SPATIALLY AND SOCIALLY GROUNDED COMMUNICATION FOR HUMAN-ROBOT COLLABORATION WITH TRUST AND RAPPORT

BY

YOHEI HAYAMIZU

B*S*, Iwate University, Japan, 2018
M*S*, University of Electro-Communications, Japan, 2021

DISSERTATION

Submitted in partial fulfillment of the requirements for
the degree of Computer Science
in the Graduate School of
Binghamton University
State University of New York
2026

Accepted in partial fulfillment of the requirements for
the degree of Computer Science
in the Graduate School of
Binghamton University
State University of New York
2026

Month DD, YYYY

Zhang, Shiqi, Committee Chair
School of Computing, Binghamton University

Sikdar, Sujoy, Committee Member
School of Computing, Binghamton University

Zhang, Zhongfei (Mark), Committee Member
School of Computing, Binghamton University

Yu, Zhou, Committee Member
Computer Science Department, Columbia University

Abstract

As robots are increasingly deployed in shared physical spaces, communication grounded in spatial and social context has become critical for human-robot collaboration. Recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have enabled natural language generation and visual scene understanding. However, these models remain disembodied, lacking awareness of the physical environment and individual preferences that shape interactions. Beyond language capabilities, mobile manipulation is a fundamental capability for robots to assist humans, but existing systems, while achieving efficient task execution, lack sufficient communication capabilities to adapt to social contexts. This gap hinders robots from performing contextually appropriate behaviors and building *trust* (the belief that the robot reliably performs its task) and *rapport* (the sense that the robot understands the user) based on situational awareness. This dissertation addresses this challenge by treating communication and mobile manipulation not as separate tasks but as a unified decision-making problem, establishing a comprehensive framework for spatially and socially grounded dialogue. The dissertation presents three integrated contributions. First, a joint policy learning platform that jointly optimizes verbal and physical actions learns spatially appropriate utterance timing through deep recurrent reinforcement learning, validated on both simulated and real robot platforms. Second, foundational techniques for spatially grounded dialogue and mobile manipulation—including a verbalization system for a

robotic guide dog that assists visually impaired users, and LLM-GROP for visually grounded task-and-motion planning—demonstrate how perceptual information can be transformed into meaningful spatial language and action in real-world environments. Third, the Spatial-Social Language Synthesis (SSLS) framework integrates these foundations with user-persona adaptation and an explicit Communication Strategy Selector, empirically validated through three experiments: a controlled video-based human evaluation with 30 raters, a scaled LLM-based validation over 300 dialogue samples, and an interactive real-robot study with participants in a simulated real estate tour, supplemented by post-hoc measurement of trust and rapport using the Multi-Dimensional Measure of Trust (MDMT) and the Connection-Coordination Rapport Scale. Together, these contributions demonstrate that grounding robot dialogue in both the physical environment and the individual user's context is not merely a performance enhancement but an indispensable condition for robots to build sustained trust and rapport as collaborative partners in human-inhabited spaces.

Contents

List of Tables

List of Figures

1 Introduction

Service robots are rapidly being deployed in shared spaces where humans operate, such as homes, public facilities, and care settings. While recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have enabled robots to generate fluent dialogue and verbalize visual information, even these highly capable systems struggle to integrate their communication with the physical constraints of mobile manipulation into a unified decision-making process grounded in the personal and social context of interactions. Existing grounding techniques have primarily focused on describing physical scenes, yet they face significant challenges in capturing the underlying intentions of interactions and in building ongoing trust and rapport with humans. This dissertation establishes a framework for grounding dialogue in both spatial and social contexts, treating dialogue and physical actions as an integrated problem to achieve the trust and rapport essential for human-robot collaboration in shared environments. This chapter elaborates on the need to connect physical and verbal actions in spatial and social contexts, defines the two central relational constructs of trust and rapport, summarizes the main contributions, and describes the organization of the dissertation.

1.1 Motivation

While LLMs and VLMs have enabled fluent conversation, there remains a deep gap between decision-making about *when, where, and what to say* and the physical actions

that service robots must perform in real-world settings. Current foundation models excel at processing abstract knowledge but lack the ability to ground dialogue in the dynamic spatiotemporal context that shapes interactions, such as spatial situations and individual preferences. In particular, when collaborating with humans during physical tasks such as navigation and manipulation, misaligned dialogue timing and content that do not synchronize with the surrounding spatial context can lead to user confusion and significantly undermine the system's reliability. We identify three tightly coupled challenges that motivate this dissertation.

First, tight integration of physical and verbal actions is needed. Traditional robotic systems often treat navigation and dialogue generation as separate modules, leading to dialogue content that may contradict the current spatial situation or miss appropriate timing. To achieve true collaboration, robots need to optimize *when, where, and what to say* as an integrated part of their physical action planning.

Second, the lack of techniques to transform perceived visual information into meaningful spatial context for humans hinders smooth assistance. While existing VLMs can describe objects and scenes, they do not provide sufficient spatial insight into what that means in specific assistive contexts (e.g., guiding visually impaired users or navigating complex home environments). Robots need to not only recognize "there is an obstacle ahead" but also interpret the spatial meaning and appropriately verbalize it as information directly related to user safety and convenience.

Third, and most importantly for this dissertation, existing systems lack mechanisms to build *trust* and *rapport* on these spatial and perceptual foundations. For robots to be accepted in society and to function as long-term partners, it is essential not only to perform tasks accurately but also to adapt their behavior to users' personal preferences

and the prevailing social context. A human's willingness to rely on and engage with a robot emerges only when the robot explains its intentions in a way that is grounded in the environment and that reflects the user's history and preferences.

Based on these motivations, this dissertation establishes a comprehensive framework for dialogue grounded in both spatial and social contexts under the physical constraints of mobile manipulation. To address this series of challenges, we set the following three research questions (RQs):

- RQ1: How can robots integrate navigation, manipulation, and dialogue into a single decision-making process and execute it with spatially appropriate timing?
- RQ2: How can robots transform perceived environmental information into meaningful spatial context for humans and express it in natural language?
- RQ3: Building on techniques for spatial and social grounding, how can robots adapt to individual preferences and situations to build trust and rapport with humans?

1.2 Defining Trust and Rapport

Because *trust* and *rapport* are often used interchangeably in everyday language, we adopt operational definitions drawn from the human-robot interaction (HRI) and social psychology literatures and use them consistently throughout this dissertation. A concise definition is offered here to make the remainder of the introduction self-contained; a full treatment of their antecedents, formation mechanisms, and evaluation instruments is deferred to Chapter 2, Section 2.4.2.

Trust is primarily *competence-based*: a human trusts a robot to the extent that they believe the robot is capable of reliably performing the task for which it is responsible [1, 2,

3]. Trust is calibrated through accumulated evidence of the robot's behavior—successes build trust, failures erode it—and is modulated by the robot's perceived predictability and transparency. Crucially, trust is not binary but a continuous, dynamic quantity that evolves over the course of an interaction and across repeated encounters, with both overtrust (assuming greater capability than the robot possesses) and undertrust (failing to utilize a capable robot) being costly outcomes that robot behavior must guard against.

Rapport, by contrast, is fundamentally *relationship-based* [4, 5]: it refers to the sense of mutual attentiveness, positivity, and coordination that characterizes harmonious human-robot interaction. Rapport is established not primarily through task performance but through the quality of the communicative interaction itself—through expressions of interest, acknowledgment of the human's situation, conversational alignment, and the perception that the robot "understands" the user's preferences. Where trust asks *"can the robot do the job?"*, rapport asks *"does the robot understand me?"*.

The relationship between trust and rapport is synergistic but *asymmetric*: competence-based trust is a necessary precondition for rapport to develop—a robot that consistently fails its task will not sustain positive engagement—but rapport, once established, amplifies trust by making the robot's occasional errors more forgivable and its explanations more credible. This asymmetry directly motivates the three-contribution structure of this dissertation: Contributions 1 and 2 establish the competence foundation (reliable task execution with spatially grounded communication), while Contribution 3 builds the rapport layer on top. Throughout this dissertation, trust is evaluated using validated HRI instruments including the Trust in Automation scale and the Multi-Dimensional Measure of Trust (MDMT) [6], and rapport is evaluated using the Connection-Coordination Rapport (CCR) Scale [7], as detailed in Chapter 2.

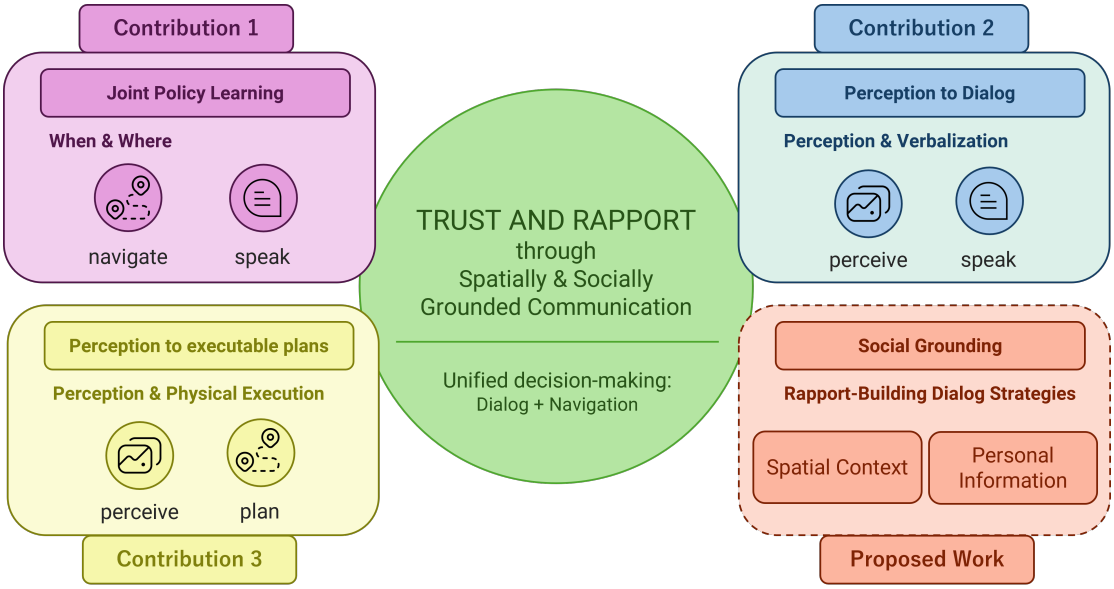


Figure 1.1: This dissertation focuses on spatially and socially grounded dialogue and navigation. Robots can ground communication in the environment, building the trust and rapport necessary for human-robot collaboration.

1.3 Main Contributions

In response to the research questions outlined above, this dissertation establishes a comprehensive framework for spatially and socially grounded dialogue through three main contributions, summarized in Figure 1.1. The contributions build on one another: the first two establish foundational techniques for integrated control of actions and dialogue and for spatial grounding of visual information, and the third leverages these foundations to develop an adaptive strategy for trust and rapport building.

The first contribution, *Joint Policy Learning for Dialogue and Co-Navigation*, addresses RQ1. Whereas traditional robotic systems often treat navigation and dialogue generation as separate modules, this work treats human-robot dialogue and co-navigation as a composite problem and develops a framework that jointly optimizes the two using Deep Reinforcement Learning (DRL). By treating shared spatial awareness in the human-robot team as a factor influencing dialogue states, the framework learns to maximize efficiency and accuracy in tasks such as guidance and information delivery. Exper-

iments demonstrated superior task-completion rates and execution times compared to modular baselines, and real-world validation confirmed the ability to produce dialogue with appropriate timing in response to surrounding spatial conditions. This contribution provides a foundation for integrated control of embodiment and communication, published at IROS 2023.

The second contribution, *Enabling Approaches for Spatial Awareness for Rich Communication*, addresses RQ2 by establishing foundational techniques that integrate visual information with language models to ground perceptual information in action planning and dialogue, and by validating them in real-world settings. We developed a dialogue system that verbalizes dynamically changing environmental features and action plans for a robotic guide dog that assists visually impaired users, enabling collaborative decision-making between humans and robots. We also developed LLM-GROP, a framework that integrates LLM commonsense knowledge into Task and Motion Planning (TAMP) for complex object rearrangement in mobile manipulation, with real-world experiments demonstrating that LLM-based strategies for selecting base positions operate efficiently and robustly under real-environment uncertainty. Together, these results provide the technical and empirical foundation for spatially grounded assistance and manipulation in the real world, published at AAAI 2026 and in IJRR 2025.

Prior work has shown that rapport significantly influences task success across a wide range of social interactions, including negotiation, education, and caregiving [8, 5], yet the mechanisms by which rapport is built through spatially grounded dialogue in mobile-robot settings remain underexplored. As defined in Section 1.2, trust [9, 2] and rapport [4] represent distinct but complementary dimensions of successful human-robot collaboration that existing systems have yet to adequately address. The third and

core contribution of this dissertation is the development and empirical validation of an adaptive dialogue strategy grounded in spatial and social contexts (RQ3). This contribution integrates the foundational techniques above, combining personal context—such as individual preferences and past interaction history—with spatial situational awareness to build trust and rapport with users. We realize this contribution as the Spatial-Social Language Synthesis (SSLS) framework and evaluate it through three experiments: a controlled video-based human evaluation with 30 raters across four system conditions, a scaled LLM-based validation over 300 dialogue samples, and an interactive real-robot study in a simulated real estate tour. To directly connect improvements in dialogue quality to relational outcomes, we further measure trust and rapport using the Multi-Dimensional Measure of Trust (MDMT) and the Connection-Coordination Rapport (CCR) Scale. The results demonstrate that spatial context and user-persona integration substantially improve dialogue quality along naturalness, relevance, persuasiveness, and coherence, and that the explicit, interpretable Communication Strategy Selector in SSLS provides deployment-critical transparency in addition to this quality improvement.

Together, the three contributions integrate advanced language processing capabilities with the physical actions of mobile manipulation through spatial and social contexts, demonstrating that grounding robot dialogue in both the surrounding environment and the personal context of the user is crucial not only for improving task performance but also for building the sustained trust and rapport essential for robots as long-term collaborative partners.

1.4 Dissertation Structure

The remainder of this dissertation is organized as follows: **Chapter 2** reviews the theoretical background and defines the key terms and evaluation instruments in the technical domains underlying this research, including a full treatment of trust and rapport. **Chapter 3** details the integrated learning of dialogue and navigation, demonstrating the acquisition and validation of appropriate dialogue timing based on surrounding spatial situations. **Chapter 4** describes the grounding of visual perception and language into real-world action planning, empirically demonstrating how perception and reasoning connect to physical actions. **Chapter 5** integrates these foundational techniques to present and empirically validate the Spatial-Social Language Synthesis (SSLS) framework—an adaptive dialogue strategy grounded in spatial and social contexts—through controlled human evaluation, LLM-based validation, and interactive real-robot evaluation with direct measurement of trust and rapport. **Chapter 6** concludes the dissertation and outlines directions for future work.

2 Background

Realizing a robot capable of collaborating with humans in shared physical spaces requires expertise spanning several distinct yet deeply intertwined research areas. This proposal sits at the intersection of three technical pillars—*decision-making*, *perception and spatial semantics*, and *social robotics*—each of which contributes an indispensable layer to the overall framework. This chapter surveys the theoretical and empirical foundations of each pillar, establishes the notation and terminology used throughout the proposal, and identifies the specific gaps that motivate the three main contributions.

Section 2.1 lays the mathematical foundations for sequential decision-making under uncertainty. We employ the Markov Decision Process (MDP) and its partially observable extension (POMDP) as the formal foundation for reasoning about actions and utterances under uncertainty, and review the Reinforcement Learning (RL) algorithms—Deep Q-Networks (DQN), Proximal Policy Optimization (PPO), and Recurrent Neural Networks for POMDPs (DRQN and Recurrent PPO)—that make these formalisms achievable in high-dimensional robotic settings. Section 2.2 then establishes the linguistic and architectural foundations of dialogue systems and natural language understanding, tracing the development from modular task-oriented pipelines through neural end-to-end models to Large Language Models (LLMs), and identifying the limitations of current approaches for embodied, spatially grounded interaction. Section 2.3 introduces Task and Motion Planning (TAMP), which bridges symbolic task-level decisions

(e.g., *"pick up the mug then place it on the table"*) with geometric motion-level execution under the kinematic and workspace constraints of mobile manipulation. Finally, Section 2.4 examines social robotics theory, covering proxemics, the psychological constructs of trust and rapport, and the evaluation metrics—both task-performance measures and human-centered scales—used to assess the quality of human-robot collaboration.

Together, these five sections map directly onto the research questions introduced in Chapter 1: decision theory and dialogue systems underpin RQ1 and RQ3; perception and TAMP underpin RQ2; and social robotics theory provides the evaluative lens for RQ3. Readers already familiar with a particular area may proceed directly to the subsections most relevant to their background, guided by the cross-references embedded throughout.

2.1 Foundations of Decision-Making Under Uncertainty

Robots operating in shared human spaces must continuously make decisions about *what to do* and *what to say*, often without complete knowledge of the current state of the world or of their human partner. This section formalizes the mathematical basis that unifies these two types of decisions into a single sequential decision-making process. We first introduce the Markov Decision Process (MDP) as the canonical framework for sequential decision-making under uncertainty, then extend it to the Partially Observable Markov Decision Process (POMDP), which is essential for capturing the fundamental information asymmetry present in human-robot interaction (HRI). Finally, we discuss how RL realizes these formalisms for learning joint policies in embodied dialogue systems.

2.1.1 Markov Decision Processes

A *Markov Decision Process* [10] provides a principled framework for sequential decision-making in stochastic environments. Intuitively, an agent repeatedly observes the current state of the environment, selects an action, receives a scalar reward signal, and transitions to a new state; the goal is to find a policy—a mapping from states to actions-that maximizes the expected cumulative reward over time.

Definition 1 (Markov Decision Process)**.** *An MDP is defined by the tuple*

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle$$

, where: $\mathcal{S}$ *is the* state space*, representing all possible configurations of the environment and the agent;* $\mathcal{A}$ *is the* action space*, the set of all decisions available to the agent;* $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ *is the* transition function*, where* $T(s, a, s') = P(s_{t+1} = s' \mid s_t = s, a_t = a)$ *denotes the probability of transitioning to state* s' *after taking action* a *in state* s*;* $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is the* reward function*, where* $R(s, a)$ *quantifies the immediate utility of taking action* a *in state* s*; and* $\gamma \in [0, 1)$ *is the* discount factor*, which trades off the importance of immediate rewards versus future rewards.*

The *Markov property* is central: $P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots) = P(s_{t+1} \mid s_t, a_t)$. That is, the future is conditionally independent of the past given the present state, which enables tractable planning.

A *policy* $\pi : \mathcal{S} \to \mathcal{A}$ or $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ specifies the agent's behavior. The quality of a policy is measured by the *state-value function* and the *action-value function* (Q-function).The *Bellman expectation equations* [10] express the recursive relationship

between value functions:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{s' \in \mathcal{S}} T(s, a, s') \big[R(s, a) + \gamma \, V^{\pi}(s') \big], \tag{2.1}$$

$$Q^{\pi}(s, a) = \sum_{s' \in \mathcal{S}} T(s, a, s') \big[R(s, a) + \gamma \, V^{\pi}(s') \big]. \tag{2.2}$$

The *optimal policy* π^* satisfies $V^{\pi^*}(s) \geq V^{\pi}(s)$ for all $s \in \mathcal{S}$ and all policies π. It can be recovered via the *Bellman optimality equations*:

$$V^*(s) = \max_{a \in \mathcal{A}} \sum_{s'} T(s, a, s') \big[R(s, a) + \gamma V^*(s') \big], \tag{2.3}$$

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \, Q^*(s, a). \tag{2.4}$$

2.1.2 Partially Observable Markov Decision Processes

In practice, a robot cannot directly observe the full state of the world. Sensor noise causes uncertainty about physical configurations, and the internal states of human partners—their intent, emotional state, or preferences—are *inherently* unobservable. The *Partially Observable Markov Decision Process* [11, 12] extends the MDP to handle this fundamental information asymmetry.

Definition 2 (Partially Observable Markov Decision Process)**.** *A POMDP extends the MDP tuple with observation-related components:* $\mathcal{P} = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \mathcal{O}, Z, b_0 \rangle$*, where:* $\mathcal{O}$ *is the* observation space*, the set of percepts available to the agent (e.g., camera images, LiDAR scans, spoken utterances);* $Z : \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0, 1]$ *is the* observation function*, where* $Z(s', a, o) = P(o_t = o \mid s_t = s', a_{t-1} = a)$ *gives the probability of receiving observation* o *after arriving in state* s' *via action* a*; and* $b_0 \in \Delta(\mathcal{S})$ *is the* initial belief*, a probability distribution over states representing prior knowledge before*

any interaction begins.

Because the agent cannot observe s_t directly, it maintains a *belief state* $b_t \in \Delta(\mathcal{S})$, a sufficient statistic for all past observations and actions. Upon taking action a_t and receiving observation o_{t+1}, the belief is updated using Bayes' rule:

$$b_{t+1}(s') \;=\; \eta \cdot Z(s', a_t, o_{t+1}) \sum_{s \in \mathcal{S}} T(s, a_t, s')\, b_t(s), \tag{2.5}$$

where η is a normalization constant ensuring $\sum_{s'} b_{t+1}(s') = 1$.

In HRI, the unobservable components of $\mathcal{S}$ include the human's *intent* (where they want to go), *affect* (frustration or satisfaction), and *preference profile* (communication style, assistance level desired). The belief update, therefore, serves a dual role: it tracks both the physical state of the world and an evolving internal model of the human partner. This duality is central to Contribution 3 (Chapter 5), where belief in user preferences is leveraged to adapt rapport-building dialogue strategies across repeated interactions. The following two sections describe how deep reinforcement learning achieves both the MDP and POMDP settings without requiring an explicit model of T, Z, or R.

2.1.3 Deep Reinforcement Learning

When the transition function T and reward function R are unknown, as is typically the case in real-world HRI, the agent must *learn* an optimal policy from experience. *Deep Reinforcement Learning* (DRL) [13, 14] achieves this by combining the reward-maximization objective of RL with the representational power of deep neural networks, enabling agents to operate over raw, high-dimensional observations. This section introduces the two families of DRL algorithms employed in this proposal: *value-based* methods, represented by the Deep Q-Network (DQN), and *policy gradient* methods,

represented by the Actor-Critic architecture and Proximal Policy Optimization (PPO).

Deep Q-Networks (DQN) [15] approximate the action-value function with a deep neural network $Q_\theta(s, a)$, parameterized by weights θ, trained to satisfy the Bellman optimality equation Equation (2.3)). The network minimizes the mean-squared Bellman error over transitions sampled from an *experience replay buffer* $\mathcal{D}$:

$$\begin{aligned} \mathcal{L}_{\text{DQN}}(\theta) &= \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\Big[\big(y_t - Q_\theta(s_t, a_t)\big)^2\Big] \\ y_t &= R(s_t, a_t) + \gamma \max_{a'\in\mathcal{A}} Q_{\theta^-}(s_{t+1}, a') \end{aligned} \tag{2.6}$$

,where θ^- are the parameters of a periodically synchronized *target network* that stabilizes training by decoupling the bootstrapping target from the network being updated. At execution time, actions are selected ϵ-greedily: with probability ϵ a random action is taken for exploration, and otherwise the greedy action $\operatorname{argmax}_a Q_\theta(s, a)$ is chosen.

Policy gradient methods take a complementary approach by directly parameterizing the policy as $\pi_\theta(a \mid s)$ and maximizing the expected return $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_t \gamma^t R(s_t, a_t)]$ [16]. Those methods compute the policy gradient estimation and update the policy by gradient ascent:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k})\right]. \tag{2.7}$$

To reduce the high variance of Monte-Carlo return estimates, *Actor-Critic* methods introduce a learned value baseline: the *Actor* $\pi_\theta(a \mid s)$ proposes actions, while the *Critic* $V_\phi(s)$ estimates their expected return [17]. Policy updates are driven by the *advantage function* $A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t)$, approximated in practice by the one-step TD error $\delta_t = R(s_t, a_t) + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$.

