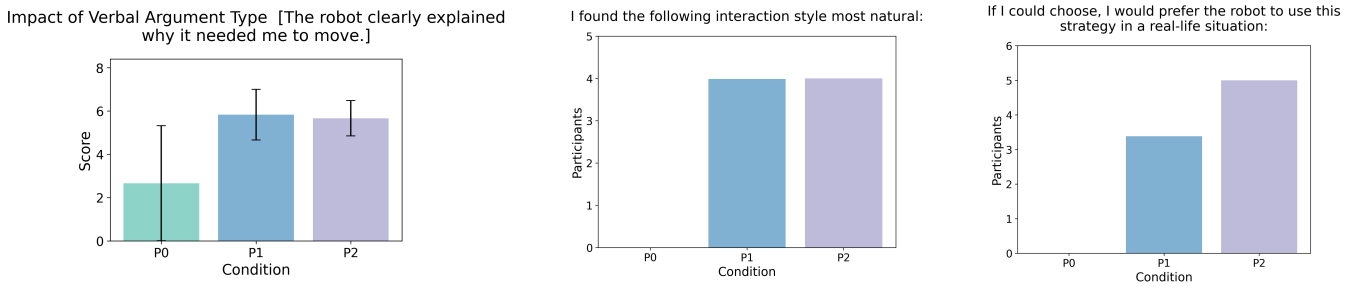


**Figure 3: Summary of subjective evaluation. (a) Clarity of the robot's explanation. (b) Perceived naturalness of the interaction style. (c) Preferred strategy for real-life robot behavior. Across all measures, participants tended to prefer PAD-enabled behaviors (P1, P2) over the SLAM-only baseline (P0).**

VLM-based perception enabled earlier detection of human blockage, preventing oscillatory behavior. Larger stopping distances further reflect safer and more socially aligned motion.

## 5.2 Differences Between Interaction Policies

Although P1 and P2 share the same perception and situational-awareness layers, their decision-making strategies led to distinct interaction patterns.

**P1 — Conservative interaction:** Produced longer utterances (35.45 tokens/turn) but required fewer dialogue turns (3.4), resulting in the shortest trials (99.95 s). Clear and empathetic explanations appear to have facilitated faster cooperation.

**P2 — Assertive interaction:** Generated shorter messages (21.36 tokens/turn) but required more dialogue turns (6.2), leading to longer trials (162.21 s). Although more goal-directed, this strategy sometimes required additional negotiation.

Despite these differences, both policies reliably resolved the scenario, demonstrating that PAD supports multiple conversational styles without compromising task success.

## 5.3 Hypotheses Validation

**H1 (Navigation effectiveness).** *Supported.* PAD-enabled behaviors (P1, P2) outperformed the SLAM-only baseline (P0) across all task-level metrics, including success rate, re-planning attempts, and stopping distance (Table 1).

**H2 (Context-grounded dialogue supports cooperation).** *Supported.* Both P1 and P2 achieved perfect task completion and were rated as providing clearer explanations (Fig. 3(a)), indicating that dialogue grounded in situational context improves human cooperation.

**H3 (Perceived naturalness and appropriateness).** *Supported as an observed trend.* Participants consistently preferred PAD-enabled behaviors over P0 in measures of naturalness and real-life suitability (Fig. 3(b),(c)). Although not statistically tested due to sample size, the observed pattern favors PAD-based navigation.

## 6 Conclusion

We introduced a Perception–Awareness–Decision (PAD) framework that structures multimodal perception and decision making for socially aware robot navigation. By combining geometric mapping with VLM-based semantic cues and language-driven interaction, PAD builds a situational map that helps the robot determine whether to continue navigating, wait, or initiate dialogue.

The corridor-blocking study illustrates the value of this structure. Compared with a SLAM-only baseline, PAD-based behaviors achieved higher task success, fewer repeated re-planning attempts, and safer distances when approaching a person. Although the two interaction policies differed in tone and efficiency, both reliably resolved the scenario, and participants consistently favored them over the baseline. These findings suggest that integrating semantic perception with explicit situation awareness can support more appropriate and effective robot behavior in human environments. Future work will examine richer multi-party interactions and explore broader decision-making behaviors beyond those required for the navigation task studied here.

## Acknowledgments

This work was supported by JST Moonshot R&D (grant JPMJMS2011) and European project TORNADO (HORIZON-CL4-2024-DIGITAL-EMERGING-01-101189557).

## References


[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[3] Weizheng Wang, Obi Ike, Soyun Choi, Sungeun Hong, and Byung-Cheol Min. Deductive chain-of-thought augmented socially-aware robot navigation world model, 10 2025.

[4] Ziyi Chen, Yingnan Guo, Zedong Chu, Minghua Luo, Yanfen Shen, Mingchao Sun, Junjun Hu, Shichao Xie, Kuan Yang, Pei Shi, Zhining Gu, Lu Liu, Honglin Han, Xiaolong Wu, Mu Xu, and Yu Zhang. Socialnav: Training human-inspired foundation model for socially-aware embodied navigation, 11 2025.

[5] J. E. Domínguez-Vidal, Nicolás Rodríguez, and Alberto Sanfeliu. Perception–intention–action cycle in human–robot collaborative tasks: The collaborative lightweight object transportation use-case. *International Journal of Social Robotics*, 17:1927–1956, 2025. Published online March 25, 2024.

[6] Gonzalo Ferrer, Anaís Garrell, Fernando Herrero, and Alberto Sanfeliu. Robot social-aware navigation framework to accompany people walking side-by-side. *Autonomous Robots*, 41(4):775–793, 2017.

[7] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, 1995.

[8] Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995.


Received 08 December 2025; revised 12 March 2009; accepted 5 June 2009