



Doctoral Thesis

**Communicating Object-handling Meaning
through Robot Manipulation**

Xiang PAN

January 2026

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Doctoral Thesis
submitted to Department of Social Informatics,
Graduate School of Informatics,
Kyoto University
in partial fulfillment of the requirements for the degree of
DOCTOR of INFORMATICS

Thesis Committee: Takayuki Kanda, Professor
Takayuki Ito, Professor
Michinori Hatayama, Professor

Communicating Object-handling Meaning through Robot Manipulation*

Xiang PAN

Abstract

Robots are expected to play increasingly important roles in future applications, and dual-arm robots with human-like structures and capabilities are particularly significant. Designed with the human body as inspiration, they aim to achieve dexterity and object manipulation abilities that support diverse and complex tasks in everyday environments. In future human-centered settings, object manipulation is expected to become a core component of a robot’s capability framework. Many daily human activities involve grasping, placing, transferring, and coordinating multiple objects, making stable and flexible manipulation essential for achieving high-quality human-robot interaction.

Dual-arm robots offer clear advantages in this regard. Their dual-arm structure enables coordinated actions that more closely resemble human behavior, making them well suited for complex and delicate manipulation tasks. As robots move from controlled industrial environments into open settings such as homes, stores, and public spaces, the manner in which they manipulate objects not only influences task success but also shapes people’s understanding and expectations of the robot and the objects it handles. This understanding, in turn, affects perceived safety and efficiency during interaction, highlighting object manipulation as a key element in enabling robots to integrate into future human environments.

However, much of the existing research on object manipulation has primarily focused on physical or functional objectives, such as how to move an object to a target location, maintain a stable grasp, or complete a task successfully. Less attention has been given to whether the manner in which a robot manipulates an object conveys meaning to human observers or reflects how the object ought to be handled in a socially appropriate way. In contrast, human manipulation behavior is shaped by object properties. For example, when handling heavy objects,

*Doctoral Thesis, Department of Social Informatics, Graduate School of Informatics, Kyoto University, KU-I-DT6960-34-2283, January 2026.

people tend to adopt more stable grasp positions and use both hands to prevent accidental drops. When handling fragile items, people often move more slowly and carefully to avoid damage. These manipulation strategies arise not only from physical requirements but also convey important information to observers about the object and its appropriate handling. How robots can similarly express such information through manipulation behavior, however, remains largely unexplored.

This thesis treats object manipulation as a non-verbal communication channel with expressive potential. The central idea is that the way a robot grasps, moves, and handles an object can communicate information about the object and shape human understanding. Through three studies, this thesis investigates whether and how robots can express such information through manipulation behavior, and examines how users perceive and interpret these cues. The goal is to enable robots to convey object-related information in ways that align with human expectations by leveraging how objects are appropriately handled.

The first study investigates whether a dual-arm robot can make meaningful use of its second hand in social contexts when a single hand is sufficient to complete the task. Inspired by the behavior of human shopkeepers, this study develops a model in which the idle hand provides supportive actions during object handling. User studies show that supportive behavior for the idle hand enhances perceived politeness and competence, demonstrating that such behavior can function as a form of social communication.

The second study examines how robot manipulation behaviors can convey physical object properties. It introduces a motion generation method that adjusts handling strategies based on object properties such as weight, temperature, or fragility. Experimental results indicate that people can infer these properties from the robot’s manipulation behavior, showing that physical object properties can be expressed through manipulation behaviors.

The third study explores whether robots can convey relations between objects through their manipulation behaviors. Using a large language model, this study develops a system that generates coordinated dual-arm behaviors to express relational meanings, such as similarity or difference. User studies reveal that people can identify these relations from the robot’s actions and that such behaviors are associated with more positive perceptions of the robot. This suggests that semantic relations between objects can be communicated through coordinated

manipulation.

Taken together, these three studies provide conceptual, theoretical, and computational foundations for understanding how robot manipulation behavior can support both task execution and communication. By demonstrating how robots can convey important information about objects through appropriate handling behaviors, this thesis advances a perspective in which a robot’s manipulation behaviors fulfill both functional and communicative roles in human-centered environments.