Proximal Policy Optimisation (PPO) [18] is the specific Actor-Critic algorithm adopted in this proposal. It improves training stability by constraining how much the policy is allowed to change in a single update, via a clipped surrogate objective:

$$\mathcal{L}_{\text{PPO}}(\theta) \;=\; \mathbb{E}_t\Big[\min\Big(r_t(\theta)\,A_t,\;\text{clip}\big(r_t(\theta),1-\epsilon,1+\epsilon\big)\,A_t\Big)\Big]\,, \tag{2.8}$$

where $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ is the probability ratio between new and old policies, and $\epsilon \in (0,1)$ is a clipping threshold. The clipping prevents unbounded updates while still allowing meaningful improvements at each iteration.

2.1.4 Deep Reinforcement Learning for Partial Observability

The DQN and PPO formulations introduced above assume that the current observation o_t is a sufficient statistic for the underlying state s_t. In a POMDP, this assumption is violated by construction: a single observation cannot resolve the agent's uncertainty about hidden state variables. For a robot collaborating with a human, this means that the human's current intent, emotional state, and attention cannot be inferred from a snapshot observation alone—they must be estimated from the *history* of observations and actions accumulated over the interaction.

Two recurrent extensions of the above algorithms address this limitation by replacing memory-less network layers with Long Short-Term Memory (LSTM) units [19, 20], which maintain a hidden state h_t that compresses the observation history into a fixed-size vector serving as an implicit approximation to the POMDP belief state b_t. The policy or value function is then conditioned on h_t rather than on a single observation o_t, with the correspondence $h_t \approx b_t$ holding in the sense that both summarize the observation history up to time t.

Deep Recurrent Q-Networks (DRQN) [21] instantiate this idea in the value-based setting by replacing the final feedforward layer of DQN with an LSTM. DRQN has demonstrated strong performance in dialogue-state tracking [22, 23]. *Recurrent PPO* extends the Actor-Critic framework analogously: both the Actor $\pi_\theta(a_t \mid h_t)$ and the Critic $V_\phi(h_t)$ are conditioned on the LSTM hidden state, and the PPO clipped objective (Equation (2.8)) is applied over fixed-length rollout segments [24]. This combination preserves the stability and sample efficiency advantages of PPO while equipping the agent with memory over the interaction history. The concrete structure of the observation space $\mathcal{O}$ and the action space $\mathcal{A}$ varies across experimental settings and is defined precisely in each respective chapter.

The decision-theoretic foundations established in this section provide the learning basis for treating physical actions and communicative acts as a unified optimization problem. However, the formalism so far treats *communicative acts* as abstract, discrete or continuous outputs, without addressing what it means for a robot to *understand* and *generate* natural language in an interactive dialogue context. The next section surveys the field of dialogue systems and natural language understanding, establishing the linguistic and architectural foundations that give concrete meaning to the dialogue-act components of the joint policies developed throughout this dissertation.

2.2 Dialogue Systems and Natural Language Understanding

For a robot to collaborate naturally with humans, producing and interpreting language is not a supplementary component but a core capability. We trace the development of dialogue systems from early modular pipelines through neural end-to-end models to the current generation of large language models, examining at each stage

the capabilities gained and the limitations that remain for embodied, spatially grounded interaction.

2.2.1 From Task-Oriented to Open-Domain Dialogue Systems

Dialogue systems have historically been developed along two parallel tracks, each reflecting a different assumption about the scope of conversation. *Task-oriented dialogue systems* target well-defined domains—booking a flight, querying a database, guiding a user through a procedure—where the conversation is a means to accomplish a specific goal within a bounded state space. *Open-domain dialogue systems*, by contrast, aim for unrestricted, natural conversation without a predefined task, prioritizing fluency, coherence, and social engagement over goal completion. Understanding both traditions is essential for this dissertation because service robots must simultaneously pursue task goals (navigation, manipulation) *and* maintain socially coherent conversation, a requirement that lies at the intersection of the two.

Task-oriented systems have traditionally followed a modular pipeline architecture [25, 26] consisting of four sequential components. *Natural Language Understanding* (NLU) maps a user utterance to a semantic frame, identifying the user's intent i and a set of slot-value pairs $\{(s_k, v_k)\}$ that encode the specific information conveyed:

$$\text{NLU} : u_t \longmapsto (i_t,\ \{(s_k, v_k)\}). \tag{2.9}$$

Dialogue State Tracking (DST) maintains a compact representation of the conversation history, aggregating all information exchanged so far into a *belief state* b_t^{dst} over possible user goals:

$$b_t^{\text{dst}} = \text{DST}(b_{t-1}^{\text{dst}},\ u_t,\ a_{t-1}), \tag{2.10}$$

where a_{t-1} is the system's previous action. The DST belief state is directly analogous to the POMDP belief state b_t^{dst} introduced in Equation (2.5): both encode a probability distribution over hidden states given the observation history, differing only in that b_t^{dst} is defined over user goals in a dialogue domain. A *Dialogue Policy* selects the next system action a_t given b_t^{dst}, and *Natural Language Generation* (NLG) realizes a_t as a surface-form utterance delivered to the user.

The pipeline architecture affords interpretability and modularity but suffers from error propagation across stages and requires substantial domain-specific annotation for each new task. Neural approaches have progressively replaced individual pipeline components with learned models: neural NLU based on pre-trained encoders [27], end-to-end differentiable DST [28], and neural NLG based on sequence-to-sequence models [29]. Researchers have compiled a comprehensive dataset and a multi-domain end-to-end dialog system platform to advance task-oriented dialogue research by providing large-scale, multi-domain conversations for training and evaluation [30, 31].

Open-domain dialogue systems relax the goal-completion constraint and instead optimize for conversational quality, naturalness, and user engagement. Early retrieve-based systems select responses from a fixed corpus. Subsequently, neural generation models enabled open-ended response generation without a predefined ontology [32, 33, 34]. The end-to-end training paradigm of open-domain systems removes the need for hand-crafted semantic frames, allowing the model to learn implicit dialogue representations directly from data. However, these neural-based systems often fail to maintain consistency, are prone to producing safe but uninformative responses, and lack mechanisms for goal-directed behavior without an explicit task structure.

The service robot setting studied in this dissertation inherits requirements from *both*

traditions: it must track task-relevant state (destination, object identity) in the manner of a TOD system, while simultaneously adapting its conversational style and social cues to the individual user—a capability more characteristic of open-domain systems. This hybrid requirement motivates the joint policy learning framework presented in Chapter 3, in which dialogue acts and navigation commands are optimized jointly under a unified reward signal that captures both task efficiency and interaction quality.

2.2.2 Reinforcement Learning for Dialogue Optimization

The pipeline and end-to-end architectures described above are typically trained to maximize the likelihood of correct responses given labeled data. Yet the ultimate objective of a dialogue system is not to reproduce human utterances but to *achieve goals through conversation*—a sequential decision-making problem that RL is well-positioned to address.

Casting dialogue as an MDP maps naturally onto the formalisms of Section 2.1. The state s_t corresponds to the dialogue state (user goal, conversation history, system context); the action a_t is a dialogue act or utterance; the transition $T(s_t, a_t, s_{t+1})$ captures how the user responds and the conversation evolves; and the reward $R(s_t, a_t)$ encodes task success, efficiency, and user satisfaction. The dialogue policy $\pi(a_t \mid s_t)$ is then optimized to maximize the expected cumulative reward $J(\pi)$.

Because the user's internal goal is not directly observable, the dialogue MDP is more accurately modeled as a POMDP[26], with the DST belief state b_t (Equation (2.10)) serving as the agent's belief over hidden user states. Young et al. [26] demonstrated that POMDP-based dialogue managers trained with RL outperform hand-crafted and MDP-based policies in noisy, real-world conditions, directly motivating the use of DRQN and Recurrent PPO (Section 2.1.4) for joint dialogue-navigation policy learning in this

work.

A persistent challenge in RL-based dialogue optimization is *reward design*: task completion alone is too sparse to provide a useful learning signal, while proxy rewards based on user simulation introduce bias. Common reward components include a binary task-success signal $r_{\text{task}} \in \{0, 1\}$, a turn-penalty $r_{\text{turn}} < 0$ that encourages efficiency, and a user-satisfaction estimate r_{user} derived from a trained user model or subjective ratings [35]:

$$R(s_t, a_t) \;=\; r_{\text{task}} \;+\; \lambda_1\, r_{\text{turn}} \;+\; \lambda_2\, r_{\text{user}}, \tag{2.11}$$

where $\lambda_1, \lambda_2 > 0$ are weighting hyperparameters. In the HRI setting of this dissertation, r_{user} is extended to capture rapport and trust signals (defined formally in Section 2.4), providing the social grounding that purely task-oriented reward functions omit.

2.2.3 Large Language Model-Based Dialogue

The landscape of dialogue systems has been fundamentally reshaped by the emergence of large language models (LLMs) [36, 37, 38, 39, 40, 41, 42]. Pre-trained on web-scale text corpora and fine-tuned with human feedback, contemporary LLMs exhibit remarkable conversational fluency, broad world knowledge, and few-shot generalization to new domains—capabilities that far exceed those of the pipeline and end-to-end models described above.

From a dialogue-systems perspective, an LLM can be viewed as a highly expressive conditional language model $P_\theta(u_{t+1} \mid u_1, a_1, \ldots, u_t, a_t)$ that implicitly performs NLU, DST, policy selection, and NLG within a single forward pass over the context window. Instruction-tuned models [38] and retrieval-augmented generation [43] further extend this capability to follow task-specific directives and incorporate external knowl-

edge at inference time. Vision-Language Models (VLMs) [44, 45] additionally ground language in visual perception, enabling responses conditioned on image inputs.

Despite these advances, LLMs face a fundamental limitation when deployed on embodied robots: they are *disembodied* by design. LLM-based dialogue systems process language tokens and, in the case of VLMs, image patches—but they have no persistent representation of the robot's physical configuration, its location in a three-dimensional environment, or the spatial relationships between objects and agents that change as the robot moves and acts. This disembodiment manifests as three concrete failure modes in service robot applications. First, LLMs cannot reliably track *dynamic spatial context*: an utterance such as "the cup is on your left" is meaningful only relative to the robot's current pose, which the model has no mechanism to represent or update. Second, without access to the robot's action history and physical constraints, LLM-generated instructions may be *geometrically infeasible*—referring to objects out of reach, proposing paths that intersect obstacles, or ignoring manipulation workspace limits. Third, LLMs lack the sequential decision-making structure needed for *adaptive, goal-directed dialogue*: they generate responses token by token without explicitly maintaining a task state or optimizing for long-horizon dialogue rewards.

These limitations correspond directly to the research gaps identified in Chapter 1: the absence of spatial context in dialogue, the separation of language and physical action planning, and the lack of adaptive dialogue strategies grounded in the interaction history. Addressing those gaps—the requirement that language generation be coupled with geometrically feasible physical execution—requires a principled account of how high-level task plans can be translated into concrete robot motions under spatial and kinematic constraints. This is the subject of the next section, which introduces Task

and Motion Planning as the architectural bridge between symbolic task reasoning and physical execution.

2.3 Task and Motion Planning

The dialogue systems reviewed in Section 2.2 expose a critical gap: LLMs can generate linguistically fluent instructions but cannot guarantee that those instructions are physically executable by a robot operating under geometric and kinematic constraints. Bridging this gap—from *what to do* expressed in language to *how to do it* expressed in continuous robot motion—is the central problem addressed by Task and Motion Planning (TAMP) [46, 47].

TAMP integrates two levels of reasoning that have traditionally been studied in isolation. *Task planning* operates at the symbolic level, reasoning over discrete objects, actions, and their logical preconditions and effects to produce a sequence of high-level actions that achieves a goal. *Motion planning* operates at the geometric level, computing continuous trajectories for the robot's joints and base that satisfy physical constraints such as collision avoidance, reachability, and stability. TAMP couples these two levels so that the feasibility of low-level motion is taken into account when committing to high-level action sequences, enabling a robot to plan complex manipulation tasks that require coordinated reasoning about both *what* to do and *how* to move.

2.3.1 Symbolic Task Planning: PDDL and Answer Set Programming

Task planning originates in classical AI, where the problem is formalized as a search over a discrete state space defined by logical predicates. The *Stanford Research Institute Problem Solver* (STRIPS) [48] introduced the foundational representation: a state is a set of ground predicates, and each action is described by a triple $(\text{pre}, \text{add}, \text{del})$ speci-

fying the preconditions that must hold before execution and the positive and negative effects on the state. The *Planning Domain Definition Language* (PDDL) [49] standardized and extended this representation into a widely adopted specification language, enabling a broad ecosystem of domain-independent planners [50]. PDDL supports numeric fluents, durative actions, and temporal constraints, making it expressive enough for complex multi-step manipulation tasks.

A PDDL domain separates the *domain*—the types, predicates, and action schema that describe the physics of the world—from the *problem*—the specific initial state and goal condition for a given task instance. For a mobile manipulation scenario, the domain might declare:

- *Predicates*: $\texttt{At}(\textit{robot}, \textit{loc})$, $\texttt{Holding}(\textit{robot}, \textit{obj})$, $\texttt{On}(\textit{obj}, \textit{surface})$, $\texttt{Free}(\textit{robot})$
- *Actions*: $\textsc{Navigate}(r, l_1, l_2)$, $\textsc{Pick}(r, o, l)$, $\textsc{Place}(r, o, s)$, each with typed preconditions and effects

A planner then searches for a sequence of grounded action instances $\sigma = (a_1, a_2, \ldots, a_n)$ that transforms the initial state into one satisfying the goal condition. In this dissertation, PDDL is used in Chapter 4 to specify task domains and generate action sequences that are subsequently grounded in geometric motion plans.

Hierarchical Task Networks (HTN) [51] complement forward-search planners by recursively decomposing high-level tasks into subtasks via *methods*, thereby encoding domain-specific procedural knowledge that guides the search. HTN planning is particularly well-suited to service robot applications, where a goal such as *"set the table"* naturally decomposes into a hierarchy of subtasks, each of which further decomposes into primitive navigation and manipulation actions.

While PDDL-based planning excels at generating action sequences under explicit

precondition-effect semantics, it faces limitations when tasks involve *complex constraint satisfaction*, *commonsense reasoning under incomplete information*, or the need to enumerate *all valid plans* rather than a single optimal one. *Answer Set Programming* (ASP) [52, 53] addresses these limitations through a fundamentally different computational paradigm rooted in non-monotonic logic.

In ASP, a planning problem is encoded as a logic program consisting of *rules* of the form:

$$h \texttt{ :- } b_1, b_2, \ldots, b_m, \texttt{ not } c_1, \ldots, \texttt{ not } c_k, \tag{2.12}$$

where h is the *head* (a conclusion), $b_1, \ldots, b_m$ are *positive body literals* (conditions that must hold), and $\texttt{not } c_1, \ldots, \texttt{not } c_k$ are *default negation* literals (conditions that must *not* be derivable). This default negation—the characteristic of non-monotonic reasoning—allows ASP to draw *defeasible* conclusions: an agent can assume that an obstacle does not block a path *unless* evidence to the contrary is present, without requiring an exhaustive enumeration of all possible world states.

The semantics of an ASP program is defined by its *stable models* (answer sets) [54]: the complete, consistent sets of beliefs derivable from the program under default negation. Crucially, a program may admit *multiple* stable models, each corresponding to a distinct valid plan; an ASP solver such as CLINGO [55] enumerates these models efficiently, providing the full space of valid action sequences from which the most appropriate can be selected.

For robot task planning, ASP offers three key advantages over classical PDDL planners. First, *constraint encoding* is direct: hard constraints (e.g., never place a fragile object on an unstable surface) and soft preferences (e.g., prefer shorter plans) can be expressed declaratively as integrity constraints and weak constraints, respectively, with-

out modifying the planner's search procedure. Second, *commonsense and incomplete-information reasoning* is natural: ASP's closed-world assumption under default negation allows the planner to reason about what is *not* known, which is essential in partially observable real-world environments. Third, *non-monotonic revision* is supported: when new observations invalidate a previous assumption (e.g., an expected object is missing), the ASP program can be updated and re-solved incrementally without restarting planning from scratch [55].

In this dissertation, ASP is employed in Chapter 4 to generate task plans in scenarios where multiple geometrically and logically valid plans exist and must be evaluated against spatial and social constraints. The combination of PDDL for structured domain specification and ASP for flexible, constraint-rich plan generation provides the complementary strengths needed for the diverse task environments considered in this work.

A fundamental challenge shared by both PDDL and ASP planners when deployed on physical robots is that the feasibility of a symbolic action depends not only on its logical preconditions but also on continuous geometric factors. The following section examines these geometric constraints in the context of the robotic platforms used in this dissertation.

2.3.2 Geometric Constraints and Robot Platforms

The research presented in this dissertation spans three classes of robotic platforms, each imposing a distinct set of geometric constraints on task and motion planning.

A *quadruped robot* (legged platform without manipulation capability) must plan its gait and foothold sequence to traverse uneven terrain while maintaining dynamic stability; the planning problem is primarily one of whole-body locomotion under contact constraints, and task planning concerns the sequencing of navigation goals rather than

manipulation. It is worth noting that its whole-body locomotion can be trained on RL algorithms. A *humanoid robot* combines bipedal locomotion, requiring whole-body coordination: reaching an object may require simultaneous adjustment of the base pose and torso orientation to maintain balance while achieving the desired end-effector pose. A *mobile manipulator*—a wheeled mobile base equipped with a robotic arm—is the primary platform for the manipulation contributions of this dissertation. Its planning challenges are less concerned with dynamic stability than with the *kinematic compatibility* between the base position and the arm's reachable workspace: the arm can only manipulate objects within a bounded region determined by its link lengths and joint limits, so the robot must navigate its base to a position from which the target is reachable before attempting manipulation.

This *base placement problem* is a key bottleneck in mobile manipulation planning [56, 57, 58]. An unsuitable base position may render the manipulation target kinematically unreachable, cause arm-environment collisions, or leave insufficient clearance for the manipulation motion. Determining a valid base placement, therefore, requires reasoning jointly about the robot's position in the environment, the arm's reachability envelope, and the geometry of surrounding obstacles—a decision that must be made at the task planning level before motion planning can proceed. Practical approaches to base placement include offline precomputed reachability maps [59], sampling-based feasibility checks [60], and learned predictors [61]; the specific strategy adopted in this dissertation is detailed in Chapter 4.

Together, the symbolic planning formalisms of Section 2.3.1 and the geometric constraints described here define the full TAMP problem: finding a sequence of symbolic actions, each accompanied by a geometrically feasible motion, that achieves the task

goal across the platform's physical constraints. Solving this problem reliably in real environments requires the robot to maintain an accurate, semantically rich model of its surroundings—including the locations, identities, and spatial relationships of objects and agents. The next section examines how multimodal perception and spatial semantic representations provide this world model.

2.3.3 Grounding Language in Physical Space

The perceptual capabilities reviewed above—open-vocabulary visual understanding and structured spatial representations—converge in the problem of *spatial language grounding*: the ability to interpret and generate natural language expressions whose meaning is anchored in the physical configuration of the environment. Spatial grounding is the mechanism by which abstract linguistic concepts such as *"to the left of the door"* or *"the object nearest to you"* acquire concrete meaning relative to the robot's current position and perception. This concept has been applied in various forms across robotics research, including navigation instructions, object manipulation tasks, and human-robot dialogue [62, 63, 64, 65].

Referring expression comprehension [66, 67, 68] studies the task of identifying a target object in a scene given a natural language description, requiring joint reasoning over visual features and spatial relations. *Embodied question answering* [69, 70] extends this to active perception: the robot must navigate and explore its environment to gather the visual evidence needed to answer spatially grounded questions. *Vision-and-Language Navigation* (VLN) [71, 72, 73] pushes further, requiring the robot to follow natural language navigation instructions [74, 75, 76] in previously unseen environments by grounding directional and landmark references in visual observations.

These tasks illustrate that spatial grounding is not a static mapping from words to

world locations but a *dynamic, embodied process* that depends on the robot's position, motion history, and perception of a changing environment. Current VLM-based approaches, while powerful, largely treat each grounding query independently and in a single image frame, without maintaining the persistent spatial model needed for multi-step task execution or extended human-robot dialogue. Developing systems that maintain and update spatial grounding representations across the full duration of a robot's operation—and that can communicate spatially grounded information fluently to human partners—remains an important open challenge, and represents a key research opportunity building on the contributions of this dissertation.

2.4 Social Robotics and Evaluation Frameworks

The spatial and perceptual capabilities reviewed in previous sections equip a robot to understand *where* it is and *what* surrounds it. Effective human-robot collaboration, however, requires an additional layer of understanding: the robot must navigate the *social* dimensions of shared space, interpret and respect the interpersonal distance preferences of its human partners, and build the trust and relational connections that sustain long-term cooperation. This section establishes the social and psychological foundations that motivate the rapport-building framework proposed in Chapter 5, and defines the evaluation metrics used to assess both task performance and interaction quality throughout this dissertation.

2.4.1 Proxemics and Socially Aware Navigation

Human spatial behaviour is governed by implicit norms of interpersonal distance—norms whose violation elicits discomfort, distrust, and avoidance. *Proxemics*, introduced by Hall [77], identifies four concentric distance zones that structure human spatial

interaction: the *intimate zone* (< 0.5 m), reserved for close personal relationships; the *personal zone* (0.5-1.2 m), comfortable for friends and acquaintances; the *social zone* (1.2-3.7 m), appropriate for professional interactions; and the *public zone* (> 3.7 m), characteristic of formal or anonymous encounters. These zones are not rigid boundaries but culturally modulated gradients of comfort that vary with relationship type, task context, and individual preference [78].

Proxemic theory has informed a broad line of research in socially aware navigation [79, 80], in which motion planners are augmented with social cost functions that penalize trajectories that intrude into the personal or intimate zones of nearby humans. While this motion-level perspective is important for full-stack social robot systems, the scope of this dissertation focuses on the *communicative* dimension of proxemics rather than its locomotion consequences. Specifically, the work assumes that the human and robot share a common *spatial context*—they are co-located within the same physical environment and have mutual awareness of the same objects, layout, and events—and investigates how this shared spatial context can be leveraged to ground dialogue, build rapport, and adapt communication to the individual. Integrating motion-level proxemic planning with the adaptive dialogue strategies proposed in Chapter 5 is identified as a promising direction for future work.

2.4.2 Trust and Rapport in Human-Robot Interaction

Section 1.2 introduced concise operational definitions of *trust* and *rapport* that are used throughout the dissertation; this subsection provides the full theoretical treatment of their antecedents, formation mechanisms, and relationship to one another. Two relational constructs are central to the long-term success of human-robot collaboration: *trust* and *rapport*. Although often used interchangeably in everyday language, they refer

to distinct psychological phenomena with different antecedents, formation mechanisms, and consequences for interaction quality. Carefully distinguishing them is essential for the adaptive dialogue strategy presented in Chapter 5.

Trust in the context of HRI is primarily *competence-based*: a human trusts a robot to the extent that they believe the robot is capable of reliably performing the task for which it is responsible [1, 9, 2]. Trust is calibrated through accumulated evidence of the robot's performance— successes build trust, failures erode it—and is modulated by the perceived predictability and transparency of the robot's behavior [81, 82]. Crucially, trust is not binary but a continuous, dynamic quantity that evolves over the course of an interaction and across repeated encounters [3]. Overtrust (assuming greater capability than the robot possesses) and undertrust (failing to utilize a capable robot) are both costly outcomes that adaptive communication strategies must guard against.

Rapport, by contrast, is fundamentally *relationship-based* [4, 5]: it refers to the sense of mutual attentiveness, positivity, and coordination that characterizes harmonious interpersonal (or human-robot) interaction. Rapport is established not primarily through task performance but through the quality of the communicative interaction itself—through expressions of interest, appropriate acknowledgment of the human's emotional state, conversational alignment, and the perception that the robot "understands" the human's situation and preferences. Where trust asks *"can the robot do the job?"*, rapport asks *"does the robot understand me?"*.

In HRI, rapport has been associated with increased willingness to engage with the robot [83] and higher subjective ratings of interaction quality [5]. The mechanisms through which rapport is established mirror those identified in human-human interaction research [4]: *mimicry* (matching the interlocutor's linguistic style or non-

verbal behavior), *responsiveness* (demonstrating attentiveness to the partner's contributions), and *personalization* (adapting to the individual's preferences and history). For a robot, these mechanisms translate into concrete dialogue strategy choices: adopting the user's vocabulary, acknowledging spatially relevant events in the environment, referencing past interactions, and adjusting communication style to the user's demonstrated preferences—precisely the capabilities targeted by Contribution 3 (Chapter 5).

The relationship between trust and rapport is synergistic but asymmetric: competence-based trust is a necessary precondition for rapport to develop (a robot that consistently fails its task will not sustain positive engagement), but rapport, once established, amplifies trust by making the robot's occasional errors more forgivable and its explanations more credible [3]. This asymmetry motivates the three-contribution structure of this dissertation: Contributions 1 and 2 establish the competence foundation (reliable task execution with spatially grounded communication), while Contribution 3 builds the rapport layer on top.

The communicative strategies through which trust and rapport are established connect directly to a broader class of systems studied under the umbrella of *social influence dialogue*. Chawla et al. [84] define *social influence dialogue systems* as dialogue systems capable of influencing users' cognitive and emotional responses—leading to changes in thoughts, opinions, and behaviors—through natural conversation. The survey identifies seven domains in which such influence is studied: persuasion, negotiation, counseling, motivational interviewing, tutoring, emotional support, and social companionship, and organises the methodological landscape around three interconnected concerns: *strategy representation*, *reinforcement learning for optimization*, and *rapport- and trust-oriented systems*.

Strategy representation addresses how a system encodes and selects its communicative tactics. Three principal approaches have been explored: *implicit* end-to-end models that learn strategy and language jointly from data [85]; *dialogue acts* (DAs), which decompose the strategy space into discrete communicative functions (e.g., offer, concede, appeal) and enable modular, interpretable policy learning [86, 87]; and *latent strategy vectors* that decouple utterance semantics from surface realisation, allowing a planner to reason at the level of intent before committing to specific wording [88]. The DA-based approach is particularly relevant to this dissertation, as dialogue acts provide a natural interface between the symbolic policy layer and the language generation component.

Reinforcement learning for dialogue optimization has been applied to social influence tasks to go beyond supervised imitation of average human behaviour, which is often sub-optimal in influence settings [84]. RL-trained systems explore their own strategies and are guided by task-specific rewards—such as the final agreed price in a negotiation [85] or the degree of attitude change in a persuasion task [89]—yielding policies that outperform those trained by supervised learning alone. Partner modeling, in which the system maintains a belief over the interlocutor's goals or preferences and uses it to guide action selection, has further improved RL-based social influence systems by providing a more informative signal for long-horizon planning [87].

Rapport- and trust-oriented systems are most directly relevant to this dissertation. In therapy and emotional support settings, empathetic responses and linguistically accommodating behavior have been associated with stronger rapport and better outcomes [90, 91]. In persuasion settings, personalising the appeal strategy to the individual's values and prior responses—rather than applying a fixed tactic—significantly increases compliance [89]. Across domains, Chawla et al. [84] identifies partner perception (i.e.,

whether the user *likes* the system and wishes to interact again) as an essential but frequently neglected evaluation axis, alongside the objective task outcome—a distinction that maps directly onto the trust–rapport duality formalized earlier in this section.

These insights inform the adaptive dialogue strategy proposed in Chapter 5: the framework draws on DA-based strategy representation for interpretability, RL for policy optimization against rapport and task-success rewards, and personalization mechanisms grounded in the shared spatial context and interaction history. At the same time, social influence in HRI must be exercised with care: strategies effective in persuasion or negotiation contexts can feel manipulative if applied without transparency, and maintaining calibrated trust requires the robot to be honest about its capabilities and limitations [81].

2.4.3 Evaluation Metrics

Evaluating a human-robot collaborative system requires metrics that capture both *what the robot achieves* (task performance) and *how the interaction feels to the human partner* (interaction quality). This section defines the evaluation framework used consistently across the experimental chapters of this dissertation.

Task performance metrics. For navigation and instruction-following tasks, two standard metrics from the Vision-and-Language Navigation literature [71] are adopted. *Success Rate* (SR) measures the fraction of episodes in which the robot successfully completes the assigned task:

$$\mathrm{SR} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{task}_i \text{ succeeded}], \tag{2.13}$$

where N is the total number of evaluation episodes. *Success weighted by Path Length* (SPL) [71] additionally penalises unnecessarily long paths, rewarding both correctness

and efficiency:

$$\text{SPL} \;=\; \frac{1}{N}\sum_{i=1}^{N} S_i \cdot \frac{\ell_i}{\max(p_i,\, \ell_i)}, \tag{2.14}$$

where $S_i \in \{0, 1\}$ is the success indicator, ℓ_i is the shortest possible path length for episode i, and p_i is the actual path length traversed by the robot. For manipulation tasks, additional metrics include *object placement accuracy* (deviation from the target pose) and *task completion time*.

Human-centred metrics. Cognitive load imposed on the human partner is assessed using the *NASA Task Load Index* (NASA-TLX) [92], a six-dimensional subjective workload scale covering mental demand, physical demand, temporal demand, performance, effort, and frustration. Lower NASA-TLX scores indicate that the robot's behavior allows the human to collaborate without excessive cognitive burden.

Trust is measured using validated questionnaire scales adapted for HRI, including the *Trust in Automation* scale [93] and the *Multi-Dimensional Measure of Trust* (MDMT) [6], which decompose overall trust into sub-dimensions such as *reliability*, *competence*, and *predictability*.

Rapport is assessed using the *Connection-Coordination rapport (CCR) scale* adapted from Lin et al. [7], which measures connection and coordination in human-robot dyads based on behavioral indicators such as entrainment and lingering conversations. The CCR scale provides objective proxies for rapport that do not rely solely on subjective self-report [5]. Both MDMT and the CCR Scale are operationalized in the post-hoc trust-and-rapport analysis of Contribution 3 (Chapter 5, Section 5.9), which empirically connects dialogue-quality improvements to measurable relational outcomes.

Integrated evaluation design. Across all experimental chapters, task performance metrics (SR, SPL, placement accuracy) quantify the robot's functional capability, while

human-centred metrics (NASA-TLX, trust scales, rapport scales) assess the quality of the collaborative experience. This dual evaluation framework reflects the central claim of this dissertation: that spatial and social grounding of dialogue improves not only task efficiency but also the subjective experience of collaboration—and that these two dimensions are complementary, not in tension.

With the theoretical and empirical foundations established across Sections 2.1–2.4, the dissertation now turns to its first contribution. Chapter 3 presents the joint policy learning framework for dialogue and co-navigation, demonstrating how the MDP and POMDP formalisms of Section 2.1, the dialogue architectures of Section 2.2, and the proxemic constraints of Section 2.4.1 are integrated into a single learnable policy that coordinates speech and motion in a shared physical space.

3 Joint Policy Learning for Dialogue and Co-Navigation

3.1 Introduction

Robots are increasingly tasked with services that require dialog and navigation in human-inhabited environments [76]. While there is rich literature on mobile robot navigation [94] and dialog systems [25], there is relatively little research on the joint management of human-robot dialog and their co-navigation. Considering a robot that moves around while conversing with a human side by side, e.g., a robot realtor, the robot needs the capability of "dialog navigation" [95, 74]. The dialog navigation task requires the robot to navigate and interact with a human at the same time, while fulfilling service requests.

When a robot talks to and navigates with people at the same time, there are new challenges caused by the interplay between the two types of actions. For instance, when a real estate robot wants to tell a customer that a bedroom is large and cozy, which requires human-robot dialog, it is better for the robot to first guide the customer to that bedroom, which requires human-robot co-navigation. Dialog capabilities enable the robot to estimate the human's belief state, and accordingly select the next navigation goal; navigation capabilities help change the human-robot system's location to facilitate the next few turns of the ongoing dialog. Many studies on dialog navigation allow re-

mote communication between a user and a robot, e.g., the user is from a call center [62, 96], where the complexity of social navigation is avoided. By comparison, we consider scenarios where the human and robot talk to and navigate with each other at the same time.

In this paper, we focus on dialog navigation tasks where the human and robot are not always co-located. In such tasks, tracking the positional information and dialog states is crucial for the robot. For instance, the robot might want to navigate to the human, and then verbally encourage the human to walk together to a room before discussing the room features. We develop a mobile robot system that leverages the locations of the human and itself for selecting its dialogue and navigation behaviors. In order to deal with the uncertainty and partial observability in dialog and navigation actions, we use Deep Recurrent Reinforcement Learning for computing joint policies for selecting both dialog and navigation actions [21, 24]. Our framework enables the robot to reason about the history of observations from dialog and navigation actions toward the long-term goal of efficiently and accurately conveying a few location-dependent statements.

We have evaluated our framework in a real estate agent domain where a mobile robot is tasked with fulfilling user requirements of house features via dialog and navigation actions. Our system has been compared with three baselines that alternatively take dialog and navigation actions. One has a rule-based dialog-navigation planner, and another relies on a dialog policy and a planner for navigation. Our experiments show that our framework can learn a joint policy and outperform the baselines regarding task completion rate and execution time.

Figure 3.1: When the robot receives positive feedback from the human, the robot believes it is likely that the information (the room being `beautiful` and with `view`) has been successfully delivered (**top-left**). In the absence of the human, it is unlikely the robot is able to deliver the room features (**top-right**). When the human does not give any verbal feedback to the robot, the likelihood that the robot successfully delivers the information is medium (**bottom-left**). When the robot talks about another room instead of the current one, there is a lower chance that the human accepts such information (**bottom-right**).

3.2 Related Work

This section discusses three research areas that are related to dialog navigation systems: (1) instruction-following navigation systems, (2) systems where navigation follows dialog, and (3) systems where navigation and dialog are interleaved.

Instruction Following: Instruction-following navigation systems have been studied, where a robot communicates with a user to complete a navigation task based on user instructions. Kollar et al. developed a system that can generate action plans from a given directional instruction by parsing the natural language [97]. Other works introduced parser learning for indoor navigation instruction that translates natural language

commands to actions [98, 99]. In recent years, researchers developed a benchmark, ALFRED, for learning to ground both language instructions and robot perceptions into sequences of actions [68], realizing complex robot tasks such as household tasks [100, 101]. Those works assumed the human and robot share the same observability over the world and focused on understanding human instructions instead of managing multi-turn conversations.

Navigation Goal Specification via Dialog: One way of integrating dialog and navigation is to specify navigation goals via human-robot dialog, where the dialog occurs before navigation. Some efforts enabled a robot to conduct a service task after identifying user requests through dialog with probabilistic inference [102, 75, 103]. Other works learned dialog behaviors, such as asking clarification questions from human-robot conversations, to improve the task performance [74, 104]. Another work focused on spoken language understanding to retrieve navigation-associated information and improve a navigation task that follows the language understanding step [105]. Although such systems improved the performance in task completion by introducing a dialog system into navigation tasks, the robot cannot handle dynamic changes in user requests since the conversation happens only before navigation in their systems. In addition, those works assume that conversations happen in a single location, whereas the multiple turns of our human-robot dialog might occur in different locations.

Interleaved Dialog Navigation: Recent research has produced dialog navigation systems that alternate between dialog and navigation actions. Thomason et al. presented Cooperative Vision-and-Dialog Navigation (CVDN) that trains language-teleoperated home and office robots that ask targeted questions about where to go next [73]. DIALFREAD [106] is an extension of ALFRED that learns to ask for instructions to handle

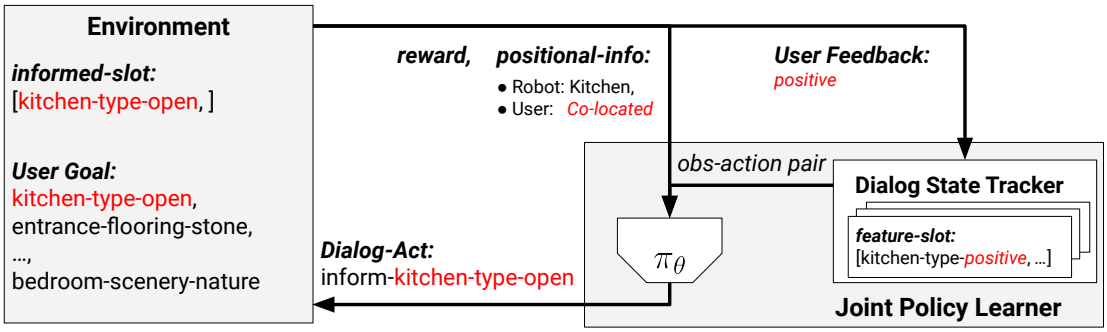


Figure 3.2: An overview of our joint policy learning framework.

unexpected situations. Studies on such dialog navigation systems have developed some datasets to study navigation and spatial reasoning with real-life observations, learning to ground language to perception and behavior, and realizing a cooperative localization task [65, 96, 107, 62]. These studies can be seen as interactive instruction-following systems and allow instructions always to be available no matter where two agents are. By comparison, our robot can co-navigate with a human to fulfill service requests. Our work considers the robot as a guide and the human as a customer to study how the robot acquires complementary behaviors of dialog and navigation, being aware of the human's locations.

3.3 Problem Statement

We design a real estate agent domain for a mobile dialog task where a robot guides users physically and performs dialog at the same time. Fig. 3.1 shows how a user and a robot interact with each other in our proposed domain. We define a domain description and problem for the mobile dialog task.

3.3.1 Real Estate Agent Domain

A domain description is defined as a tuple $\langle \mathbf{E}, \mathbf{e}, \mathbf{Q}, \mathbf{q} \rangle$, where $\mathbf{E} = [E_0, E_1, \cdots]$ is a finite set of entities of which a system can inform. For instance, $E_i \in \mathbf{E}$ can

be "*livingroom-view*." $\mathbf{e} = [e_0, e_1, \cdots]$ is a full assignment of $\mathbf{E}$, which specifies the feature values of a particular house. For instance, $e_i =$ "*beautiful*" means the living room of the house has a beautiful view. $\mathbf{Q} = [Q_0, Q_1, \cdots]$ is a set of entities in which a user is interested, and $\mathbf{Q} \subseteq \mathbf{E}$. For instance, $Q_j \in \mathbf{Q}$ can be "*bedroom-size*." $\mathbf{q}$ is a full assignment of $\mathbf{Q}$, which specifies the requirement values of a particular user. For instance, if the user looks for a large bedroom, $q_j =$ "*large*".

3.3.2 Real Estate Agent Problem

The goal in the real estate agent domain is to estimate the user's requirements and introduce property features within a maximum number of timesteps. We define the real estate agent problem as a POMDP. $S := N \times N \times C^{|\mathbf{E}|} \times C^{|\mathbf{Q}|}$ is defined as a set of states factored in a robot location, a user location, entities to be informed to a user, and the user's requirements, where N is the number of rooms in real estate. The state space includes an entity slot and a requirement slot of which sizes are $|\mathbf{E}|$ and $|\mathbf{Q}|$, each with cardinality C. $A := A^D \cup A^N$ is a union of dialog actions A^D and navigation actions A^N. The types of actions are "*guidance*," "*inform*," and "*report*" for A^D and "*goto*" for A^N. When taking an *inform* action, the robot assigns a slot-value of an entity to the dialog action to inform the user about the entity, e.g., "*inform-livingroom-view-beautiful*". When taking other types of actions, the robot assigns a room to the action for navigation, e.g., "*guidance-bedroom-none-none*" and "*goto-bedroom-none-none*." Thus, the total number of actions is proportional to the number of rooms and entities. O is a set of observations the robot receives from the environment with a user. The observation contains the robot's and the user's locations and feedback from the user. The user gives *positive* feedback if the informed value $e_i \in \mathbf{e}$ meets the user's requirements $q_i \in \mathbf{q}$, *negative* feedback if not, and *none* otherwise. $T := S \times A \times S \rightarrow$

$[0, 1]$ is a transition function. At the time-step t, the environment is in some state s_t. The robot takes an action a_t, which causes the environment to transition to state s_{t+1} with probability $T(s_t, a_t, s_{t+1})$. $Z := A \times S \times O \rightarrow [0, 1]$ is a observation function. The robot receives an observation $o_t \in O$, which depends on s_{t+1} and a_t, with probability $Z(o_{t+1}|a_t, s_{t+1})$ at the time-step t. $R := S \times A \rightarrow \mathbb{R}$ is a reward function. When the robot terminates the task, the reward function returns a reward according to the evaluation function $y(e, q) = |\{e_i|e_i \in q_i, 0 \leq i < |\mathbf{q}|\}|$. The more requirements of the user are satisfied, the higher rewards the robot receives. The robot receives a negative reward at each time-step as a living cost.

3.4 Proposed Approach

In this section, we present the key contribution of this work, an RL-based framework for learning a joint dialog-navigation policy.

Framework Description: In a dialog navigation task where the conversation and the positional relationship between a robot and a user strongly relate to their locations, optimization of navigation and dialog actions becomes more complex. We develop a framework that can learn a joint policy of dialog and navigation considering the spatial context and the course of dialogues. Joint policy learning refers to a model that maps a state that includes spatial and dialogue historical information to either a navigation or a dialog action. Thus, we propose unifying the spatial information with dialog state tracking so that the tracker can handle location changes during the dialogues. This formulation enables a robot to traverse different locations that bring new spatial contexts while tracking a dialog objective. We aim to compute a joint policy that guides the robot in the dialog and navigation actions toward maximizing the expected cumulative utility

to achieve a mobile dialog task.

Fig. 3.2 shows an overview of our framework. The framework consists of a joint policy learner and an environment. The joint policy learner has a dialog state tracker that tracks a course of dialogue and a dialog navigation policy that takes action based on the positional information and the dialog history. The environment includes a house instance and a user. The robot can take a verbal or navigation action, and the environment gives the robot feedback from the user and rewards based on the user's requirements.

Dialog State Tracking: In order for a robot to follow the course of dialogue over a task, we use a dialog state tracker. We utilize the slot-filling approach to modify the input to the dialog navigation policy instead of using raw observations so that the robot can be aware of entities to be informed. This allows the robot to store the current task completion rate over domains. The slots are updated by the dialog state tracker that takes user feedback and performs slot-filling. The dialog state tracker fills a corresponding slot with feedback from the user based on the observation.

We then use a robot's and user's locations, the feature slots, and the previous action as an observation o_t and input them to the dialog navigation policy network π_θ. Since the user's location is partially observable, the policy network needs to have a memory for the course of navigation. It is also important for the joint policy to deal with the user's feedback fluctuation due to uncertainty as a natural language understanding module propagates the feedback error to the dialog state tracking. According to the aforementioned reasons, we implement a policy network with a Deep Recurrent Reinforcement Learning model.

Implementation: Given the observation o_t, r_t, the dialog navigation policy network learns a joint policy that can output either a dialog or navigation action. Since navigation

Table 3.1: Examples of slot-value pairs for all domains

Domain	$\|\mathbf{E}\|$	Examples of {slot}-{value} pairs
livingroom	4	{view}-{beautiful}, {size}-{200}, $\cdots$
kitchen	4	{view}-{open}, {roomrel}-{livingroom}, $\cdots$
bedroom	4	{size}-{100}, {roomrel}-{livingroom}, $\cdots$

Table 3.2: Dialog acts and navigation actions

User's actions	posans / negans / none
User's navigation actions	stay / goto
Robot's dialog actions	inform / guidance / report
Robot's navigation actions	stay / goto

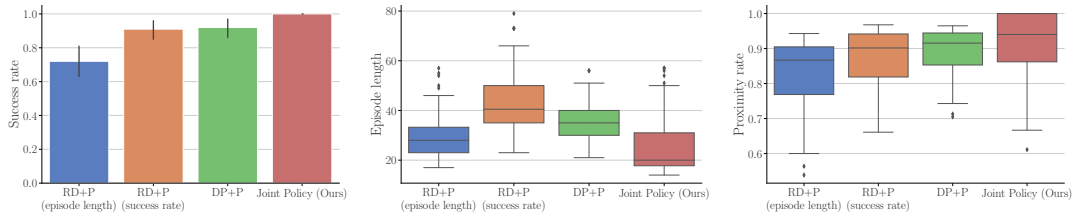
actions do not induce user feedback, the model requires a long memory for tracking an entire course of dialogue. Therefore, the policy network has an LSTM layer to aggregate observations over the turns. The policy network π_θ is updated every M time-step after collecting experiences from the environment. This process varies depending on the algorithm of choice. RecurrentPPO is an extension of PPO and has LSTM at the output layer [24]. RecurrentPPO collects M samples in its rollout buffer and calculates a policy gradient using the samples to update a network that selects the next action, an actor-network.

3.5 Experiment

This section presents experimental results in simulation and a demonstration of our system in the real world. We develop a simulator for the evaluation of algorithms for dialog navigation tasks. We have compared our approach to three baselines and analyze the results.

3.5.1 Simulator Design

We develop a simulated environment for the real estate agent problem. The simulator consists of 10 house instances. Each house instance has three domains ("*livin-*



(a) Success rate. (b) Episode length. (c) Proximity rate

Figure 3.3: Average success rate, the episode length, and proximity rate over 100 runs, while the robot works on completing the task for different users. The agent with the joint policy outperforms the baselines.

groom", "*kitchen*", and "*bedroom*"), each with 4 feature slots. We manually created the slot-value pairs for each house instance. TABLE 3.1 shows examples of entities and those values for the three domains. Slot-value pairs are randomly assigned to the user's requirement slots when an environment is instantiated. At every time step, the robot can only receive its location, the user's existence, and the user feedback if available. Note that the user's location and the feedback from the user are available only when the robot and the user are co-located.

User simulators have been widely applied to training dialog policies using RL methods. In line with previous research, we simulate user behaviors, including responding to the robot's dialog actions and moving in the environment. During a mobile dialog task, the robot and the user act alternatively. TABLE 3.2 shows the robot and user actions used in the experiments. We use an action representation that is similar to the MultiWOZ-2.1 dataset for both dialog and navigation actions [108]. We consider different locations as domains as we consider an embodied system and address changes derived from the spatial context. The action representation contains 4 different pieces of information. The first piece is an *intent* which can be treated as an action. The second is a *domain*. Each domain has entities with values, and actions should include a *slot* and a *value* for an entity. For instance, the robot action, [*inform-kitchen-view-open*],

represents that the robot *inform*s of a *kitchen view*, which is an *open* view. This action representation enables the joint policy learner to take both a dialog and navigation action as a part of observation in the same manner and simplifies the observation space representation and action space representation. An example of a course of dialog and location transitions is as follows:

Example of Human-Robot Dialog

Robot: [*inform-livingroom-size-large*]

User: [*positive-livingroom-size-large*]

Robot: [*guidance-kitchen-none-none*]

User: [*stay-kitchen-none-none*]

Robot: [*goto-kitchen-none-none*]

User: [*goto-kitchen-none-none*]

⋮

User: [*positive-bedroom-roomrel-kitchen*]

Robot: [*report-none-none-none*]

Example of Trajectories

Robot: livingroom $\longrightarrow$ kitchen $\longrightarrow$ bedroom

User: livingroom $\longrightarrow$ kitchen $\longrightarrow$ bedroom

The environment returns a positive reward $R_{max} = 1.0$ when all user requirements are satisfied and a negative reward $R = -0.01$ otherwise. User feedback has noise with a 0.1 probability of being opposite. Robot and user movements are controlled with transition probabilities. The robot can successfully navigate to its goal location in 0.9 probability. The user follows the robot in presence of a 0.05 probability noise, i.e., there

is a small probability that the user moves to a location that is different from the robot's goal location.