Keywords: Human-robot Interaction, Non-verbal Communication, Object-handling Meaning, Object Manipulation, Idle Hand Motion, Support Hand Behaviors, Object Properties, Property-based Motion Planner, Object Relations, Relational Gestures, Large Language Models (LLMs)

CONTENTS

1	Introduction	1
1.1	Background and Motivation	1
1.2	Research Gap and Research Questions	2
1.3	Key Challenges	5
1.4	Research Contributions	6
1.5	Thesis Structure	7
2	Modeling Support Behavior for Robot's Idle Hand	9
2.1	Introduction	9
2.2	Related Work	11
2.2.1	Gestures in Human Communications	11
2.2.2	Idle Motions	11
2.2.3	Two-handed Motions	12
2.2.4	Social Cues of Hand Motions	13
2.3	Model of Support Hand Behavior	13
2.3.1	Observation of Human Shopkeeper Behaviors	13
2.3.2	Analysis of Support Hand Behavior	15
2.3.3	Reasons for Support Hand	16
2.3.4	Modeling	16
2.4	Implementation	18
2.4.1	Robot Hardware	18
2.4.2	Architecture	19
2.4.3	Collision Avoidance	19
2.4.4	Support Motion	20

Contents

2.4.5	Motion Planner	20
2.4.6	Trajectory Execution	21
2.5	Evaluation	21
2.5.1	Hypotheses	21
2.5.2	Participants	22
2.5.3	Conditions	22
2.5.4	Procedure	23
2.5.5	Measurement	23
2.6	Results	24
2.6.1	Observation	24
2.6.2	Verification of Hypotheses	25
2.6.3	Interview Results	26
2.7	Discussion	27
2.7.1	Design Implications	27
2.7.2	Should Robots Exhibit Support Hand Behavior?	28
2.7.3	Which Occasions Should Robots Exhibit Support Hand Behavior?	28
2.7.4	Generalizability across Cultures	28
2.7.5	Limitations	29
2.8	Conclusion	29
3	Communicating Physical Properties Through Robot Object Manipulation	31
3.1	Introduction	31
3.2	Related Work	33
3.2.1	Gestures for Communications	33
3.2.2	Legibility	33
3.2.3	Communication of Object Properties	34
3.3	How Do People Communicate Properties through Object Manipulation?	34
3.3.1	Selection of Object Properties	34
3.3.2	Observations of Human Hand Behaviors	35
3.4	Property-based Planner for Communicating Object Properties	41
3.4.1	Architecture	41
3.4.2	Robot Key Elements	42

Contents

3.4.3	Utility	42
3.4.4	Action Sequence Generator	44
3.4.5	Motion Planner	44
3.4.6	Implementation	45
3.4.7	Examples of Planning Results	45
3.5	Evaluation	45
3.5.1	Participants	48
3.5.2	Conditions	48
3.5.3	Procedure	48
3.6	Results	49
3.6.1	Recognition Results	49
3.6.2	Interview Results	50
3.7	Discussion	52
3.7.1	Generalizability	52
3.7.2	Optimization for System Performance	53
3.7.3	Limitations	54
3.8	Conclusion	54
4	Communicating Object Relations Through Robot Gestures	55
4.1	Introduction	55
4.2	Related Work	57
4.2.1	Gestures in Human Communications	57
4.2.2	Gestures in Human-Robot Interaction	57
4.2.3	Co-speech Gesture Generation	57
4.2.4	Large Foundation Models in Robotics	58
4.3	Relational Gestures During Object Comparisons	58
4.3.1	Data Collection from Human Demonstrations	58
4.3.2	Gesture Analysis	60
4.4	Relational Gesture Generation System	64
4.4.1	Overview	64
4.4.2	Co-Speech Gesture Generator	65
4.4.3	Relational Gesture Planner	66
4.4.4	Gesture Integrator	67
4.4.5	Robot	67
4.5	Study I: Relational Gestures Recognition	68

Contents

4.5.1	Participants	68
4.5.2	Stimuli	68
4.5.3	Procedure	69
4.5.4	Results	70
4.6	Study II: Co-Speech Relational Gesture Evaluation	71
4.6.1	Hypotheses and Predictions	71
4.6.2	Participants	72
4.6.3	Conditions	72
4.6.4	Procedure	73
4.6.5	Measurement	74
4.6.6	Results	74
4.7	Discussion	76
4.7.1	Design Implications	76
4.7.2	Limitations	76
4.8	Conclusion	77
5	General Discussion	78
5.1	Unifying Framework	78
5.2	Technical Differences	79
5.3	Design Principles for Scalable Expressive Manipulation	80
5.4	Is Manipulation Necessary for Conveying Information?	80
5.5	Cognitive Foundations of Interpreting Expressive Manipulation	81
5.6	Generalizability	82
6	Conclusions and Future Directions	84
6.1	Conclusions	84
6.2	Future Directions	85
6.2.1	Multimodal and Context-Rich Expression	85
6.2.2	Cognitive Modeling of Human Interpretation	85
6.2.3	Human-Robot-Object Triadic Interaction Framework	86
	Acknowledgements	87
	Selected List of Publications	102