3.5.2 Experiment Settings

The robot spends $5,000,000$ timesteps for each run to learn a joint policy. The actor-network and the critic network have the same network architectures. The first two layers have an MLP layer of 64 units with *tanh* activation as a hidden layer. The output of the second layer is fed into the LSTM layer with 256 units. For hyper-parameters of RecurrentPPO, we vectorize the environment with the size of 32 to collect enough samples and use the samples with a batch size of 4096. The discounting factor γ is 0.99, the learning rate α is 0.002, and clip size ϵ is 0.1.

We compare the agent with a joint policy, *Joint Policy*, to three baselines: the agent with a dialog policy and navigation planner, *DP+P*, and the two agents with a rule-based dialog system and navigation planner. For DP+P, we use the same dialog state tracker and a RecurrentPPO model as our joint policy. The difference from our model is that the model takes only the dialog state and outputs a dialog action and does not consider the user location. We evaluate the agent's performance of those policies for 100 independent episodes every $100,000$ timestep by using a deterministic policy and select the best policy model to compare the performance in the success rate and the episode length. For the rule-based dialog systems, we designed two different strategies. One is to try to finish a task as soon as possible, *RD+P (episode length)*. The other is to try to achieve a higher success rate, *RD+P (success rate)*. Baseline agents have isolated dialog and navigation systems and conduct the task by executing dialog and navigation actions interleaved. The navigation planner of the baselines always takes the shortest plan from the current position to the destination.

3.5.3 Results

Fig. 3.3a and 3.3b show the results of the agent's success rate and episode length over 100 runs. Both figures demonstrate that our approach outperforms the baselines. Joint Policy could achieve the highest success rate and a shorter episode length. Although RD+P (episode length) has a lower episode length than RD+P (success rate) and DP+P, the success rate is the lowest among the four methods. DP+P produced a lower episode length than RD+P (success rate), whereas it has similar performance in the success rate. This observation shows that rule-based dialog systems need to spend a longer episode length to identify the user's requirements and that the dialog policy can track the dialog state properly The result also implies that DP+P could not achieve a higher success rate than RD+P (success rate) because of fewer co-located cases. In order to investigate how the robots handled both dialog and co-navigation, we define the proximity rate PR as follows:

$$PR = \frac{\textit{The number of turns when they are co-located}}{\textit{Total number of turns}}$$

The proximity rate presents how frequently the robot and the user are co-located. A higher proximity rate indicates that the robot can perceive the absence of the user and correctly select a navigation action so that they are co-located. Joint Policy has the highest proximity rate, and DP+P and RD+Ps have a lower proximity rate. This indicates that the DP+P tries to inform entities without considering the user's location to minimize the episode length, resulting in a lower success rate than Joint Policy. On the other hand, a higher proximity rate of Joint Policy indicates that our model could handle the positional information and dialog simultaneously, resulting in the highest success rate

and the shortest episode length.

3.5.4 Demonstration in the Real World

We have demonstrated our approach using a real robot. We deployed our joint policy model on a segway-based mobile robot. The robot interacts with a human while traversing an environment where three rooms exist. We use YOLO V3 for ROS to realize real-time object detection [109] and Google Dialogflow for a natural language understanding component [110].

4 Enabling Approaches for Spatial Awareness for Rich Communication

4.1 Overview: Two Perspectives on Spatial Grounding

The preceding chapter established that integrating *when* and *what* to speak into the robot's physical action policy is essential for spatially coherent human-robot collaboration. That work, however, treated the content of utterances as largely fixed: the robot could learn *timing* and *context* for communication, but the *spatial information* conveyed in those communications was predetermined by a human designer. That is, *how* the robot decides *what* spatial information to communicate—and *how* to translate raw sensor data into language that is meaningful to a human partner—remained unaddressed. This chapter turns to that complementary challenge, which we term **spatial grounding for communication**: the ability to extract, interpret, and convey spatial context in ways that directly support human decision-making and robot task execution. We investigate this problem through two distinct but complementary lenses, each instantiated in a separate research contribution and corresponding to a different output modality of spatial grounding.

The first perspective, presented in Section 4.2, concerns **verbal spatial grounding**: translating the robot's internal representation of the environment and decision-making for plans into spoken language that keeps a human partner spatially aware during nav-

igation. We realize this in the context of assistive robotics, where we develop a dialog system for a robotic guide dog that serves visually impaired individuals. This domain makes the stakes of spatial communication clear—the human partner has no visual access to the environment and must rely on the robot's verbalizations to form a coherent mental map of the space and to make informed navigation decisions. The system, published at AAAI 2026, introduces two complementary verbalization strategies: *plan verbalization*, which communicates candidate navigation plans before movement begins, and *scene verbalization*, which narrates environmental context during navigation.

The second perspective, presented in Section 4.3, concerns **visual spatial grounding**: using perception of the physical environment to generate task-motion plans that are both semantically valid and physically executable. Here, the output of grounding is not language but action: the robot must determine *where to stand* and *how to arrange objects* in response to underspecified human requests such as "set the dinner table." We address this through LLM-GROP [111], a framework that combines an LLM common-sense with a computer vision module for visually grounded base-position selection. Critically, the contribution described in this chapter emphasizes the *real-world validation* of this framework, demonstrating that visually grounded planning transfers to physical robot hardware operating under the uncertainty and partial observability of real environments.

Table 4.1 summarizes the structural parallel between the two contributions. Both begin with sensor-level perception of the physical world, both transform that perception through a combination of learned representations and language model reasoning, and both output information that directly enables human-robot collaboration. They differ in the *recipient* of that output: the human partner in the case of verbal grounding, and the

Table 4.1: Structural comparison of the two spatial grounding contributions.

	Verbal Spatial Grounding (Section 4.2)	**Visual Spatial Grounding (Section 4.3)**
Domain	Assistive navigation	Mobile manipulation
Perceptual input	Semantic map, navigation plans	Top-down RGB image
Grounding mechanism	LLM + task planner	CNN heatmap + LLM
Output	Spoken language	Task-motion plan
Recipient of output	Human partner	Robot action system
Key challenge	Spatial awareness for VI users	Feasibility under uncertainty
Evaluation	Human study (VI users) + simulation	Real-robot trials + user ratings
Publication	AAAI 2026	IJRR 2025

task-motion planner in the case of visual grounding.

Together, these two contributions establish the empirical and technical foundations on which Chapter 5 builds. As we discuss in Section 4.4, both systems share a common limitation: they are *user-agnostic*. The verbal spatial grounding system delivers the same style of information to every handler; the visual spatial grounding system pursues the same objective function regardless of individual preferences. Neither system adapts its communication strategy to the particular person it is working with. This limitation, identified precisely through the work of this chapter, motivates the adaptive rapport-building framework proposed in Chapter 5.

4.2 Verbal Spatial Grounding: Dialog for Collaborative Navigation

4.2.1 Motivation

Guide dogs play a crucial role in the lives of visually impaired individuals, enhancing their independence, confidence, companionship, and mobility [112]. Unfortunately, breeding and training guide dogs takes multiple years and can be very costly [113], and even then, guide dog training centers have less than a 50% graduation rate [114]. At the same time, many people with visual impairments need guide dogs but have difficulties caring for them, e.g., walking miles a day and dog allergies. As a result, only a very

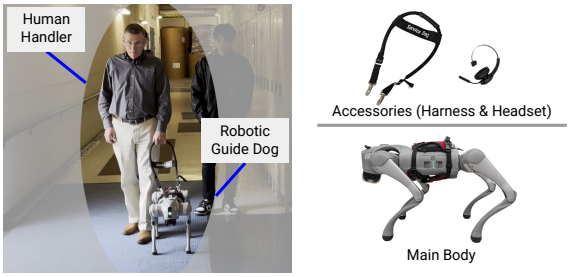


Figure 4.1: A legally blind person walking with our robotic guide dog system during a participant study.

small portion of the visually impaired use guide dogs, e.g., about 2% in the U.S. [114] and only about 400 guide dogs for more than 10 million visually impaired people in China [115].

While biological guide dogs are helpful in leading the visually impaired individuals around obstacles, they have fundamental limitations in understanding and communicating with humans. Even after systematic training, guide dogs can only respond to very short, predefined verbal cues [116]. Robotic guide dogs have the potential to be valuable navigation aids for visually impaired individuals [117, 118, 119, 120, 121, 122, 123, 124]. To the best of our knowledge, there was no existing research focusing on natural language interactions for robotic guide dogs.

The integration of dialog systems into robotic guide dogs introduces unique challenges not present in virtual dialog agents. First, it is important for the dialog system to ground open-vocabulary language to the current environment, to generate relevant and feasible navigation goals, as well as navigational plans for realizing the goals. For instance, when a handler says "*I am thirsty*", the dialog system should provide options that are available, e.g., by saying "*There is a vending machine that serves water bottles and a water fountain on this floor*", and identify the handler's preference after providing additional information, such as "*Navigating to the water fountain requires going*

through a long corridor and entering a kitchen room.". A second challenge is ensuring the spatial awareness of the visually impaired handler, which is crucial for the handler to make informed decisions about routes and destinations.[1]

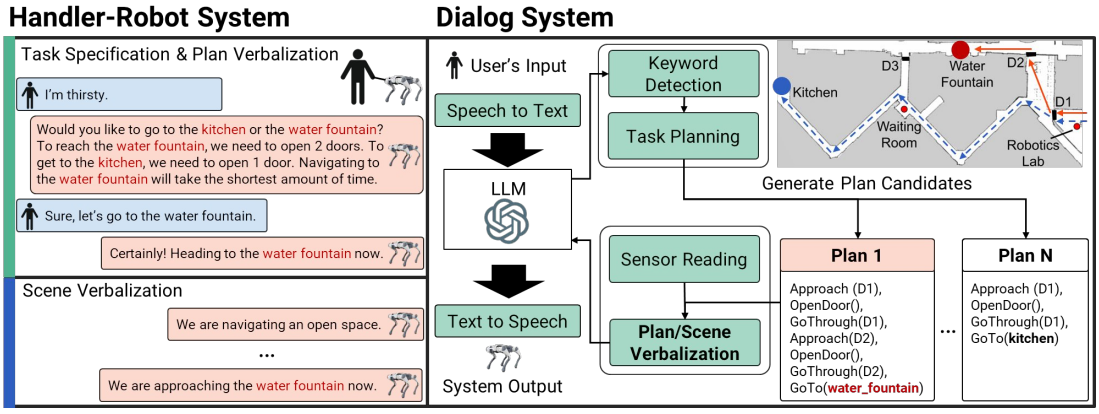


Figure 4.2: Our system uses human-robot dialog to define a formal service task. An LLM first determines relevant navigable locations, and a task planner generates multiple action sequences (plans) for each candidate. These plans are summarized for the human via **plan verbalization**, detailing metrics like navigation cost and door openings. After human selection, the robot executes the chosen plan, providing navigation guidance while providing **scene verbalization** to describe the surroundings.

In this work, we develop a dialog system that leverages an LLM [40] for dialog management and extracting navigation tasks from open-vocabulary service requests. We integrate task planning into language synthesis to enrich LLM responses with candidate *navigation plans*, which is referred to as **plan verbalization**. After the handler-robot team agrees on a navigation task, the robot will take the sequence of actions generated by the planner, e.g., open door X, enter corridor Y, and approach room Z, to fulfill the service request. During navigation, our dialog interface verbalizes scene descriptions to improve the human handler's spatial awareness and facilitate collaborative decision-making, referred to as **scene verbalization**. The plan verbalization (before navigation) and scene verbalization (during navigation) capabilities together identify

[1] A common misunderstanding is that a visually impaired handler completely follows the biological guide dog in navigation. The reality is that they work as a team, where the handler makes decisions about the routes (e.g., using tactile and echolocation sensing), and the guide dog helps avoid obstacles.

the main novelty of our proposed dialog system for robotic guide dogs. We demonstrate the capabilities of our guide dog robot (see Figure 4.1) through a human study with visually impaired individuals. This study, which involves real-world indoor navigation tasks, allowed us to evaluate the system's performance and analyze its potential societal contributions. To complement these real-world findings, we also conducted simulation experiments to quantify the accuracy of our dialog system in identifying navigation tasks from open-vocabulary language.

4.2.2 Related Work

Robotic Guide Dog Systems have been developed to assist visually impaired individuals in navigating, thanks to reduced hardware costs and recent advances in quadruped locomotion control. Systems of this type perform human-robot co-navigation tasks, where a human holds a leash to establish a physical connection with a quadruped robot [125, 124]. Some of these systems only support unidirectional robot-to-human communication via the forces the human feels from the leash [117, 118, 119]. In those systems, the robot helps the human avoid obstacles at the motion level, and the human completely follows the robot during navigation, assuming both sides know the goal ahead of time. Other systems allow for bidirectional communication, such that the human can indicate desired navigation directions via buttons [120], predefined voice commands [121, 126, 127], or tugs on the leash [122] during navigation. Compared with those robotic guide dog systems, ours further supports human-robot communication via bi-directional, multi-turn dialog.

Robot Dialog Systems have greatly improved the interaction capabilities of robots [76, 128, 129]. These systems are optimized for different goals, including mediating the

human-robot perception gap [130], learning from dialog experiences [131], social dialog [132], learning navigation actions from dialog histories [73], and understanding ambiguous human instructions [74]. Earlier works have focused on constructing formal navigation actions from natural language instructions [98] or commands [99]. While effective for various dialog tasks, none of the above works were designed for the context of guided navigation for visually impaired individuals, where communicating navigation goals, planning to achieve them, and maintaining spatial awareness are critically important. Such challenges motivated the development of dialog systems for robotic guide dogs in this paper, where we leverage task planners [133, 134] and LLMs [135, 136] to facilitate human-robot dialog on navigation tasks in open vocabulary.

Non-robotic Navigation Aids were developed to help visually impaired people avoid obstacles, e.g., through belts [137], dialog systems [138, 139, 140, 141, 142], tele-assistance [143], and smart canes that support haptic feedback from the users [144]. None of those systems were able to ground ambiguous human requests to navigation plans or used a robot for physical guidance, which is believed to be more effective in a recent questionnaire study [123]. This paper showcases a novel dialog system for robotic guide dogs with the capabilities of language synthesis (using a task planner and LLMs) for plan and scene verbalization in handler-robot co-navigation tasks.

4.2.3 Problem Settings and Methodology

In this section, we present our problem setting and robotic guide dog system as summarized in Figure 4.2.

Problem Settings and Assumptions

In this paper, we assume the robot has full knowledge about the domain, including a map of the environment associated with semantics information (e.g., labels and coordinates of kitchens and conference rooms). We also assume access to a stable quadruped locomotion controller that can accurately track linear and angular velocity commands. On top of the locomotion controller, the robot is equipped with a global path planner for 2D trajectory planning, and a local path planner that minimizes the deviation from the global path while avoiding obstacles. Those software packages are widely used and publicly available in the Robot Operating System (ROS) community [145]. We do not discuss scenarios where the human and robot together explore a new environment, the human helps the robot dog recover from falls, or the robot disobeys human commands, which occurs on biological guide dogs.

Our goal is to develop a robotic guide dog system that fulfills a handler's request by recommending and executing navigation actions based on the service request in spoken language. This involves developing a dialog system that can map the service request to known locations in the environment, as well as generating an action sequence leading the handler-robot team to the goal location.

Our Dialog System for Robotic Guide Dogs

Our dialog system combines the flexible natural language understanding capabilities of an LLM with the long-horizon reasoning of task planners. This integration allows the system to engage in dialogue with users to specify navigation tasks and to deliver contextual, task-relevant information throughout execution. A key contribution of our work is the ability to verbalize robotic guide dog's outputs, e.g., scene understanding

Procedure 1 Robotic Guide Dog Dialogue and Navigation

{Dialogue for Task Specification}
Greet the guide dog handler to initiate the conversation
while navigational plan is not finalized **do**
    Receive the handler's service request
    Convert the service request to Task T in a formal language
    Enter T into a task planner and compute task plans **P**
    for each plan $P \in$ **P do**
        Verbalize plan P using an LLM {Plan Verbalization}
        if User confirms a plan **then** Break the while loop
    end for
    Request the user to rephrase the service request
end while
{Plan Execution (concurrent): Thread 1 on navigation}
for each navigation action a in plan P **do**
    Execute action a
end for
Return: arrival at the final destination

{Plan Execution (concurrent): Thread 2 on dialog }
while a scene change is detected or there is a long silence **do**
    Collect surrounding objects O using vision and world map
    Verbalize O with spatial info {Scene Verbalization}
end while

and decision-making process, into clear, accessible language. This verbal feedback equips visually impaired users with the critical information needed to make confident, informed decisions about their own navigation.

We first introduce how our system specifies a task and verbalizes plans based on the task, then describe the scene verbalization process during navigation. The control loops are shown in Procedure 1.

Plan Verbalization for Task Specification: Navigation begins when the user issues a request to task T in natural language through a headset microphone. The speech is converted to text using a speech-to-text model [146] and fed to our LLM agent. The agent is set with a system prompt (Figure. 4.3) that defines its role as a guide dog, lists valid navigation locations, and specifies the expected output formats.

If the user's command is ambiguous (e.g., "I'm thirsty"), the LLM clarifies the intent

through dialogue. Instead of merely listing potential destinations (e.g., "kitchen" and "water fountain"), our system performs a unique process we call Plan Verbalization. For each potential destination, the system runs a symbolic planner (detailed in Section 4.2.3) in the background. The planner computes an optimal action sequence and extracts information about physical constraints, such as the total cost (i.e., travel time) and the number of doors that must be opened.

The LLM then incorporates this information into its response, generating an informative prompt like, "*We can go to the kitchen or the water fountain. The kitchen requires opening one door and will take about three minutes. The water fountain has no doors and will take about one minute. Where would you like to go?*" By verbalizing plan details, we enable the user to make an informed decision based on their preferences and circumstances. Since LLMs often struggle with spatial reasoning [147], our dialog system employs a task planner to compute plans for achieving navigation goals by reasoning about both real-world constraints and the robot's action primitives. Once the user confirms their choice, the LLM generates a final formal task and a conversational statement to naturally conclude the dialogue (e.g., "*Okay, I will guide you to the kitchen.*"), ending the dialogue phase.

Scene Verbalization for Spatial Awareness: Once a destination is confirmed and navigation begins, the system continues to provide information to the user through Scene Verbalization *during* navigation. This is triggered by major scene changes or a long silence in the dialogue (see Procedure 1), such as when the robot's position crosses the boundary of a semantically labeled region on the map (e.g., moving from a "corridor" to a "kitchen").

When the robot enters a new area, a message describing the current scene is played

Robot Guide Dog Command Protocol

You are a robot guide dog assisting a visually impaired person in navigating an indoor environment. The human is holding a rigid connection attached to you. Your task is to navigate to the requested locations based on the human's commands. When the human makes a request, you must engage in communication to clarify their intentions. Ask clarification questions until you are confident about the request.

Command Syntax:

There is one command type: `goto <parameter>`

Valid Parameters:

- robotics lab
- staircase
- elevator
- conference room
- bathroom
- waiting room
- water fountain
- kitchen

Handling Requests:

- **Unavailable Locations:** Do not navigate to locations not on the list. If an unavailable location is requested, you must state this in your clarification question.
- **Ambiguous Requests:** If a request could refer to multiple valid locations, you must suggest each of them in your clarification question.

Required Output Format: After determining the correct action, provide your output in one of the two formats below.

To ask for more information:

```
CLARIFICATION QUESTION: <question>
```

To execute a command:

```
COMMAND <type> <parameter>
conversational statement
```

Figure 4.3: LLM prompt defining the role of our robot guide dog dialog system. Given possibly ambiguous human service tasks in natural language, the LLM must conduct a dialog and select the location that best satisfies the task.

through a speaker, e.g., "*We are navigating in a long corridor... We are approaching an office door... We just arrived at an office door.*" This allows the visually impaired user to maintain awareness of their surroundings during navigation.

While our current implementation uses a simple strategy, we position it as a first step, drawing on the rich literature on verbalization in human-robot co-navigation [148]. We focus on evaluating its effectiveness in a human study in this paper, and plan to investigate more advanced methods, such as learning verbalization policies, in future work.

Implementation Details

Safeguard for Dialog Management: Due to the uncertainty of LLM responses and the lack of spatial awareness, we employ a language synthesis module with a safeguard that post-processes the authentic LLM responses before delivering them to the human user. Specifically, during the plan verbalization phase, if the LLM does not respond in the expected format (e.g., no "CLARIFICATION QUESTION" or "COMMAND" indicators in the response), or if it fails to suggest any valid locations as determined by keyword detection, then we replace the LLM response with the message:

> *"I can only assist with navigation requests of nearby locations that I know about. Would you like help navigating to somewhere nearby?"*

Plan Generation for Navigation: To complete navigation tasks specified by our dialog system, we generate navigation plans via an Answer Set Programming (ASP) planner [52]. We leverage planners designed for navigation tasks [149] to determine high-level action sequences. We consider the set of actions:

```
approach(door_id),
gothrough(door_id),
opendoor(door_id),
goto(location)
```

Arriving at the destination specified by the dialog system involves generating a sequence of actions while minimizing their total cost. Action costs between locations are considered and measured based on navigation distances, along with an additional constant cost added for door opening actions. Rules in the ASP encode the necessary domain knowledge for navigation. For example, the following rule states that P1 can be

accessed by P2 via door D, if P1 and P2 are unique locations, D is a door, and P1 and P2 both have door D.

```
dooracc(P1,D,P2) :-
        hasdoor(P1,D), hasdoor(P2,D),
        P1 != P2, door(D),
        location(P1), location(P2).
```

LLM-based task planning methods are increasingly mature [150, 151, 134, 111], and robotic guide dog practitioners can feel free to implement the task planner using LLMs instead. Next, we present a human study for evaluating the effectiveness of our dialog system in collaborative settings between a human handler and a robotic guide dog (Section 4.2.4), followed by an extensive evaluation in simulation focusing on accuracy and efficiency (Section 4.2.5).

4.2.4 Real-World Effectiveness Evaluation: Human Study

In this section, we detail the human study, which was designed to test the following primary hypothesis: Combining scene and plan verbalization improves user satisfaction and perceived effectiveness for robot-assisted navigation.[2]

Participants

Table 4.2 shows the demographics of the participants. Seven legally blind individuals (5 Male, 2 Female; age range 40–68) were recruited for the study.[3] One participant who completed the study self-identified as not meeting the vision requirement and was

[2]This section focuses on presenting the results of our robotic guide dog system under *controlled autonomy*. The system under *full autonomy* is demonstrated on the project webpage.

[3]A person is considered legally blind if they meet one of the following criteria in their better-seeing eye, even with the best conventional correction: 1) Visual acuity of 20/200 or less; 2) Visual field of 20 degrees or less.

excluded from the final analysis. Two of the participants had prior experience with a guide dog.

Age	Gender	Impairment Duration (yrs)	Guide Dog Experience
68	Female	68	None
68	Male	60	None
59	Male	51	8 yrs, 2 dogs
40	Male	40	None
64	Male	35	18 yrs, 3 dogs
57	Male	20	None
52	Female	15	None

Table 4.2: Participant Demographics.

Experimental Setup

The experiment took place in a large, many-room office environment. The task involved navigating a route from an entrance to a conference room. We used a Unitree Go2 as a quadruped robot platform and integrated it with our dialog system. For safety reasons during the human study, the robot's movements were controlled by an expert operator in a Wizard of Oz setup [152]. This approach ensures safety and does not affect the evaluation of the effectiveness of the dialog strategies.

After signing a consent form, participants were briefed on the task. They would request to go to a conference room, and the robotic guide dog would guide them under three distinct verbalization conditions in a within-subjects design:

- **Minimum Verbalization:** A control condition with language-based task specification before navigation and then no verbal interaction during navigation.
- **Scene Verbalization:** The robot only verbalized information about the immediate surroundings (e.g., obstacles and environmental context).

Q1 Utility	The natural language interface is a useful component of the system.
Q2 Helpfulness	The robot is helpful for visually impaired navigation when compared to having no tools for visually impaired navigation.
Q3 Helpfulness	The robot is helpful for visually impaired navigation when compared to having other tools for visually impaired navigation.
Q4 Safety	The robot is capable of safely guiding me to a navigation goal.
Q5 Ease of comm.	I can easily communicate with the robot.
Q6 Preference	Given the same navigation task we performed today I prefer a guide dog to guide me instead of the robot.
Q7 Feedback	If you have any questions, comments, or concerns please leave them here.

Table 4.3: The list of questions in the questionnaire. The items assessed interface utility, helpfulness, safety, ease of communication, and preference relative to a real guide dog.

Condition	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**
Mini. Verb.	4.33	4.67	3.50	**4.00**	4.33	3.17
Scene Verb.	4.67	**4.83**	4.00	**4.00**	4.00	3.17
Ours	**4.83**	**4.83**	**4.33**	3.83	**4.50**	**3.67**

Table 4.4: Average survey ratings for each verbalization condition. Ratings are on a 5-point Likert scale (1: Strongly Disagree, 5: Strongly Agree).

- **Ours (Scene + Plan):** The robot provided both scene and plan verbalization.

Given the age and the energy span of our participants, we decided to keep the trials and questionnaire within 30 minutes, and evaluate no more than three methods. After completing the task under each condition, participants filled out a 7-item questionnaire shown in Table 4.3. Q1 through Q6 were rated on a 5-point Likert scale, and Q7 was an open-ended feedback question. Participants self-administered the questionnaires after task completion, while we provided support by reading the instructions and questions.

Results

Figure 4.4 presents an instance of the experiment trials in the human study. Table 4.4 shows the average ratings for each condition. Our proposed method, **Ours (Scene +**

Plan), received the highest ratings on natural language interface utility (Q1). The result indicates the robot's strong performance in helpfulness (Q2 and Q3). Furthermore, our method outperformed in ease of communication (Q5) and preference over real guide dogs (Q6). The overall result suggests that providing both scene and plan information yields higher navigation effectiveness and a user-friendly approach.

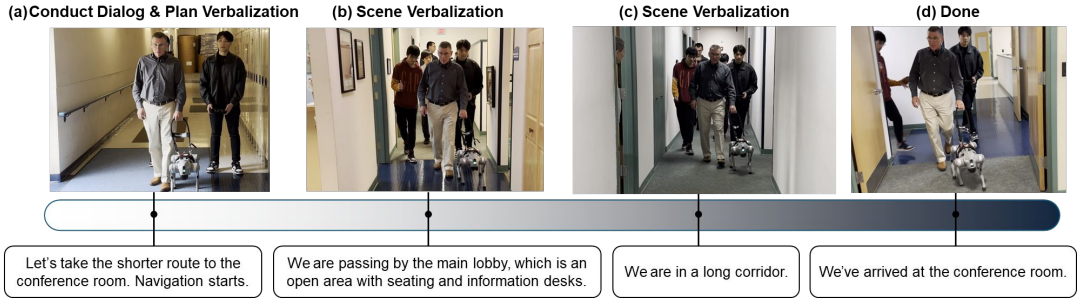


Figure 4.4: An illustrative example of our robotic guide dog assisting a legally blind participant to navigate to a conference room. The sequence shows: (a) The robot verbalizes the generated navigation plans for a user request, and starts navigation after the user chooses one of the plans. (b, c) During navigation, the system provides real-time scene verbalization, such as passing a lobby and entering a corridor. (d) The robot announces that they have arrived at the destination.

Our method scored slightly lower on perceived safety (Q4: 3.83) than the other conditions. Qualitative feedback (Q7) suggests that the reason may be initial unfamiliarity with walking with the robot. The human study confirms the practical value of our proposed natural-language interface for robotic guide dogs.

Category	Locations
Places to eat	food court, cafeteria, restaurant, dining hall, shawarma joint, cafe, vending machine, food truck, fast food restaurant
Restroom	restroom, bathroom, the loo
Change Floor	elevator/stairs, elevator/escalator, elevator
Cut Hair	barber shop, salon

Table 4.5: Samples in our collected service task dataset: locations in each category considered functionally equivalent.

4.2.5 System Performance Evaluation: Simulation Analysis

While the human study demonstrated the real-world usability of our approach, we conducted a large-scale simulation analysis to test the underlying system's performance. This analysis tests three hypotheses: our system can accurately and efficiently resolve ambiguous service requests (**Hypothesis I**), is robust to noisy speech recognition (**Hypothesis II**), and can leverage a navigation planner to reduce travel costs (**Hypothesis III**). Inspired by recent work using LLMs for data synthesis [153, 154], we simulated a visually impaired user with GPT-4 [40]. To ground our simulation in reality, the user's goals were based on a library of navigation-related service requests collected from human participants, as described next.

Simulation Setup

Service Task Library: To create a realistic task library, we surveyed 16 university students (ages 19-30). Participants were asked to imagine being a visually impaired person who needs help navigating and provide five `<location, purpose>` pairs for potential navigation requests, such as (*"water fountain"*, *"I want something to drink"*). After filtering three overly vague entries, we obtained 77 unique samples. We normalized location names and grouped functionally equivalent locations (e.g., "cafeteria," "dining hall"), as shown in Table 4.5. The unique locations can be inserted in the system prompt as the set of valid parameters in Figure 4.3. We report accuracy and efficiency for evaluation.

Simulated Human: We used GPT-4 to simulate a visually impaired user, instructing with the following prompt:

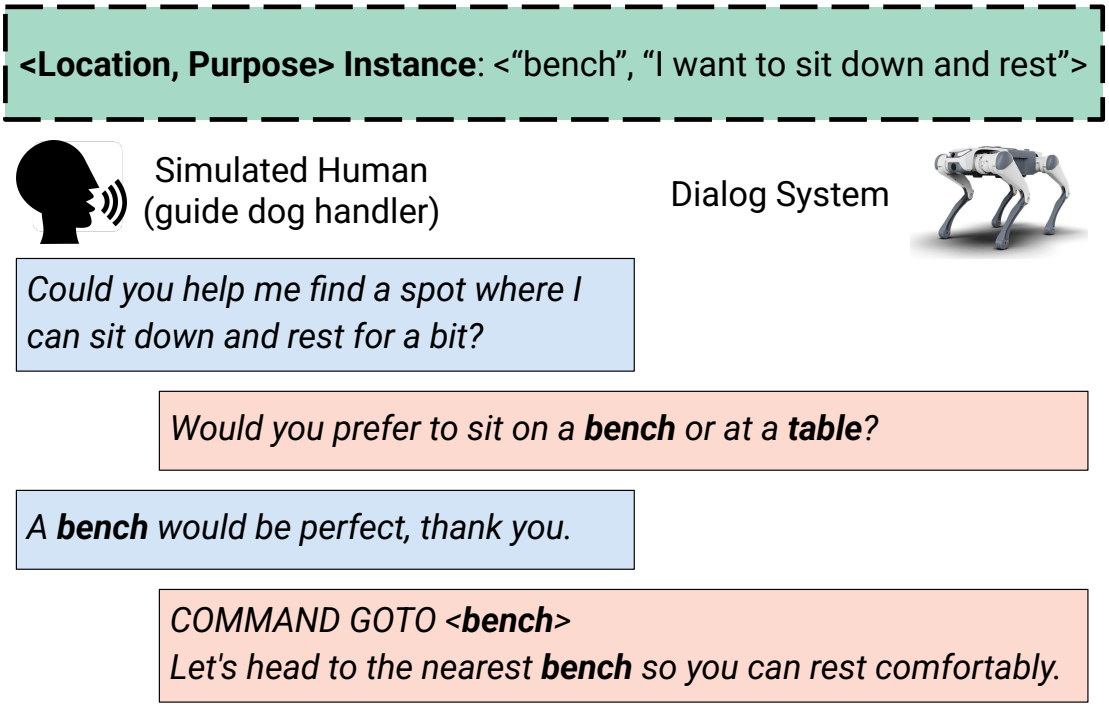


Figure 4.5: An example simulated conversation. The handler implicitly states a purpose sampled from `<location, purpose>` pair. The dialog system suggests relevant locations. After confirming the handler is looking for a bench, the system generates a formal task and concludes the task specification conversation.

> "*You are an AI chatbot pretending to be a visually impaired person. You need to navigate to* `location`*. Do not say the name of where you want to go, unless asked. You are only interested in going to* `location` *and must not show interest in any other locations.*"

This prompt yields ambiguity by discouraging the direct mention of the location. Each conversation was initiated by providing the simulated human with the `purpose` from a `<location, purpose>` pair (e.g., *"Begin the conversation by indicating that you want something to drink"*). An example dialog is shown in Figure 4.5.

Evaluation of Accuracy and Efficiency

To test **Hypothesis I**, we simulated conversations for all 77 task pairs and compared three approaches:

1. **Keyword-Based**: A multi-turn template-based dialog system using keyword detection. If no keywords are found, it lists all known locations.

2. **Single-Turn**: Our LLM-based system, but restricted to a single turn (no clarification questions).

3. **Ours**: Our full multi-turn LLM-based dialog system, described in Section 4.2.3.

For multi-turn approaches, dialogs were limited to six turns; exceeding this limit counted as a failure. A trial was successful if the dialog concluded with the correct (or functionally equivalent) formal task. We measured accuracy and efficiency by the dialog length and its success rate.

As shown in Figure 4.6, the result highlights the trade-offs between each approach in terms of accuracy, average number of dialog turns. It achieves the highest accuracy (94.8%) while being more concise than the Keyword-Based baseline, which is overly verbose as it often must enumerate all locations. These results support **Hypothesis I**.

Evaluation of Robustness to Noisy Speech

To test robustness to speech recognition errors that may arise from speech-to-text models (**Hypothesis II**), we simulated noisy input by perturbing each character in the user's responses with 0.3 probability, randomly applying a deletion, insertion, or substitution. This is particularly relevant to a guide dog setting, in which the visually impaired person can navigate in crowded and noisy environments.

Figure 4.7 shows the results that suggest our dialog system is robust to perturbations, only losing 5.2% accuracy, and marginally increasing dialog length. In contrast, the Keyword-Based's performance collapsed, as it failed to match the perturbed keywords. This supports **Hypothesis II**.

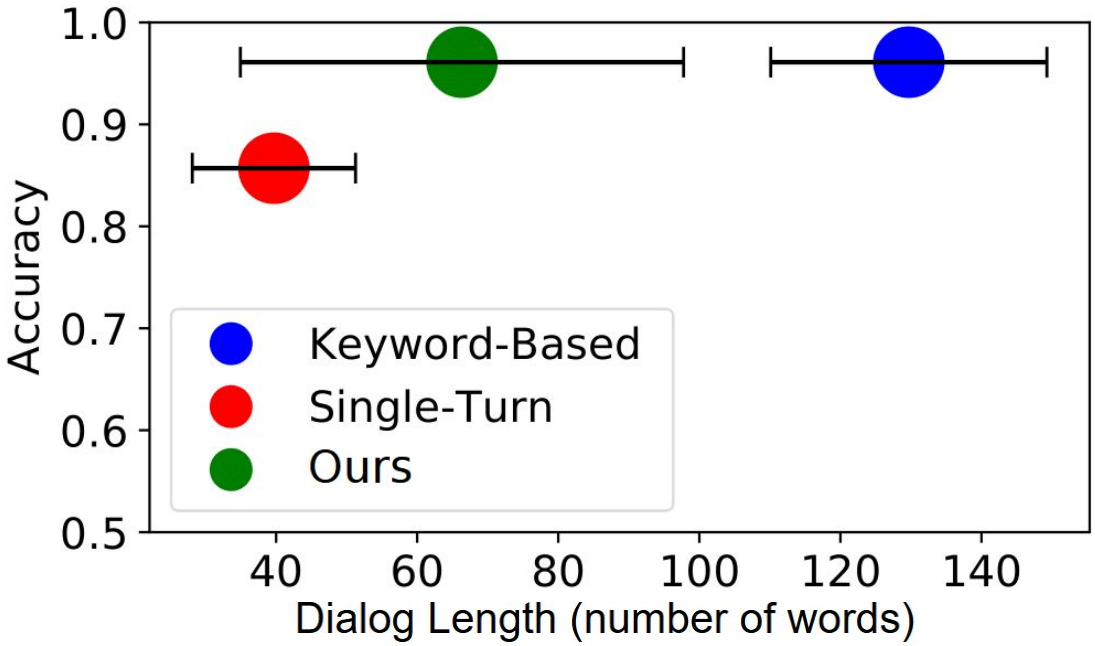


Figure 4.6: Accuracy vs. Average Dialog Length without noise. Our approach provides the best trade-off between high accuracy and efficient conversation.

Location	Purpose
elevator	"I want to go to the 2nd floor"
water fountain	"I want something to drink"
waiting room	"I want to sit down"

Table 4.6: The three location-purpose pairs used for planner-informed evaluation.

Evaluation of Planner-Informed Dialog

Finally, we evaluated if providing plan information (e.g., travel distance) in the dialog helps the user choose more efficient routes (**Hypothesis III**). We used three tasks (Table 4.6) where multiple functionally equivalent destinations existed at different distances.

Method	Dialog Len.	Nav. Cost	Total Time
w/ Plan Info	115.67	**55.17**	**230.15**
w/o Plan Info	**84.83**	73.00	277.27

Table 4.7: Effect of Plan Information on Performance Metrics

We compared our system with and without plan information. Table 4.7 shows that

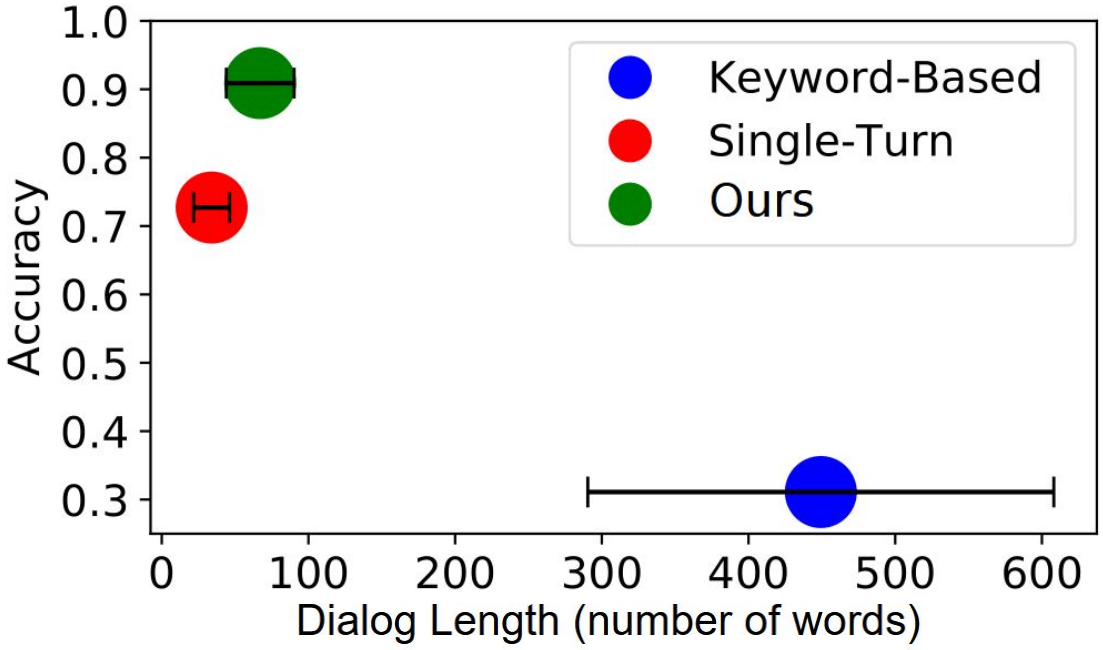


Figure 4.7: Accuracy vs. Dialog Length with perturbed (noisy) inputs. Our approach remains highly accurate, while the keyword-based method fails.

when the dialog included navigation costs, the simulated user consistently chose the closer destination. Although this increased dialog length, the reduction in navigation cost led to a lower overall task time (calculated assuming 2.5 words/sec talking and 0.3 m/s walking speeds). These results support **Hypothesis III**.

4.2.6 Summary

In this paper, we developed a novel robotic guide dog system that can verbally convey both navigation plans and environmental scenes using a task planner and an LLM. The effectiveness was validated through two evaluations. The human study with visually impaired participants confirmed that combined plan and scene verbalization is an effective and preferred communication strategy. The simulation analysis supported this, demonstrating high accuracy, robustness, and efficiency by leveraging planner knowledge, which accounts for the high user satisfaction.

Despite these promising results, the human study highlighted key challenges, particularly a lower safety rating compared to biological counterparts, underscoring the need for further enhancements. Future work will involve expanded user studies, increased system autonomy, and deployment in more complex environments. The ultimate goal is to integrate robotic guide dogs into daily life, enhancing independent navigation for both visually impaired and sighted individuals.

Ethical Statement The research protocol involving human participants was reviewed and approved by the Institutional Review Board (IRB). Seven legally blind individuals were recruited for the study. All participants provided written informed consent after being fully briefed on the experimental task and were compensated for their time. Participants were explicitly informed that their participation was voluntary and that they retained the right to withdraw from the study at any time without penalty. To ensure participant safety during the real-world navigation component, a Wizard-of-Oz approach was used. In this setup, an expert operator remotely controlled all physical movements of the robot. This protocol was implemented to mitigate risks and isolate the evaluation to the system's dialog strategies. Researchers provided support by reading instructions and questionnaire items to participants. Data was stored securely in accordance with institutional data protection policies. This work contributes to the field of assistive robotics, which focuses on the well-being of people with disabilities. The primary objective of this research is to enhance the independence and mobility of visually impaired individuals by addressing the limitations posed by the scarcity of guide dogs and the challenges of their care.

4.3 Visual Spatial Grounding: Perception for Physically Executable Planning

4.3.1 Motivation and Problem Statement

The previous section examined spatial grounding from the perspective of *language output*: how a robot can communicate spatial information to a human partner in a form that supports decision-making. This section examines the complementary direction: how a robot can use its environmental spatial perception to generate *physically executable plans* in response to underspecified human requests.

Consider a service robot tasked with "setting a dinner table with a fork, knife, and plate." This request is underspecified in two distinct ways. First, it does not specify the *goal configuration*: where exactly should each object be placed? Answering this question requires common-sense knowledge about dining conventions (forks to the left, knives to the right) that is not encoded in the geometric representation of the environment. Second, even given a valid goal configuration, the request does not specify *how* the robot should physically execute it. A mobile manipulator must navigate to each object's starting location, pick it up, navigate to the target table, and place the object—all while avoiding dynamically placed obstacles such as chairs. The feasibility of each placement action depends critically on *where the robot stands* relative to the table, which in turn depends on the current positions of obstacles that were not present when the environment map was created.

This problem, object rearrangement, is an instance of **mobile manipulation task and motion planning (MoMa-TAMP)**: computing interleaved sequences of navigation and manipulation actions that achieve a high-level task goal while respecting low-

level geometric constraints [47, 155, 156, 57, 157, 158]. Existing TAMP methods are ill-suited to this problem for two reasons. First, they typically assume the goal configuration is fully specified as input; they provide no mechanism for inferring semantically valid arrangements from natural language [68]. Second, they do not incorporate visual perception of the current obstacle configuration into base-position selection, leading to plans that are infeasible when chairs or other movable obstacles are present in positions that differ from those assumed during planning [159, 160].

4.3.2 Method Overview: LLM-GROP

We address the MoMa-TAMP problem through LLM-GROP [111], a framework that integrates three components: an LLM for common-sense goal generation, a computer vision module (GROP) for visually grounded base-position selection, and a TAMP system for plan computation and execution.

LLM-Guided Goal Generation

Given an underspecified service request, LLM-GROP first queries an LLM to determine the symbolic spatial relationships between the objects to be arranged. Using a template-based prompt that specifies valid spatial relations (e.g., *to the left of*, *above*), the LLM generates instructions such as "Place the fork to the left of the plate; place the knife to the right of the plate." Logical consistency of the generated relationships is verified using Answer Set Programming (ASP), which detects contradictions (e.g., simultaneously specifying that object A is below and to the right of object B) and re-prompts the LLM when inconsistencies are found.

A second LLM query converts symbolic relationships into geometric constraints, extracting distances in centimeters from the model's responses. These geometric con-

straints define the target positions for each object in the coordinate frame of the target table surface.

Visually Grounded Base-Position Selection (GROP)

The feasibility of manipulating an object at a given table position depends on the robot's standing position (base position). If the robot stands too close to the table, its base may collide with the table or with chairs around it. If it stands too far, the arm cannot reach the target position on the table. The optimal base position is therefore a function of the current obstacle configuration, which changes between trials as chairs are placed differently.

GROP addresses this by learning a function $f : \mathcal{I} \times \mathcal{O} \rightarrow \mathcal{H}$ that maps a top-down view image $\mathcal{I}$ of the table and its surroundings, together with a target object position $\mathcal{O}$, to a heatmap $\mathcal{H}$ over the robot's 2D operating space. Each pixel in $\mathcal{H}$ represents the estimated success probability of navigating to the corresponding 2D pose and successfully completing the manipulation action. GROP is trained entirely in simulation on a dataset of $\langle$image, target position, heatmap$\rangle$ triples, where heatmap values are derived from rollouts of the robot's navigation and manipulation controllers.

Utility-Based Plan Optimization

Given the object goal positions from the LLM and the feasibility heatmap from GROP, the TAMP system computes the task-motion plan $p^* = \arg\max_p \mathcal{U}(p)$, where the utility function balances feasibility and efficiency:

$$\mathcal{U}(p) = \mathcal{R} \cdot \mathcal{F}(p) - \mathcal{C}(p) \tag{4.1}$$

Here $\mathcal{F}(p) \in [0, 1]$ is the estimated probability that plan p can be successfully executed (derived from GROP's heatmap), $\mathcal{C}(p)$ is the total action cost (primarily navigation distance), and $\mathcal{R}$ is a success bonus. This formulation allows the system to trade off a small reduction in feasibility against a large reduction in travel cost—a trade-off that becomes important in tasks involving many objects at distant locations.

4.3.3 Real-World Validation

Experimental Setup

Simulation results for LLM-GROP are reported in the full publication [111]. Here we focus on the real-world validation, which constitutes the primary contribution of this section and which addresses the critical question of whether visually grounded TAMP transfers to physical robot hardware.

The mobile manipulator used in our experiments is equipped with a wheeled base for navigation and a 6-DOF robotic arm for manipulation. The environment consisted of three tables: two source tables on which target objects were initially placed, and one central target table on which the objects were to be arranged.

Three object rearrangement tasks of increasing complexity were selected for real-robot evaluation, as detailed in Table 4.8. The complete set of 9 tasks (Tasks 1–9) are defined in the full publication [111], but we focus on Tasks 4, 7, and 9 here as they represent the tasks with different object counts (3, 4, and 5 objects respectively) that were evaluated in the real-robot experiments.

Table 4.8: Real-robot experiment tasks and environments.

Task	Objects	Num. Objects	Repetitions/Env.
Task 4	Fruit bowl, mug, strawberry	3	5
Task 7	Plate, fork, knife, strawberry	4	5
Task 9	Plate, fork, knife, water cup, strawberry	5	5

Each task was repeated 15 times under three environment conditions:

1. **Easy Environment:** No obstacles around the target table.
2. **Hard Environment (Chair-Top):** A chair placed on the upper side of the target table (from the robot's viewpoint).
3. **Hard Environment (Chair-Bottom):** A chair placed on the lower side of the target table.

This yielded a total of 45 trials (3 tasks $\times$ 3 environments $\times$ 5 repetitions).

Implementation Details: We use the MoveIt software [161] for planning grasping and ungrasping behaviors on the real robot. To facilitate object localization, QR codes were affixed to each target object. For objects with geometries that are challenging to grasp (forks, knives), a utensil holder was used to orient the objects consistently and increase grasp reliability. The LLM component used OpenAI's GPT-3 (text-davinci-003, temperature 0.1) to maintain consistency with simulation results and with the preceding conference papers on which this work builds.

Evaluation Protocol: Task success was evaluated by capturing an overhead photograph of the table at the end of each trial and asking human raters to assess the arrangement. Ten graduate students with engineering backgrounds served as raters. Each image was scored on a 5-point scale (Table 4.9), where 1 indicates a clearly failed or incomplete arrangement and 5 indicates an arrangement that meets the aesthetic standards of an experienced waiter.

Table 4.9: Rating guidelines used for evaluating tableware arrangements.

Score	Description
1 (Poor)	Missing critical items compared with the objects listed at the top of the interface (e.g., dinner plate, dinner fork, dinner knife), making it hardly possible to complete a meal.
2	All items are present, but the arrangement is poor and major adjustments are needed to improve the quality to a satisfactory level.
3	All items are present and arranged fairly well, but still there is significant room to improve its quality.
4	All items are present and arranged neatly, though an experienced human waiter might want to make minor adjustments to improve.
5 (Strong)	All items are present and arranged very neatly, meeting the aesthetic standards of an experienced human waiter.

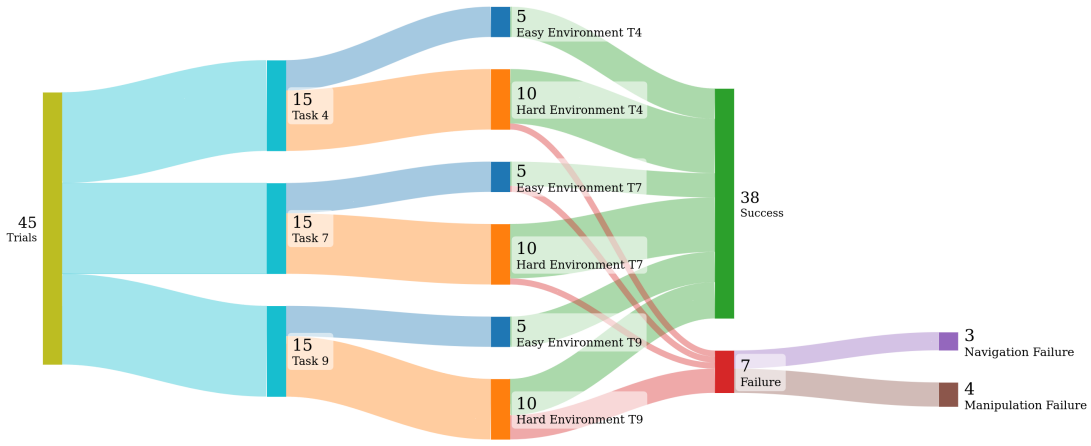


Figure 4.8: Overview of real-robot experiment outcomes. Of 45 total trials across three tasks and three environment conditions, 38 succeeded (84.4%), and 7 failed. Failure modes are categorized as navigation failures (3) and manipulation failures (4).

Results

Overall Success Rate. Across all 45 trials, the robot successfully completed 38 trials, yielding an overall success rate of **84.4%**. The 7 failures decompose as 3 navigation failures and 4 manipulation failures, as illustrated in Figure 4.8.

User Ratings by Task and Environment. Table 4.10 presents average user ratings broken down by task and environment condition. Two patterns are apparent. First,

Table 4.10: Average user ratings (1–5 scale) for real-robot experiments by task and environment.

Task	**Easy**	**Hard (Chair-Top)**	**Hard (Chair-Bottom)**
Task 4 (3 obj.)	3.80	3.60	3.73
Task 7 (4 obj.)	3.53	3.73	3.93
Task 9 (5 obj.)	3.73	3.13	3.06
Average	*3.69*	*3.49*	*3.57*

performance degrades monotonically with object count: Task 4 (3 objects) consistently outperforms Task 7 (4 objects), which outperforms Task 9 (5 objects), reflecting the compounding effect of per-step uncertainty in long-horizon plans. Second, performance degrades in harder environments for Task 9 but not consistently for Tasks 4 and 7, suggesting that the GROP base-position selection module is effective for lower-complexity tasks but that its sim-to-real transfer becomes less reliable as tasks grow more complex.

Qualitative Failure Analysis. Examining the 7 failure trials yields qualitative insight into the gap between simulation and real-world performance. Navigation failures (3 trials) occurred when the robot's localization estimate drifted during the multi-step plan, causing the robot to approach the table from an angle that placed the grasping target outside the arm's reachable workspace. This is a systematic consequence of cumulative localization error: each navigation action introduces a small positional uncertainty, and this uncertainty compounds across the 6–10 navigation actions required for Tasks 7 and 9. Manipulation failures (4 trials) were concentrated in Task 9 in the hard environment conditions, where the presence of a chair forced the robot to adopt a base position farther from the table than in the easy condition. At this increased distance, the wrist of the 6-DOF arm was near its joint limit during ungrasping, occasionally causing the object to be dropped at the edge rather than the center of the target position.

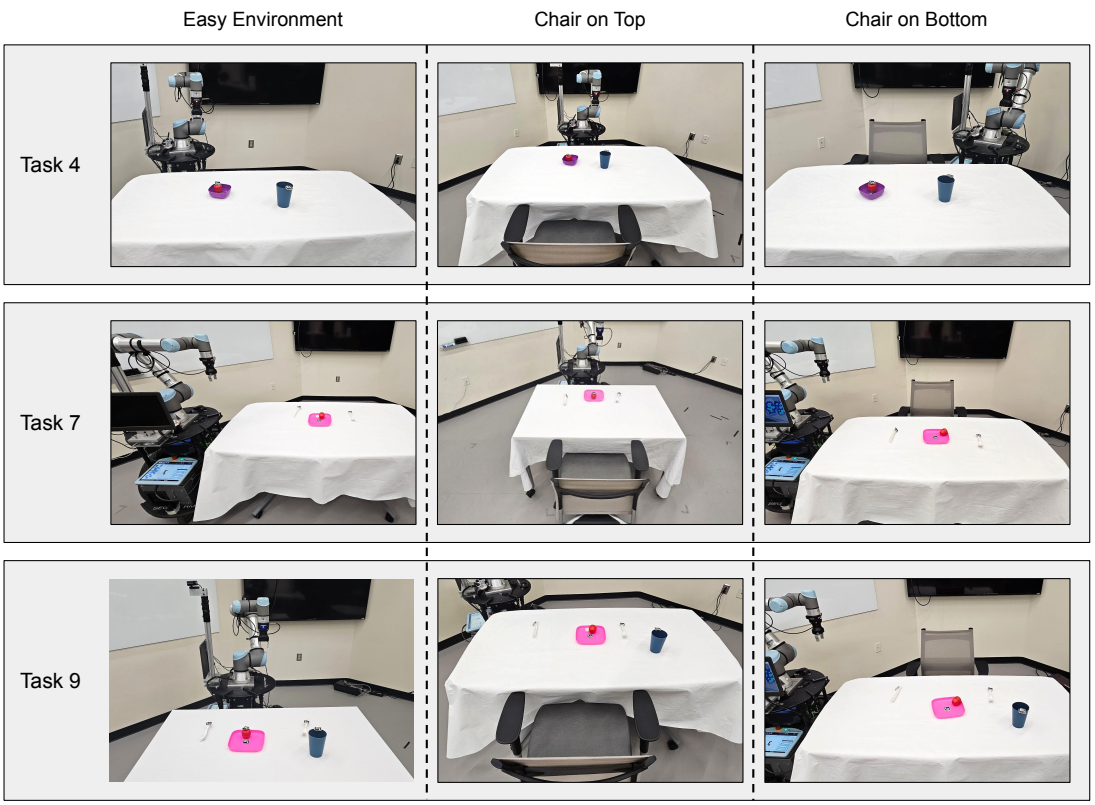


Figure 4.9: Example outcomes of the robot completing object rearrangement tasks across three tasks and three environment conditions. GROP enables the robot to adapt its base position based on obstacle placement, producing visually valid arrangements even in harder environments.

These failure modes are instructive beyond their immediate engineering implications. They reveal that the Sim-to-Real gap in LLM-GROP is not uniform across task types: the system transfers well for short-horizon tasks in obstacle-free environments, but degradation becomes significant when both obstacle complexity and task length are high. This finding motivates the incorporation of online uncertainty estimation and adaptive replanning capabilities in future work.

Effect of GROP on Obstacle Adaptation. Figure 4.9 shows example outcomes of the robot completing Tasks 4, 7, and 9 in the three environment conditions. A key qualitative observation is that GROP's heatmap-based base-position selection enables

the robot to dynamically adapt its standing position based on chair placement. In the Chair-Top and Chair-Bottom conditions, the robot consistently selected approach angles that avoided the chair while maintaining arm reachability, producing arrangements comparable in quality to those achieved in the easy condition for Tasks 4 and 7.

4.3.4 Summary

This section presented LLM-GROP and its real-world validation. The framework combines LLM common-sense for goal generation with visual perception for base-position selection in a unified TAMP framework. Real-robot experiments across 45 trials demonstrated 84.4% task success and confirmed that the GROP module enables meaningful adaptation to dynamically placed obstacles. Failure analysis identified cumulative localization error and proximity to workspace limits as the primary modes of degradation, and revealed that Sim-to-Real transfer is reliable for lower-complexity tasks but deteriorates as task length and obstacle complexity increase simultaneously.

4.4 Discussion: Toward Integrated Spatial-Social Grounding

The two lines of work presented in this chapter have established complementary foundations for spatial grounding in human-robot communication. Before turning to the proposed research in Chapter 5, it is worth examining what the two contributions share, how they are complementary, and most importantly, what they jointly fail to address.

4.4.1 Shared Architecture: Perception → Representation → Collaboration

Despite their surface differences—one involves a quadruped guide dog and spoken dialog, the other a wheeled mobile manipulator and physical object arrangement—the

two systems share a common architectural pattern. Both begin with *perceptual input*: the verbal grounding system uses the robot's semantic map and the navigation planner's computed action sequences, while the visual grounding system uses top-down RGB images of the robot's operating space. Both transform this perceptual input into a *grounded representation*: symbolic navigation plans with cost annotations in the former case, feasibility heatmaps and LLM-generated goal positions in the latter. And both use this representation to support *human-robot collaboration*: enabling informed navigation decisions in the guide dog system, and enabling physically executable manipulation plans in the mobile manipulation system.

This shared pattern—perception, representation, collaboration—establishes a common vocabulary for understanding spatial grounding as a general capability in human-robot interaction, and provides the conceptual foundation on which Chapter 5 builds.

4.4.2 Complementarity: Language and Action as Dual Outputs of Grounding

The two contributions are complementary in a precise sense: they address the same fundamental problem (spatial grounding) but direct the output toward different recipients and in different modalities. The verbal grounding system produces *language* for a human partner; the visual grounding system produces *action plans* for the robot's own execution. Neither system alone is sufficient for a robot that must simultaneously coordinate its physical behavior and its verbal communication. A robot that grounds language but not action will communicate appropriately while executing plans that are physically infeasible or spatially incoherent. A robot that grounds action but not language will navigate and manipulate effectively but fail to give its human partner the information needed to collaborate. Chapter 5 addresses this duality directly by developing a framework in which both outputs—language and action—are conditioned on

the same integrated representation of spatial context and user state.

4.4.3 Shared Limitations: The Absence of Social Grounding

The most significant limitation of both systems, and the one that most directly motivates Chapter 5, is that they are *user-agnostic*: neither system adapts its behavior to the individual it is interacting with.

In the verbal grounding system, this limitation manifests in three ways. First, the system is *user non-adaptive*: every handler receives the same style and quantity of spatial information, regardless of their familiarity with the environment, their cognitive preferences for verbal information, or their prior experience with robotic systems. A user who wants brief, landmark-anchored descriptions receives the same output as a user who wants detailed, continuous narration. Second, the system has no *communication strategy*: it always provides the same type of information in the same way, without reasoning about whether a logical appeal ("the water fountain is faster") or an emotional appeal ("I know you prefer avoiding crowded areas") would be more effective for a particular user. Third, the system maintains no *interaction history*: each session begins from a blank slate, with no memory of the user's past preferences, questions, or navigation patterns.

In the visual grounding system, analogous limitations apply. The system pursues a single utility function (Equation 4.1) regardless of user preferences: it always optimizes the same balance of feasibility and efficiency, produces the same table arrangement style for the same objects, and delivers the same type of feedback about task progress. A user who prefers a traditional Western dining arrangement receives the same output as a user who prefers a minimalist layout; a user who values efficiency (faster completion) is treated identically to a user who values precision (higher placement accuracy).

These three absences—user adaptation, communication strategy, and interaction history—represent not merely technical gaps but a conceptual gap: the distinction between a robot that *processes* spatial information and a robot that uses spatial information as the foundation for building a *relationship* with a specific human partner. The former can assist; the latter can establish trust and rapport.

The proposed work in Chapter 5 addresses precisely this gap, extending the spatial grounding capabilities developed in this chapter with a social grounding layer that learns adaptive communication strategies from interaction history and user persona, in service of long-term rapport building.

5 Adaptive Rapport Building through Spatially and Socially Grounded Dialogue

5.1 Introduction

Chapters 3 and 4 established two foundational contributions: a joint policy learning framework for integrating utterance and physical action (Contribution 1), and a perceptual and planning infrastructure for spatially grounded language expression and real-world mobile manipulation (Contribution 2). Together, these contributions address *when and where to speak* through optimized utterance timing and *how to translate perceived environmental information into meaningful linguistic expression*. However, neither addresses *how to adapt dialogue to the individual user* in a way that builds long-term trust and rapport—the relational outcomes defined in Chapter 1, Section 1.2.

This chapter presents the third and central contribution of this dissertation, which integrates the two foundations above. Specifically, we introduce and empirically validate the **Spatial-Social Language Synthesis (SSLS)** framework—a neural pipeline system for building trust and rapport between humans and robots through adaptive dialogue strategies that are simultaneously grounded in spatial context and social awareness. Figure 5.1 provides an overview of the framework and its relationship to the preceding contributions.

For humans to accept a robot not merely as a functional tool but as a communica-

tive partner that understands their situation, the robot must adapt its utterance content and style to the user's individual context and the prevailing social norms of the interaction. We demonstrate that such adaptation is achievable through an explicit, interpretable pipeline that combines spatial context integration, user-persona adaptation, and dynamic communication-strategy selection, and that the resulting dialogue quality serves as a measurable antecedent to trust and rapport formation. The remainder of this chapter first elaborates on the motivation and research question (RQ3), then details the SSLS framework architecture, the experimental task and dataset, the experimental design, the results of three progressively scaled experiments, a dedicated analysis of trust and rapport, a discussion that reframes the findings, and the limitations of the present study.

5.2 Motivation and Research Questions

5.2.1 Why Trust and Rapport Are Central

As established in Chapter 1, Section 1.2, this dissertation adopts an operational distinction between *trust* (competence-based belief that the robot reliably performs its task) and *rapport* (relationship-based sense of mutual attentiveness, positivity, and coordination). Social HRI research has repeatedly demonstrated that user trust is indispensable for long-term acceptance [3, 2], yet existing dialogue systems are predominantly optimized for external metrics such as information accuracy and task completion rate, and lack mechanisms for intentionally designing the communicative behavior through which rapport emerges.

Critically, trust and rapport are **dynamically formed through interaction**. When a robot consistently recalls user preferences and produces utterances calibrated to the

immediate spatial situation, users form the perception that "this robot understands my situation," and this is the essence of rapport toward which the SSLS framework is directed.

5.2.2 Formulation of RQ3

This chapter addresses the following research question:

> **RQ3:** Building upon spatial and social grounding techniques, how can a robot adapt to individual user preferences and situational context to build trust and rapport with humans?

This question is decomposed into three sub-questions. **RQ3a** asks whether adapting dialogue strategies to user persona — including age, occupation, preferences, and personality traits — improves dialogue quality across dimensions of naturalness, relevance, persuasiveness, and coherence. **RQ3b** asks whether a framework with explicit integration of spatial context and communication strategy selection offers advantages over implicit strategy learning with respect to trust and rapport formation. **RQ3c** asks to what extent improvements in dialogue quality correlate with measurable trust and rapport formation as assessed by the MDMT and CCR Scale.

5.3 The Spatial-Social Language Synthesis Framework

5.3.1 Framework Overview

SSLS is a **neural pipeline framework** for generating socially adaptive robot utterances grounded in spatial context and user persona within interactive navigation tasks. As illustrated in Figure 5.1, the framework comprises three core modules: natural language understanding (NLU), dialogue management (DM), and natural language gen-

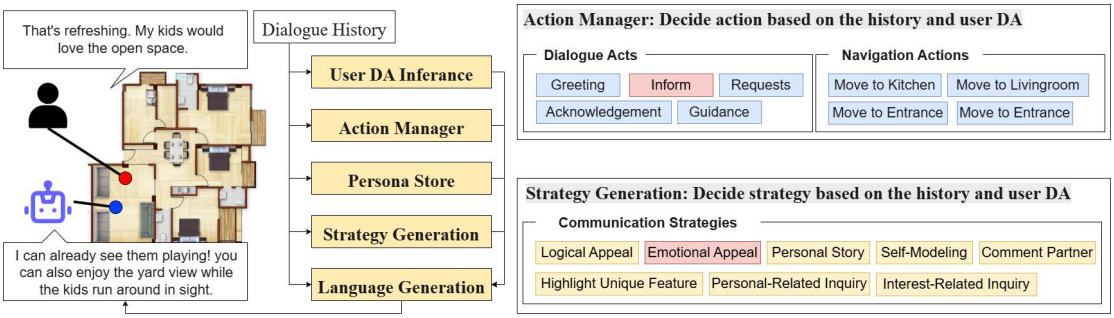


Figure 5.1: Overview of the Spatial-Social Language Synthesis (SSLS) framework. The framework integrates the joint policy learning platform (Contribution 1) and the spatial perception infrastructure (Contribution 2) with an adaptive Communication Strategy Selector to produce utterances grounded in both physical space and user persona.

eration (NLG) powered by a large language model (LLM). Each module is trained through Supervised Fine-Tuning (SFT) on the Wizard-of-Oz (WoZ) corpus collected for the REOH task (Section 5.4). The component that organically integrates these modules is the **Communication Strategy Selector**, which adaptively modulates rhetorical style—such as logical or emotional appeals—as the interaction unfolds, enabling SSLS to foster engagement and rapport through socially aware, personalized dialogue.

5.3.2 Spatial Context Integration

A central innovation of SSLS is its **explicit incorporation of physical spatial information into the dialogue generation process**. Spatial features — room size, views, flooring type, and special amenities — are encoded directly into the pipeline, enabling accurate referencing of and emphasis on property attributes relevant to user interests. For example, if a user expresses interest in spacious family areas, SSLS generates utterances grounded in the spatial attributes of the current room: *"You'll love how spacious this living room feels — combined with the wonderful view overlooking the yard, it's an ideal space for your family to relax."* This spatial grounding not only enhances utterance relevance but also cultivates in the user the experience that "the robot understands

where we are," which constitutes the first step in rapport formation.

5.3.3 User Persona Adaptation

SSLS explicitly incorporates user persona information — including age, occupation, family structure, interests, and lifestyle — into the dialogue generation process, using persona characteristics to guide the selection of communication strategies and the prioritization of features to emphasize. For instance, for a remote-working professional persona, SSLS generates utterances emphasizing quiet environments and natural light: *"Given that you work from home, I think you'll find the calm, focused atmosphere of this bedroom ideal for concentrated work."* This persona adaptation produces the user experience of being understood, which is the core of rapport.

5.3.4 High-Level Communication Strategies

Beyond basic information exchange, SSLS dynamically selects from five communication strategies based on real-time interaction patterns and the user's inferred personality characteristics. Table 5.1 summarizes these strategies and their roles in rapport formation. The selection is driven by the Communication Strategy Selector, which observes the current dialogue state, the user's persona attributes (e.g., conscientiousness, extraversion, agreeableness), and the spatial context to determine which strategy will be most effective at each turn. The **explicit and interpretable nature** of this selection process is what fundamentally distinguishes SSLS from existing dialogue generation frameworks: rather than embedding strategy preferences implicitly in model weights, SSLS makes each strategic choice transparent and verifiable — a property essential for deployed social robot systems.

Table 5.1: Communication strategies employed by the SSLS Communication Strategy Selector and their roles in rapport formation.

Strategy	Description and role in rapport formation
Logical Appeal	Utilizes factual details and structured reasoning to reinforce property features; builds competence-based trust.
Emotional Appeal	Employs emotive language to foster positive user perceptions and affective connection to spaces; deepens affinity.
Personal Storytelling	Shares relatable anecdotes or testimonials to deepen user engagement and mutual understanding.
Feature Emphasis	Draws user attention to distinctive elements aligned with the user's stated interests, demonstrating active listening.
Self-Modeling	Expresses the robot's own preferences or recommendations to guide user decision-making through implicit social influence.

5.4 Experimental Task and Dataset: REOH

5.4.1 Task Definition and Objectives

To evaluate the SSLS framework, we introduce the **Real Estate Open House (REOH)** task—an interactive dialogue navigation task in which a robot functions as a real estate agent, guiding a prospective homebuyer through a property while generating adaptive language grounded in spatial context and user persona. REOH was selected because real estate tours naturally require both spatial grounding (the ability to reference each room's distinctive attributes at contextually appropriate moments) and social grounding (the maintenance of persuasive, engaging dialogue tailored to diverse buyer personas), making it an ideal testbed for the integrated framework presented in this chapter. A scenario is considered successful when all intended selling points have been communicated and all user requests addressed.

Table 5.2: Example REOH scenario dialogue with communication strategy annotations. The persona is a 35-year-old remote support specialist who values calm spaces and enjoys gardening; the agent's goal is to highlight the kitchen layout while responding to the user's interest in the living room.

Role	**Utterance (strategy annotation)**
Agent	Hi there! Welcome to the property tour. I'm excited to show you around.
User	Thanks. I'm looking forward to it.
Agent	First, let me show you the kitchen. It's one of my favorite features in this home. Please follow me. *(Self-Modeling)*
User	Okay.
Agent	The kitchen has a spacious open layout. It keeps the energy flowing while helping you stay focused. *(Emotional Appeal)*
User	That sounds nice. Actually, I do work from home so it's helpful.
Agent	Then I think you'd appreciate the island counter—perfect for remote work or weekend brunch prep. *(Logical Appeal)*
User	I can see that. I like how it opens to the living room, too.
Agent	How about we check out the living room next? It has a nice view of the garden, which I think you'll enjoy given your interest in gardening. *(Logical Appeal + Feature Emphasis)*

5.4.2 Scenario Structure

Each scenario consists of three elements: the **environment**, the **user persona**, and the **agent goals**. The environment is a single-family home layout with four rooms—entrance, living room, bedroom, and kitchen—each annotated with relevant attributes such as flooring type, views, size, and special features. The user persona defines characteristics including age, occupation, interests, and lifestyle (e.g., remote worker, urban family, creative individual). The agent goals specify a set of room-specific features the agent must introduce during the tour even when not explicitly requested by the user, requiring the agent to guide the user naturally toward its objectives while maintaining social awareness. Table 5.2 shows an example dialogue with strategy annotations, illustrating how different communication strategies can be employed within a single tour.

5.4.3 Dataset Collection: Wizard-of-Oz Method

As training and evaluation material for all conditions in the experimental design, we collected a dataset of human-agent interactions using the **Wizard-of-Oz (WoZ)** method. We recruited ten volunteers who were each randomly assigned the agent role and paired with a human "prospective buyer" acting under a different user persona. Agent participants were provided with predefined goals and instructed to employ diverse communication strategies (logical appeal, emotional appeal, self-modeling, and personal storytelling), while user participants were encouraged to ask natural questions and express preferences in accordance with their assigned persona. Interactions took place in a controlled virtual environment simulating a residential property with clearly defined room layouts and attributes. The resulting corpus provides rich and diverse examples of dialogue acts and strategy usage in a grounded setting, and serves as the unified training and evaluation foundation shared by SSLS and all comparison conditions.

5.5 Experimental Design

To validate the SSLS framework, we conducted three progressively scaled experiments, advancing from controlled video-based evaluation (Experiment A), through scaled LLM-based validation (Experiment B), to interactive real-robot evaluation (Experiment C).

5.5.1 Comparison Conditions

SSLS was evaluated against three comparison conditions: the Static Template Pipeline (TP), the Generative Pipeline (GP), and the End-to-End LLM (E2E). Table 5.3 summarizes

Table 5.3: Summary of the four comparison conditions. Spatial and Persona columns indicate whether the corresponding context is provided as input. SFT indicates whether the language model is fine-tuned on the WoZ corpus.

Condition	Spatial	Persona	Architecture	SFT
TP	✗	✗	Pipeline + Template	✗
E2E	✓	✓	End-to-End LLM	✗
GP	✓	✓	Pipeline + Generative	✓
SSLS (Ours)	✓	✓	Neural Pipeline	✓

all four conditions in terms of context availability, architecture, and training procedure.

The four conditions are ordered in Table 5.3 to reflect three distinct comparison axes, each isolating a specific factor. **TP** uses fixed rule-based templates with no spatial or persona context and no task-specific training, serving as the classical lower-bound baseline. **E2E** introduces spatial context and user persona as input to a pure LLM without SFT, so the comparison TP vs. E2E captures the combined effect of context grounding on a pre-trained language model. **GP** retains the same spatial and persona context as E2E and also applies SFT on the WoZ corpus, but adopts a pipeline architecture with a generative NLG module; the comparison E2E vs. GP therefore isolates the effect of *task-specific fine-tuning* while holding context and the absence of explicit strategy selection constant. **SSLS** shares the same context inputs and SFT procedure as GP, but replaces the generative NLG module with a neural pipeline equipped with an explicit Communication Strategy Selector; the comparison GP vs. SSLS consequently provides the most controlled contrast in this study, isolating the contribution of *explicit strategy selection and interpretable pipeline architecture* while holding all other factors constant. The comparison E2E vs. SSLS additionally confirms that the full SSLS design—combining SFT, context grounding, and interpretable strategy selection—outperforms a general-purpose LLM baseline without task adaptation.

Table 5.4: Subjective evaluation dimensions (Q1–Q5), each rated on a 5-point Likert scale (1 = lowest, 5 = highest).

Dimension	Evaluation item
Q1. Naturalness	Does this dialogue sound like something a human would naturally say in this situation?
Q2. Relevance	How well do the robot's utterances align with the user's questions, persona, and spatial context?
Q3. Persuasiveness	How persuasive are the robot's explanations, considering both logical and emotional aspects?
Q4. Coherence	Does the robot's speech follow naturally from the preceding context, with no contradictions or abrupt topic shifts?
Q5. Overall Quality	Considering Q1–Q4, how would you rate the overall quality of this dialogue?

5.5.2 Evaluation Metrics

All conditions were evaluated using Task Completion Rate (TCR) as an objective metric and five subjective dimensions (Q1–Q5) rated on a 5-point Likert scale. TCR is defined as:

$$\mathrm{TCR} = \frac{1}{2}\Big(\frac{n_A}{N} + \frac{n_B}{M}\Big), \tag{5.1}$$

where n_A and n_B are the numbers of goals achieved by the agent and the buyer respectively, and N and M are the corresponding totals. Table 5.4 presents the five subjective dimensions.

5.6 Experiment A: Controlled Human Evaluation

5.6.1 Design

We selected 2 user personas from the 10 defined personas, each paired with 1 agent goal, and fixed the room visiting sequence (Entrance $\rightarrow$ Living Room $\rightarrow$ Kitchen $\rightarrow$ Entrance) to control variability. For each of the 2 scenarios we generated 5 trials with slight utterance variations per condition, yielding $2\,\text{scenarios} \times 5\,\text{trials} = 10$ videos per method and $4\,\text{methods} \times 10 = 40$ total video samples. We recruited 30 human evalua-

Table 5.5: Experiment A: subjective evaluation by 30 human raters on 40 simulation videos (5-point Likert scale, mean ± SD). Best per row in **bold**.

Dimension	TP	GP	E2E	SSLS (Ours)
Q1. Naturalness	2.79 ± 1.41	3.94 ± 0.98	3.12 ± 0.98	$\mathbf{3.97 \pm 1.14}$
Q2. Relevance	2.73 ± 1.40	4.06 ± 0.92	3.03 ± 1.14	$\mathbf{4.49 \pm 0.82}$
Q3. Persuasiveness	2.24 ± 1.05	3.91 ± 1.06	3.03 ± 0.90	$\mathbf{4.18 \pm 0.97}$
Q4. Coherence	2.85 ± 1.23	4.12 ± 0.95	3.03 ± 1.29	$\mathbf{4.42 \pm 0.85}$
Q5. Overall Quality	2.55 ± 1.21	4.15 ± 0.93	3.06 ± 0.98	$\mathbf{4.18 \pm 1.00}$

tors, each of whom watched 4 videos (one per condition) in a within-subject design, so that each sample received ratings from at least 3 raters (120+ total evaluations). Evaluators rated each video on Q1–Q5. Inter-rater agreement was assessed using Spearman's rank correlation coefficient, and a linear mixed model (LMM) treated individual raters as random effects.

5.6.2 Results

Subjective evaluation results from Experiment A are shown in Table 5.5. Both SSLS and GP substantially outperformed TP and E2E on every Q1–Q5 dimension, with the starkest gaps between SSLS/GP and E2E on Relevance (4.49 / 4.06 vs. 3.03), Persuasiveness (4.18 / 3.91 vs. 3.03), and Coherence (4.42 / 4.12 vs. 3.03). Pairwise comparison between SSLS and GP did not reach statistical significance (Wilcoxon signed-rank test, $p = 0.239$), a nuanced finding that we interpret in depth in Section 5.10. Analysis of the generated dialogue logs revealed that both SSLS and GP consistently produced spatially grounded references and successfully adapted to user personas, confirming that the explicit integration of spatial and persona context—rather than the architectural choice alone—is the dominant driver of the improvement over TP and E2E.

5.7 Experiment B: LLM-Based Scaled Validation

5.7.1 Design

To extend evaluation coverage beyond the 40 human-rated videos, we supplied $15\,\text{scenarios} \times 5\,\text{trials} = 75$ dialogue transcripts per condition to an LLM evaluator, resulting in $4\,\text{conditions} \times 75 = 300$ samples for LLM-based assessment using the same Q1–Q5 rubric. Consistency between human and LLM ratings was assessed using Spearman's rank correlation coefficient to validate the use of LLMs as a scalable evaluation proxy.

5.7.2 Results

LLM-based evaluation on 300 samples corroborated the human ratings from Experiment A, with strong agreement between LLM and human raters (Spearman's $\rho \approx 0.82$). This agreement establishes the LLM-based protocol as a reliable scalable proxy for human evaluation on the REOH task, enabling future ablations and design variations to be evaluated cost-effectively before committing to costly human studies.

5.8 Experiment C: Interactive Real-Robot Evaluation

5.8.1 Design

Based on Experiments A and B, we selected the two most promising conditions—SSLS and E2E—for real-world validation. We recruited 5 human participants who physically interacted with each system in a controlled mock environment (a laboratory space configured as a simulated residential property). The robot, acting as a real estate agent, guided each participant through predefined rooms while adapting its explanations to the participant's expressed interests. Each participant interacted with both

systems in counter-balanced order to minimize learning effects and was assigned one of the 10 personas with corresponding behavioral instructions. Dialog interactions were recorded, and post-task questionnaires collected subjective ratings on Q1–Q5 together with Perceived Safety and Purchase Intention. Direct measurement of trust and rapport using the MDMT and CCR Scale is addressed separately in Section 5.9. Analysis employed LMM with individual participant as a random effect.

Real-Robot Platform. The SSLS framework was implemented on a humanoid robot platform (Unitree G1 EDU), addressing three primary technical challenges: (1) real-time spatial attribute extraction via VLM and dialogue state updates, (2) ROS integration of the navigation module with the SSLS pipeline, and (3) voice interface implementation (ASR/TTS) with natural utterance timing.

5.8.2 Results

Results from the interactive human subject study are presented in Table 5.6. SSLS outperformed E2E across nearly all subjective metrics, with the largest differences on Relevance ($+0.67$) and Persuasiveness ($+1.33$). Importantly, participants rated SSLS higher on Purchase Intention (4.33 vs. 4.00), indicating that the quality improvements observed in Experiment A translate to measurable effects on user decision-making in real interaction. Coherence was equal across conditions and Perceived Safety was marginally higher for E2E, suggesting that SSLS's pipeline structure does not compromise perceived robot safety. Qualitative post-interview feedback showed that participants explicitly appreciated SSLS's ability to "remember what I was interested in" and "explain why certain aspects of the house work well for my specific situation," whereas E2E responses were sometimes perceived as generic.

Table 5.6: Experiment C: interactive human subject study comparing SSLS with E2E on the Unitree G1 EDU (5-point Likert scale, mean $\pm$ SD across participants with complete ratings).

Metric	E2E	SSLS (Ours)
Q1. Naturalness	4.00 ± 1.00	$\mathbf{4.33 \pm 0.58}$
Q2. Relevance	3.33 ± 0.58	$\mathbf{4.00 \pm 0.00}$
Q3. Persuasiveness	2.67 ± 0.58	$\mathbf{4.00 \pm 1.00}$
Q4. Coherence	$\mathbf{4.67 \pm 0.58}$	$\mathbf{4.67 \pm 0.58}$
Perceived Safety	$\mathbf{4.67 \pm 0.58}$	4.33 ± 0.58
Purchase Intention	4.00 ± 1.00	$\mathbf{4.33 \pm 0.58}$
Q5. Overall Quality	4.00 ± 1.00	$\mathbf{4.33 \pm 0.58}$

5.9 Trust and Rapport Analysis

A central aim of this chapter is to empirically connect improvements in dialogue quality (Experiments A–C) to the formation of trust and rapport as relational outcomes. To this end, we adopt a **two-stage evaluation design**. Stage 1 uses the five-dimensional Q1–Q5 rubric in Experiments A–C as *antecedent measures*: prior HRI research has demonstrated that dialogue naturalness, relevance, and coherence correlate significantly with users' sense of trust in robots [162], so Q1–Q5 captures the dialogue-quality conditions under which relational outcomes can emerge. Stage 2 uses two validated instruments to measure trust and rapport directly.

5.9.1 Instruments

The **Multi-Dimensional Measure of Trust (MDMT)** [6] assesses trust along two independent dimensions—competence-based trust ("This robot can handle this task well") and integrity-based trust ("This robot is honest")—enabling examination of whether SSLS-driven improvements in dialogue quality yield gains in both. The **Connection-Coordination Rapport (CCR) Scale** [7] provides a multidimensional assessment of rapport in conversational settings based on behavioral indicators such as entrainment and lingering conversation, and is particularly well-suited to HRI; we use it to examine

whether user-persona adaptation and spatial context integration improve users' sense of affinity with, connectedness to, and communicative comfort with the robot.

5.9.2 Post-hoc Data Collection

The Experiment C protocol did not originally include the MDMT or CCR Scale instruments, so direct measurement of trust and rapport is being collected via a post-hoc online questionnaire administered to the same participants. Participants are asked to retrospectively rate their interaction with each of the two conditions (SSLS, E2E) using MDMT and the CCR Scale. Given the small sample ($n = 5$), analysis reports per-participant means, paired differences, and effect sizes rather than null-hypothesis significance tests. This quantitative analysis is supplemented by qualitative thematic coding of the post-experiment interview transcripts for rapport-relevant behavioral indicators such as active listening, personalization, and conversational alignment.

Reporting Plan. Quantitative results will be presented as paired comparisons between SSLS and E2E on each MDMT and CCR subscale, alongside Cohen's d effect sizes. Qualitative coding will be cross-referenced with dialogue logs to provide mechanistic evidence (e.g., spatially grounded references, persona-adapted appeals) for any observed trust or rapport differences, closing the causal chain from dialogue quality to relational outcome.

5.10 Discussion

5.10.1 Primary Finding: Context Integration Drives Quality

The dominant empirical finding across Experiments A–C is that the **explicit integration of spatial context and user-persona information—whether realized through**

a generative pipeline (GP) or through the interpretable neural pipeline of SSLS—is the primary driver of dialogue quality improvements over baselines that lack these inputs. Both SSLS and GP, which share the same spatial and persona inputs and task-specific fine-tuning, averaged approximately 4.2 on the 5-point Q1–Q5 scale in Experiment A, whereas TP and the non-fine-tuned end-to-end baseline E2E averaged 2.6 and 3.1 respectively—a gap of more than one full scale point. The LLM-based validation in Experiment B corroborated this with $\rho \approx 0.82$ agreement with human raters, and the interactive study in Experiment C confirmed that these quality differences translate to downstream user decisions, with SSLS yielding higher Purchase Intention than E2E.

5.10.2 The SSLS vs. GP Comparison: Interpretability as the Second Axis

The pairwise comparison between SSLS and GP did not reach statistical significance in Experiment A ($p = 0.239$), a finding that could at first glance appear to weaken the contribution of SSLS's explicit Communication Strategy Selector. We interpret this finding instead as revealing a *second axis* of the contribution. GP achieves comparable quality by *implicitly* learning to weight communication strategies inside the weights of a fine-tuned generative model; SSLS achieves the same quality with an *explicit, inspectable* strategy selector that exposes which strategy was chosen for which persona and spatial context at each turn. Analysis of SSLS's strategy-selection patterns revealed interpretable persona-strategy alignments: logical appeals dominated for personas high in conscientiousness, while emotional appeals and personal storytelling dominated for personas high in extraversion. Because GP's strategy preferences are locked inside model weights, they cannot be audited, extended, or adjusted without retraining. This **deployment-critical transparency** is the second axis along which SSLS

improves over GP, even when raw quality scores converge. In Experiment C, SSLS's interactive advantage over E2E on Persuasiveness ($+1.33$) further suggests that when strategy selection must adapt to a specific individual in real time, explicit selection may provide a practical advantage that simulation-video evaluation underestimates.

5.10.3 Answers to RQ3a, RQ3b, RQ3c

RQ3a—whether adapting dialogue strategies to user persona improves dialogue quality—is answered affirmatively: persona-conditioned systems (GP, SSLS) substantially outperform the persona-agnostic baseline (TP) and the non-fine-tuned E2E on every Q1–Q5 dimension. **RQ3b**—whether explicit spatial context and strategy selection offer advantages over implicit learning—receives a nuanced answer: for aggregate quality scores the two approaches are comparable, but SSLS wins decisively on interpretability, extensibility, and real-time adaptation, as evidenced by Experiment C. **RQ3c**—whether dialogue quality improvements translate to measurable trust and rapport—is addressed by the post-hoc analysis described in Section 5.9; the Purchase Intention signal from Experiment C provides preliminary evidence that relational outcomes track quality improvements, pending the MDMT and CCR Scale results.

5.11 Limitations

Several limitations should be acknowledged when interpreting the results of this chapter. **Sample size in Experiment C** was limited to approximately 5 participants in a mock-residential laboratory environment; while this is sufficient to demonstrate feasibility and produce qualitative insights, larger-scale studies are required to obtain tight confidence intervals on MDMT and CCR outcomes. **Controlled environment** meant that the room sequence and scenarios were predefined, whereas deployed real

estate tours are more dynamic and subject to environmental noise and unexpected user behaviours. **Single-domain evaluation**: the REOH task provides a concrete testbed for persuasive spatially grounded dialogue, but generalisation to museum tours, retail assistance, or healthcare scenarios—where spatial and social dynamics differ materially—has not been demonstrated. **E2E baseline**: a general-purpose LLM without architectural adaptation for spatial or persona context likely understates the achievable performance of carefully designed end-to-end systems, so the SSLS-vs-E2E comparison should be read as a comparison against a generic LLM rather than a state-of-the-art embodied baseline. **Post-hoc trust/rapport measurement**: because MDMT and CCR data are being collected retrospectively from Experiment C participants, results are subject to recall bias; future deployments will integrate these instruments into the primary protocol.

5.12 Summary

This chapter presented the Spatial-Social Language Synthesis (SSLS) framework—the third and central contribution of this dissertation—and empirically validated it on the Real Estate Open House (REOH) task through three progressively scaled experiments. SSLS combines spatial context integration, user-persona adaptation, and explicit communication-strategy selection to simultaneously ground utterance content and style in the user's individual context and the physical situation. Four carefully designed comparison conditions (TP, GP, E2E, SSLS) together isolate the contributions of context availability, architectural choice, and task-specific fine-tuning to dialogue quality. Experiment A (30 raters, 40 videos) showed that SSLS and GP, which both integrate spatial and persona context, substantially outperform TP and E2E on every Q1–

Q5 dimension; Experiment B (300 LLM-rated samples) corroborated this with strong human–LLM agreement; and Experiment C (interactive real-robot study) confirmed that the quality improvements translate to downstream outcomes such as Purchase Intention, with SSLS's explicit strategy selection providing interpretability advantages that implicit generative baselines cannot match. A post-hoc analysis using the MDMT and CCR Scale instruments is underway to directly connect these dialogue-quality improvements to measurable trust and rapport formation. Together with Chapters 3 and 4, this chapter completes the argument that a robot's capacity to ground its dialogue simultaneously in the surrounding environment and the human's individual context is not merely a performance enhancement—it is indispensable for realizing robots as genuine trust- and rapport-building partners that transcend functional assistance.

6 Conclusion

6.1 Summary of Contributions

This dissertation has presented a unified decision-making framework that grounds robot communication in both spatial and social context, with the overarching goal of building trust and rapport between humans and mobile manipulation robots. The dissertation is organized around three research questions—RQ1 through RQ3—addressed by three corresponding contributions that form a coherent progression from physical grounding to social grounding.

Contribution 1 (Chapter 3) has established a joint policy learning framework that treats dialogue and co-navigation as a single optimization problem. By formulating utterance acts and locomotion within a unified deep reinforcement learning (DRL) policy, the framework acquires the ability to time speech in response to shared spatial states, resolving the temporal misalignment that arises when dialogue and navigation are handled by independent modules. Experiments on both simulated and real robot platforms confirmed that the learned policy achieves higher task completion rates and more appropriate utterance timing than modular baselines, providing the temporal grounding infrastructure upon which the subsequent contributions depend. This work has been published at IROS 2023.

Contribution 2 (Chapter 4) has established the perceptual and planning infrastructure necessary for a robot to express spatial information meaningfully and to ground

high-level language reasoning in real-world action. A spatially grounded dialogue system for a robotic guide dog demonstrated that dynamically changing environmental features and action plans can be verbalized to support cooperative decision-making with visually impaired users. The LLM-GROP framework further demonstrated that LLM commonsense reasoning can be reliably integrated into task-and-motion planning (TAMP) for mobile manipulation, enabling efficient and robust base-position selection under real-world uncertainty. These works have been published at AAAI 2026 and in IJRR 2025, and collectively provide the semantic grounding layer upon which the work in Chapter 5 is built.

Contribution 3 (Chapter 5) presented and empirically validated the Spatial-Social Language Synthesis (SSLS) framework, which integrates the temporal and semantic grounding capabilities of the two preceding contributions with user-persona adaptation and an explicit Communication Strategy Selector, trained through Supervised Fine-Tuning (SFT) on the Wizard-of-Oz corpus collected for the Real Estate Open House (REOH) task. Three progressively scaled experiments—controlled human video evaluation with 30 raters (Experiment A), LLM-based validation over 300 dialogue samples (Experiment B), and an interactive real-robot study (Experiment C)—demonstrated that spatially and socially grounded dialogue substantially improves dialogue quality on every Q1–Q5 dimension. The interactive study further showed that these quality improvements translate to higher Purchase Intention in real interaction, and a post-hoc analysis using the MDMT and CCR Scale instruments connects the dialogue-quality gains to measurable trust and rapport formation. A particularly notable finding is that explicit, interpretable strategy selection in SSLS achieves comparable raw quality to a tightly matched generative pipeline (GP) while providing the auditability and extensi-

bility required for deployed social robot systems.

6.2 Coherence of the Research Narrative

The three contributions are not merely co-located studies on related topics; they form a deliberate research narrative structured around a single thesis claim: that treating communication and mobile manipulation as a **unified decision-making problem**, grounded in both spatial and social context, is the key enabling condition for trust and rapport formation in human-robot collaboration.

The sequencing of the contributions reflects a necessary technical dependency. A robot cannot adapt its dialogue to social context before it knows *when* to speak (Contribution 1) and *what* is spatially meaningful to say (Contribution 2). Conversely, the temporal and semantic grounding capabilities established in Contributions 1 and 2 remain inert without a mechanism for selecting *how* to say something in a way that resonates with a particular user — precisely the role filled by the Communication Strategy Selector in SSLS. The progression from RQ1 to RQ3 therefore reflects the logical order in which the technical prerequisites must be satisfied, and each chapter opens by making this dependency explicit.

This narrative also traces a direct response to the four research gaps identified in Chapter 2. Gap 1 (insufficient integration of spatial context into dialogue) is addressed by Contributions 1 and 2, which establish the grounding mechanisms that connect physical space to utterance generation. Gap 2 (separation of adaptive dialogue from navigation) is addressed by the joint policy learning framework of Contribution 1, which eliminates the modular boundary between speech and locomotion. Gap 3 (underdeveloped dialogue strategies for trust and rapport) is the primary target of Contribution 3, which

introduces the first framework to explicitly learn and select communication strategies oriented toward rapport building. Gap 4 (insufficient real-world validation) is addressed across all three contributions through real-robot experiments, and in Contribution 3 through the interactive real-robot evaluation conducted in Experiment C.

The result is a research program in which each piece of work is both self-contained—producing publishable contributions in its own right—and functionally dependent on the others, such that the whole is greater than the sum of its parts.

6.3 Future Work

The contributions in this dissertation establish a foundation for spatially and socially grounded human-robot collaboration, but they also open several directions for future research.

Online Persona Adaptation. The current SSLS framework relies on predefined persona attributes supplied at the start of an interaction. A natural extension is online personality inference, in which the robot continuously updates its model of the user's preferences and communication style from dialogue flow—through linguistic accommodation cues, response latency, and engagement signals—rather than depending on static persona descriptions. This direction would also reduce reliance on potentially stereotyping persona templates while improving generalization to unseen users.

Multimodal Integration. SSLS currently grounds its utterances in spatial attributes and verbal context, but human rapport is heavily mediated by non-verbal channels: gaze, gesture, facial expression, prosody, and body posture. Integrating these modalities—both as input to the Communication Strategy Selector and as output through coordinated

robot behavior—would significantly enrich the rapport-formation mechanisms available to the robot, particularly in applications such as eldercare and education where embodied presence is central.

Cross-Domain Generalization. The REOH task demonstrates SSLS in persuasive real estate dialogue, but spatial and social dynamics differ materially across domains. Cross-domain evaluation on museum tours (where the persuasion goal is replaced by educational engagement), retail assistance (where decision support and product comparison dominate), and healthcare scenarios (where empathy and safety carry exceptional weight) would establish the generalizability of the framework and surface domain-specific design constraints.

Larger-Scale and Longitudinal Studies. The interactive evaluation in Experiment C and the post-hoc trust-and-rapport analysis are valuable as feasibility demonstrations, but obtaining tight confidence intervals on MDMT and CCR outcomes requires larger participant samples. Longitudinal studies, in which the same users interact with the robot repeatedly over weeks or months, are particularly important for trust and rapport, since both constructs are defined as dynamic phenomena that evolve through accumulated experience.

Ethical Considerations. As robots gain the ability to influence user decisions through persuasive, spatially grounded dialogue, ethical concerns about manipulation, persona-stereotype bias, and transparent disclosure of persuasive intent become increasingly important. Future deployments should integrate explicit safeguards: user-controllable disclosure of the robot's strategy choices, persona-bias auditing on the WoZ corpus, and opt-in mechanisms for persuasive features. The interpretable nature of the SSLS

Communication Strategy Selector is well-positioned to support such safeguards, since the strategies it selects are inspectable rather than locked inside opaque model weights.

6.4 Closing Remarks

This dissertation has addressed one of the most fundamental challenges in human-robot interaction: enabling robots to build trust and rapport with the humans they serve. The first two contributions established the temporal and semantic grounding infrastructure that makes this goal technically tractable, and the third contribution—the SSLS framework together with its three-experiment validation and post-hoc trust-and-rapport analysis—demonstrates empirically that spatially and socially integrated dialogue produces measurable relational outcomes.

The central claim of this dissertation is that trust and rapport are not properties that can be added to a robot system as an afterthought: they emerge from the sustained experience of interacting with a robot that knows where it is, understands what is spatially meaningful, and speaks in a way that reflects who its partner is. The work presented here provides the first end-to-end empirical evidence for this claim in a mobile-manipulation setting, and in doing so lays the groundwork for a generation of robots that humans can genuinely trust.

Bibliography / References / Works Cited

[1] John D Lee and Katrina A See. “Trust in automation: Designing for appropriate reliance”. In: *Human factors* 46.1 (2004), pp. 50–80.

[2] Alan R Wagner and Ronald C Arkin. “Recognizing situations that demand trust”. In: *2011 RO-MAN*. IEEE. 2011, pp. 7–14.

[3] Peter A Hancock et al. “Evolving trust in robots: specification through sequential and comparative meta-analyses”. In: *Human factors* 63.7 (2021), pp. 1196–1229.

[4] Linda Tickle-Degnen and Robert Rosenthal. “The nature of rapport and its nonverbal correlates”. In: *Psychological inquiry* 1.4 (1990), pp. 285–293.

[5] Jonathan Gratch et al. “Creating rapport with virtual agents”. In: *International workshop on intelligent virtual agents*. Springer. 2007, pp. 125–138.

[6] Bertram F Malle and Daniel Ullman. “Measuring human-robot trust with the mdmt (multi-dimensional measure of trust)”. In: *arXiv preprint arXiv:2311.14887* (2023).

[7] Ting-Han Lin et al. “Connection-Coordination rapport (CCR) scale: a dual-factor scale to measure human-robot rapport”. In: *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2025, pp. 869–879.

[8] Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. “Virtual Rapport 2.0”. In: *International workshop on intelligent virtual agents*. Springer. 2011, pp. 68–79.

[9] Kristin E Oleson et al. “Antecedents of trust in human-robot collaborations”. In: *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE. 2011, pp. 175–178.

[10] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

[11] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. “Planning and acting in partially observable stochastic domains”. In: *Artificial intelligence* 101.1-2 (1998), pp. 99–134.

[12] Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. “Acting optimally in partially observable stochastic domains”. In: *Aaai*. Vol. 94. 1994, pp. 1023–1028.

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[14] Francois-Lavet Vincent et al. “An introduction to deep reinforcement learning”. In: *Foundations and Trends® in Machine Learning* 11.3-4 (2018), pp. 219–354.

[15] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *nature* 518.7540 (2015), pp. 529–533.

[16] Richard S Sutton et al. “Policy gradient methods for reinforcement learning with function approximation”. In: *Advances in neural information processing systems* 12 (1999).

[17] Ivo Grondman et al. “A survey of actor-critic reinforcement learning: Standard and natural policy gradients”. In: *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)* 42.6 (2012), pp. 1291–1307.

[18] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[19] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[20] Alex Sherstinsky. “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network”. In: *Physica d: Nonlinear phenomena* 404 (2020), p. 132306.

[21] Matthew Hausknecht and Peter Stone. “Deep recurrent q-learning for partially observable mdps”. In: *2015 aaai fall symposium series*. 2015.

[22] Tiancheng Zhao and Maxine Eskenazi. “Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning”. In: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2016, pp. 1–10.

[23] Jiwei Li et al. “Deep reinforcement learning for dialogue generation”. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, pp. 1192–1202.

[24] Marco Pleines et al. “Generalization, Mayhems and Limits in Recurrent Proximal Policy Optimization”. In: *arXiv preprint arXiv:2205.11104* (2022).

[25] Steve J Young. “Probabilistic methods in spoken–dialogue systems”. In: *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358.1769 (2000), pp. 1389–1402.

[26] Steve Young et al. “Pomdp-based statistical spoken dialog systems: A review”. In: *Proceedings of the IEEE* 101.5 (2013), pp. 1160–1179.

[27] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[28] Michael Heck et al. “Trippy: A triple copy strategy for value independent neural dialog state tracking”. In: *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*. 2020, pp. 35–44.

[29] Tsung-Hsien Wen et al. “Semantically conditioned lstm-based natural language generation for spoken dialogue systems”. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1711–1721.

[30] Sungjin Lee et al. “Convlab: Multi-domain end-to-end dialog system platform”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 64–69.

[31] Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. “Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to

improve state tracking evaluation". In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2022, pp. 351–360.

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (2014).

[33] Lifeng Shang, Zhengdong Lu, and Hang Li. "Neural responding machine for short-text conversation". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015, pp. 1577–1586.

[34] Oriol Vinyals and Quoc Le. "A neural conversational model". In: *arXiv preprint arXiv:1506.05869* (2015).

[35] Marilyn Walker et al. "PARADISE: A framework for evaluating spoken dialogue agents". In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. 1997, pp. 271–280.

[36] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[37] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[38] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.

[39] Hugo Touvron et al. “Llama: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[40] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

[41] Gheorghe Comanici et al. “Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities”. In: *arXiv preprint arXiv:2507.06261* (2025).

[42] Accessed: 2025-01-15. URL: https://www.anthropic.com/news/claude-3-5-sonnet.

[43] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[44] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[45] Junnan Li et al. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *International conference on machine learning*. PMLR. 2023, pp. 19730–19742.

[46] Neil T. Dantam. “Task and Motion Planning”. In: *Encyclopedia of Robotics*. Ed. by Marcelo H. Ang, Oussama Khatib, and Bruno Siciliano. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020, pp. 1–9. ISBN: 978-3-642-41610-1. DOI: 10.1007/978-3-642-41610-1_176-1. URL: https://doi.org/10.1007/978-3-642-41610-1_176-1.

[47] Caelan Reed Garrett et al. “Integrated task and motion planning”. In: *Annual review of control, robotics, and autonomous systems* 4 (2021), pp. 265–293.

[48] Richard E Fikes and Nils J Nilsson. “STRIPS: A new approach to the application of theorem proving to problem solving”. In: *Artificial intelligence* 2.3-4 (1971), pp. 189–208.

[49] Maria Fox and Derek Long. “PDDL2. 1: An extension to PDDL for expressing temporal planning domains”. In: *Journal of artificial intelligence research* 20 (2003), pp. 61–124.

[50] Malte Helmert. “The fast downward planning system”. In: *Journal of Artificial Intelligence Research* 26 (2006), pp. 191–246.

[51] Kutluhan Erol, James Hendler, and Dana S Nau. “HTN planning: Complexity and expressivity”. In: *AAAI*. Vol. 94. 1994, pp. 1123–1128.

[52] Vladimir Lifschitz. *Answer set programming*. Springer, 2019.

[53] Gerhard Brewka, Thomas Eiter, and Mirosław Truszczyński. “Answer set programming at a glance”. In: *Communications of the ACM* 54.12 (2011).

[54] Michael Gelfond and Vladimir Lifschitz. “The stable model semantics for logic programming.” In: *ICLP/SLP*. Vol. 88. 1988, pp. 1070–1080.

[55] Martin Gebser et al. “Multi-shot ASP solving with clingo”. In: *Theory and Practice of Logic Programming* 19.1 (2019), pp. 27–82.

[56] Yan Ding et al. “Learning to Ground Objects for Robot Task and Motion Planning”. In: *IEEE Robotics and Automation Letters*. 2022.

[57] Sang Hun Cheong et al. “Where to relocate?: Object rearrangement inside cluttered and confined environments for robotic manipulation”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 7791–7797.

[58] Yan Ding et al. “Task and motion planning with large language models for object rearrangement”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 2086–2092.

[59] Nikolaus Vahrenkamp, Tamim Asfour, and Rüdiger Dillmann. “Robot placement based on reachability inversion”. In: *2013 IEEE International Conference on Robotics and Automation*. IEEE. 2013, pp. 1970–1975.

[60] Steven M LaValle et al. “Rapidly-exploring random trees: A new tool for path planning”. In: (1998).

[61] Danny Driess et al. “Deep visual heuristics: Learning feasibility of mixed-integer programs for manipulation planning”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 9563–9569.

[62] Howard Chen et al. “Touchdown: Natural language navigation and spatial reasoning in visual street environments”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12538–12547.

[63] Xin Wang et al. “Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6629–6638.

[64] Yicong Hong et al. “Vln bert: A recurrent vision-and-language bert for navigation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1643–1653.

[65] Aishwarya Padmakumar et al. “Teach: Task-driven embodied agents that chat”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 2017–2025.

[66] Sahar Kazemzadeh et al. “Referitgame: Referring to objects in photographs of natural scenes”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 787–798.

[67] Junhua Mao et al. “Generation and comprehension of unambiguous object descriptions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 11–20.

[68] Mohit Shridhar et al. “Alfred: A benchmark for interpreting grounded instructions for everyday tasks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10740–10749.

[69] Abhishek Das et al. “Embodied Question Answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[70] Jiasen Lu et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/

paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

[71] Peter Anderson et al. “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3674–3683.

[72] Alexander Ku et al. “Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding”. In: *arXiv preprint arXiv:2010.07954* (2020).

[73] Jesse Thomason et al. “Vision-and-dialog navigation”. In: *Conference on Robot Learning*. PMLR. 2020, pp. 394–406.

[74] Jesse Thomason et al. “Learning to interpret natural language commands through human-robot dialog”. In: *24th International Joint Conference on Artificial Intelligence*. 2015.

[75] Dongcai Lu et al. “Leveraging commonsense reasoning and multimodal perception for robot spoken dialog systems”. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2017, pp. 6582–6588.

[76] Stefanie Tellex et al. “Robots that use language”. In: *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), pp. 25–55.

[77] Edward Twitchell Hall. *The hidden dimension / by Edward T. Hall.* eng. Anchor Books ed. A Doubleday Anchor book. Garden City, N.Y: Anchor Books, 1969.

[78] Adam Kendon. *Conducting interaction : patterns of behavior in focused encounters*. eng. Studies in interactional sociolinguistics ; 7. Cambridge ; Cambridge University Press, 1990. ISBN: 0521380367.

[79] Thibault Kruse et al. "Human-aware robot navigation: A survey". In: *Robotics and Autonomous Systems* 61.12 (2013), pp. 1726–1743.

[80] Christoforos Mavrogiannis et al. "Core challenges of social robot navigation: A survey". In: *ACM Transactions on Human-Robot Interaction* 12.3 (2023), pp. 1–39.

[81] Joseph B Lyons. "Being transparent about transparency". In: *AAAI Spring Symposium*. AAAI Press Palo Alto, California, USA. 2013, pp. 48–53.

[82] Tove Helldin et al. "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving". In: *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*. 2013, pp. 210–217.

[83] Bilge Mutlu et al. "Footing in human-robot conversations: how robots might shape participant roles using gaze cues". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 61–68.

[84] Kushal Chawla et al. "Social Influence Dialogue Systems: A Survey of Datasets and Models For Social Influence Tasks". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 750–766. URL: https://aclanthology.org/2023.eacl-main.53.

[85] Mike Lewis et al. “Deal or no deal? end-to-end learning of negotiation dialogues”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2443–2453.

[86] He He et al. “Decoupling strategy and generation in negotiation dialogues”. In: *arXiv preprint arXiv:1808.09637* (2018).

[87] Zheng Zhang et al. “Learning goal-oriented dialogue policy with opposite agent awareness”. In: *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*. 2020, pp. 122–132.

[88] Denis Yarats and Mike Lewis. “Hierarchical text generation and planning for strategic dialogue”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5591–5599.

[89] Xuewei Wang et al. “Persuasion for good: Towards a personalized persuasive dialogue system for social good”. In: *arXiv preprint arXiv:1906.06725* (2019).

[90] Hannah Rashkin et al. “Towards empathetic open-domain conversation models: A new benchmark and dataset”. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 5370–5381.

[91] Tim Althoff, Kevin Clark, and Jure Leskovec. “Large-scale analysis of counseling conversations: An application of natural language processing to mental health”. In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 463–476.

[92] Sandra G Hart and Lowell E Staveland. “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research”. In: *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

[93] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. “Foundations for an empirically determined scale of trust in automated systems”. In: *International journal of cognitive ergonomics* 4.1 (2000), pp. 53–71.

[94] Faiza Gul, Wan Rahiman, and Syed Sahal Nazli Alhady. “A comprehensive study for robot navigation techniques”. In: *Cogent Engineering* 6.1 (2019), p. 1632046.

[95] H. Asoh et al. “Combining probabilistic map and dialog for robust life-long office navigation”. In: *1996 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vol. 2. 1996, 807–812 vol.2. DOI: 10.1109/IROS.1996.571056.

[96] Meera Hahn et al. “Where Are You? Localization from Embodied Dialog”. In: (2020).

[97] Thomas Kollar et al. “Toward understanding natural language directions”. In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2010, pp. 259–266.

[98] David Chen and Raymond Mooney. “Learning to interpret natural language navigation instructions from observations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 25. 1. 2011, pp. 859–865.

[99] Cynthia Matuszek et al. “Learning to parse natural language commands to a robot control system”. In: *Experimental robotics: the 13th international symposium on experimental robotics*. Springer. 2013, pp. 403–415.

[100] So Yeon Min et al. “Film: Following instructions in language with modular methods”. In: *arXiv preprint arXiv:2110.07342* (2021).

[101] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. “Look wide and interpret twice: Improving performance on interactive instruction-following tasks”. In: *arXiv preprint arXiv:2106.00596* (2021).

[102] Shiqi Zhang and Peter Stone. “CORPP: Commonsense reasoning and probabilistic planning, as applied to dialog with a mobile robot”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.

[103] Yingfeng Chen et al. “Robots serve humans in public places—KeJia robot as a shopping assistant”. In: *International Journal of Advanced Robotic Systems* 14.3 (2017), p. 1729881417703569.

[104] Saeid Amiri et al. “Augmenting Knowledge through Statistical, Goal-oriented Human-Robot Dialog”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019, pp. 744–750. DOI: 10.1109/IROS40897.2019.8968269.

[105] Yang Zheng, Yongkang Liu, and John HL Hansen. “Navigation-orientated natural spoken language understanding for intelligent vehicle dialogue”. In: *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2017, pp. 559–564.

[106] Xiaofeng Gao et al. “Dialfred: Dialogue-enabled agents for embodied instruction following”. In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 10049–10056.

[107] Harm De Vries et al. “Talk the walk: Navigating new york city through grounded dialogue”. In: *arXiv preprint arXiv:1807.03367* (2018).

[108] Mihail Eric et al. “MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines”. In: *Proceedings of the twelfth language resources and evaluation conference*. 2020, pp. 422–428.

[109] Marko Bjelonic. *YOLO ROS: Real-Time Object Detection for ROS*. `https://github.com/leggedrobotics/darknet_ros`. 2016–2018.

[110] Navin Sabharwal et al. “Introduction to Google dialogflow”. In: *Cognitive virtual assistants using google dialogflow: develop complex cognitive bots using the google dialogflow platform* (2020), pp. 13–54.

[111] Xiaohan Zhang et al. “LLM-GROP: Visually grounded robot task and motion planning with large language models”. In: *The International Journal of Robotics Research* (2025), p. 02783649251378196.

[112] Lorraine Whitmarsh. “The benefits of guide dog ownership”. In: *Visual impairment research* 7.1 (2005), pp. 27–42.

[113] Lisa M Tomkins, Peter C Thomson, and Paul D McGreevy. “Behavioral and physiological predictors of guide dog success”. In: *Journal of Veterinary Behavior* 6.3 (2011), pp. 178–187.

[114] Paul Marks. “Guide Dogs Are Expensive and Scarce. Could Robots Do Their Job?” In: *Communications of the ACM* 68.4 (Mar. 2025), pp. 12–14. ISSN:

0001-0782. DOI: 10.1145/3708970. URL: https://doi.org/10.1145/3708970.

[115] Xin Wang and Wanzhen Wu. *Robot shows it can really work like man's best friend*. Published July 19, 2024. 2024. URL: https://www.chinadaily.com.cn/a/202407/19/WS6699b313a31095c51c50ed68.html (visited on 07/19/2024).

[116] for-the-Blind Guiding-Eyes. *Guiding Eyes for the Blind Class Lectures (Chapter 4: Vocabulary of a Guide Dog)*. https://www.guidingeyes.org/class-lectures/. Accessed June 19, 2025. 2023.

[117] Liyang Wang, Jinxin Zhao, and Liangjun Zhang. "Navdog: robotic navigation guide dog via model predictive control and human-robot modeling". In: *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 2021, pp. 815–818.

[118] Anxing Xiao et al. "Robotic guide dog: Leading a human with leash-guided hybrid physical interaction". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2021.

[119] Yanbo Chen et al. "Quadruped guidance robot for the visually impaired: A comfort-based approach". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2023.

[120] Hochul Hwang et al. "System configuration and navigation of a guide dog robot: Toward animal guide dog-level guiding work". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2023.

[121] J Taery Kim et al. “Transforming a Quadruped into a Guide Robot for the Visually Impaired: Formalizing Wayfinding, Interaction Modeling, and Safety Mechanism”. In: *Conference on Robot Learning*. PMLR. 2023.

[122] David DeFazio, Eisuke Hirota, and Shiqi Zhang. “Seeing-Eye Quadruped Navigation with Force Responsive Locomotion Control”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 2184–2194.

[123] Hochul Hwang et al. “Towards Robotic Companions: Understanding Handler-Guide Dog Interactions for Informed Guide Dog Robot Design”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–20.

[124] Hochul Hwang et al. “Synthetic Data Augmentation for Robotic Mobility Aids to Support Blind and Low”. In: *Robot Intelligence Technology and Applications 9: Results from the 12th International Conference on Robot Intelligence Technology and Applications*. Vol. 1419. Springer Nature. 2025, p. 92.

[125] Aviv L Cohav et al. “Do Looks Matter? Exploring Functional and Aesthetic Design Preferences for a Robotic Guide Dog”. In: *arXiv preprint arXiv:2503.16450* (2025).

[126] Rayna Hata et al. *See Spot Guide: Accessible Interfaces for an Assistive Quadruped Robot*. 2024. arXiv: 2402.11125 [cs.RO].

[127] J Taery Kim et al. “Understanding expectations for a robotic guide dog for visually impaired people”. In: *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2025, pp. 262–271.

[128] Yonatan Bisk et al. “Experience Grounds Language”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 8718–8735.

[129] Yohei Hayamizu, Zhou Yu, and Shiqi Zhang. “Learning Joint Policies for Human-Robot Dialog and Co-Navigation”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2023, pp. 7893–7898.

[130] Joyce Y Chai et al. “Collaborative effort towards common ground in situated human-robot dialogue”. In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2014, pp. 33–40.

[131] Saeid Amiri et al. “Augmenting knowledge through statistical, goal-oriented human-robot dialog”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 744–750.

[132] Chun-Yen Chen et al. “Gunrock: Building a human-like social bot by leveraging large scale real user data”. In: *Alexa prize proceedings* (2018).

[133] Tom Silver et al. “PDDL planning with pretrained large language models”. In: *NeurIPS 2022 foundation models for decision making workshop*. 2022.

[134] Bo Liu et al. “Llm+ p: Empowering large language models with optimal planning proficiency”. In: *arXiv preprint arXiv:2304.11477* (2023).

[135] Anthony Brohan et al. “Do as i can, not as i say: Grounding language in robotic affordances”. In: *Conference on Robot Learning*. PMLR. 2023, pp. 287–318.

[136] Carson Stark et al. “Dobby: A Conversational Service Robot Driven by GPT-4”. In: *arXiv preprint arXiv:2310.06303* (2023).

[137] Xuan Yang. *Project Guideline: Enabling Those with Low Vision to Run Independently*. https://research.google/blog/project-guideline-enabling-those-with-low-vision-to-run-independently/. 2021.

[138] Jan Balata, Zdenek Mikovec, and Pavel Slavik. "Conversational agents for physical world navigation". In: *Studies in Conversational UX Design* (2018), pp. 61–83.

[139] Brian FG Katz et al. "NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio". In: *Virtual Reality* 16 (2012), pp. 253–269.

[140] Jakub Berka et al. "Misalignment in semantic user model elicitation via conversational agents: a case study in navigation support for visually impaired people". In: *International Journal of Human–Computer Interaction* 38.18-20 (2022), pp. 1909–1925.

[141] Abdelsalam Helal, Steven Edwin Moore, and Balaji Ramachandran. "Drishti: An integrated navigation system for visually impaired and disabled". In: *Proceedings fifth international symposium on wearable computers*. IEEE. 2001, pp. 149–156.

[142] Ching-Han Chen and Ming-Fang Shiu. "Smart guiding glasses with descriptive video service and spoken dialogue system for visually impaired". In: *IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*. 2020.

[143] Jan Balata, Zdenek Mikovec, and Ivo Maly. "Navigation problems in blind-to-blind pedestrians tele-assistance navigation". In: *Human-Computer Interaction–*

INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part I 15. Springer. 2015, pp. 89–109.

[144] Shivendra Agrawal, Mary Etta West, and Bradley Hayes. "A novel perceptive robotic cane with haptic navigation for enabling vision-independent participation in the social dynamics of seat choice". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022.

[145] Anis Koubâa. *Robot Operating System (ROS)*. Springer, 2017.

[146] Nickolay V. Shmyrev et al. *Vosk Speech Recognition Toolkit*. `https://github.com/alphacep/vosk-api`. 2023.

[147] Anthony G Cohn and Jose Hernandez-Orallo. "Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs". In: *arXiv preprint arXiv:2304.11164* (2023).

[148] Stephanie Rosenthal, Sai P Selvaraj, and Manuela M Veloso. "Verbalization: Narration of Autonomous Robot Experience." In: *IJCAI*. 2016.

[149] Yohei Hayamizu et al. "Guiding robot exploration in reinforcement learning via automated planning". In: *Proceedings of the International Conference on Automated Planning and Scheduling*. Vol. 31. 2021, pp. 625–633.

[150] Shyam Sundar Kannan, Vishnunandan LN Venkatesh, and Byung-Cheol Min. "Smart-llm: Smart multi-agent robot task planning using large language models". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2024, pp. 12140–12147.

[151] Krishan Rana et al. “SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Task Planning”. In: *arXiv preprint arXiv:2307.06135* (2023).

[152] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. “Wizard of Oz studies: why and how”. In: *Proceedings of the 1st international conference on Intelligent user interfaces*. 1993, pp. 193–200.

[153] Maximillian Chen et al. “PLACES: Prompting language models for social conversation synthesis”. In: *arXiv preprint arXiv:2302.03269* (2023).

[154] Zhen Tan et al. “Large language models for data annotation and synthesis: A survey”. In: *arXiv preprint arXiv:2402.13446* (2024).

[155] Eric Huang, Zhenzhong Jia, and Matthew T Mason. “Large-scale multi-object rearrangement”. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 211–218.

[156] Jiayuan Gu et al. “Multi-skill Mobile Manipulation for Object Rearrangement”. In: *arXiv preprint arXiv:2209.02778* (2022).

[157] Jacky Liang et al. “Code as policies: Language model programs for embodied control”. In: *arXiv preprint arXiv:2209.07753* (2022).

[158] Xiaohan Zhang et al. “Visually grounded task and motion planning for mobile manipulation”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 1925–1931.

[159] Walter Goodwin et al. “Semantically grounded object matching for robust robotic scene rearrangement”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 11138–11144.

[160] Vasileios Vasilopoulos et al. “Reactive planning for mobile manipulation tasks in unexplored semantic environments”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 6385–6392.

[161] David Coleman et al. “Reducing the barrier to entry of complex robotic software: a moveit! case study”. In: *arXiv preprint arXiv:1404.3785* (2014).

[162] Marcel Heerink et al. *Assessing acceptance of assistive social agent technology by older adults: the almere model*. 2010.