LIST OF FIGURES

1.1	A conceptual framework illustrating the relationship between expression modality and communicative focus in non-verbal robot communication through motion. This thesis focuses on the upper-right quadrant, where robots convey object-handling meaning through expressive manipulation.	3
1.2	Overview of research structure.	6
2.1	An example highlighting the importance of hand motion subtleties.	10
2.2	Situations where the support hand behavior occurs during shopkeeper-customer interaction.	14
2.3	An example of how a human shopkeeper returns to support an object with the idle hand.	15
2.4	Schematic diagram of support hand behavior.	17
2.5	Diagram of implementation of proposed model.	19
2.6	Robot in proposed condition.	24
2.7	Robot in baseline condition.	24
2.8	Social behaviors toward the robot.	25
2.9	Results. The error bars show the standard error for the mean at $\pm 1 SE$. ($+ : .05 < p < .1$, $* : p < .05$, $** : p < .01$)	26
3.1	Communication of object properties during handovers.	32
3.2	Examples of human key elements for communicating object properties.	37
3.3	The Diagram of proposed planner.	42

List of Figures

3.4	Robot key elements corresponding to the human key elements illustrated in Fig. 3.2.	43
3.5	The planning results for <i>light</i> at different time ratios.	46
3.6	The planning results for <i>hot</i> at different time ratios.	47
3.7	Setup for experiments.	49
3.8	Relationship between recognition rate and time ratio.	49
4.1	Robot gestures during object comparison.	56
4.2	Representative gestures for each object relation.	64
4.3	The diagram of proposed system.	65
4.4	Examples of robot relational gestures.	68
4.5	Setup for the experiment in Study I.	69
4.6	Examples of the robot's behavior in Study II.	73
4.7	Results from our quantitative measures. The error bars show the standard error for the mean at $\pm 1 SE$. (** : $p < .01$, *** : $p < .001$)	75

LIST OF TABLES

2.1	Observed hand behaviors	16
3.1	Eight categories of object properties	36
3.2	Key elements for object properties	38
3.3	The recognition results for object properties at different time ratios	51
4.1	Hand pattern	62
4.2	Hand direction	62
4.3	Focus of attention	63
4.4	The recognition accuracy for object relations	70

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

As social robots become increasingly integrated into public environments such as retail stores [20, 60], museums [27, 34], schools [38, 86], and healthcare facilities [1, 54], their roles are evolving from purely functional assistants to socially expressive partners capable of engaging people through natural, multimodal interaction.

To interact fluently with humans, robots must go beyond verbal communication and incorporate non-verbal modalities that convey intentions, emotions, and personalities. Among these, hand movements play a central role, as they attract attention, complement speech, and reveal intentions [30, 44]. For example, symbolic gestures communicate explicit and language-like meaning, such as a waving gesture used to express greetings, similar to saying “hello.” Expressive gestures, in contrast, primarily communicate affective or interpersonal qualities rather than explicit semantics. For example, slow and smooth arm movements can appear gentle and polite. Importantly, human hand movements are not limited to symbolic or expressive gestures. In everyday life, people naturally use their hands to manipulate objects in ways that, whether consciously or subconsciously, reveal their underlying intentions and shape how others perceive them. For example, when a person quickly switches an object between two hands, the movement may indicate that the object is hot. Similarly, when a person brings two held ob-

jects closer together, the movement potentially suggests that the two objects are similar in certain aspects.

This suggests that manipulation behaviors can extend beyond mechanical execution, functioning instead as a medium for social communication. This thesis is motivated by the question of whether, and in what ways, robots can achieve comparable expressivity in their physical interactions with objects.

1.2 Research Gap and Research Questions

Robotic manipulation is defined not only by the physical motions robots produce, but also by the way they treat and handle objects during those motions. Objects, as the targets of manipulation, play a central role in human-robot interaction (HRI), especially in collaborative tasks such as handovers, where success depends not only on physical coordination but also on a shared understanding of how objects are handled. For example, when an object is handled slowly and near the ground, this may suggest that it is heavy. To achieve natural and safe collaboration, robots must therefore not only act upon objects but also communicate to humans how objects should be handled through their manipulation behaviors.

However, research on communicating how objects should be handled through manipulation remains limited. Existing work on non-verbal robot communication through motion in HRI can be broadly characterized along two conceptual dimensions: (1) *Expression modality*, spanning from non-manipulative cues, such as gestures, gaze, and posture, to manipulation-based actions; and (2) *Communicative focus*, ranging from representational meaning, which conveys what the robot or object is, as well as what the robot feels or intends, to object-handling meaning, which refers to the physical and social information implicating how objects should be handled or interacted with.

Prior work has primarily focused on representational meaning. For example, expressive gesture studies coordinate hand and body movements to convey internal states such as emotion or personality [47, 56, 58], while iconic gesture research uses hand motion to represent visually perceivable object properties such as shape or size [32, 33, 81]. Legible motion studies design one-handed manipulation behaviors that make robots’ goals visible and interpretable to humans [22, 23]. These studies improve transparency and social expressivity but focus on communicating what the robot or object is, rather than how the object should

be handled.

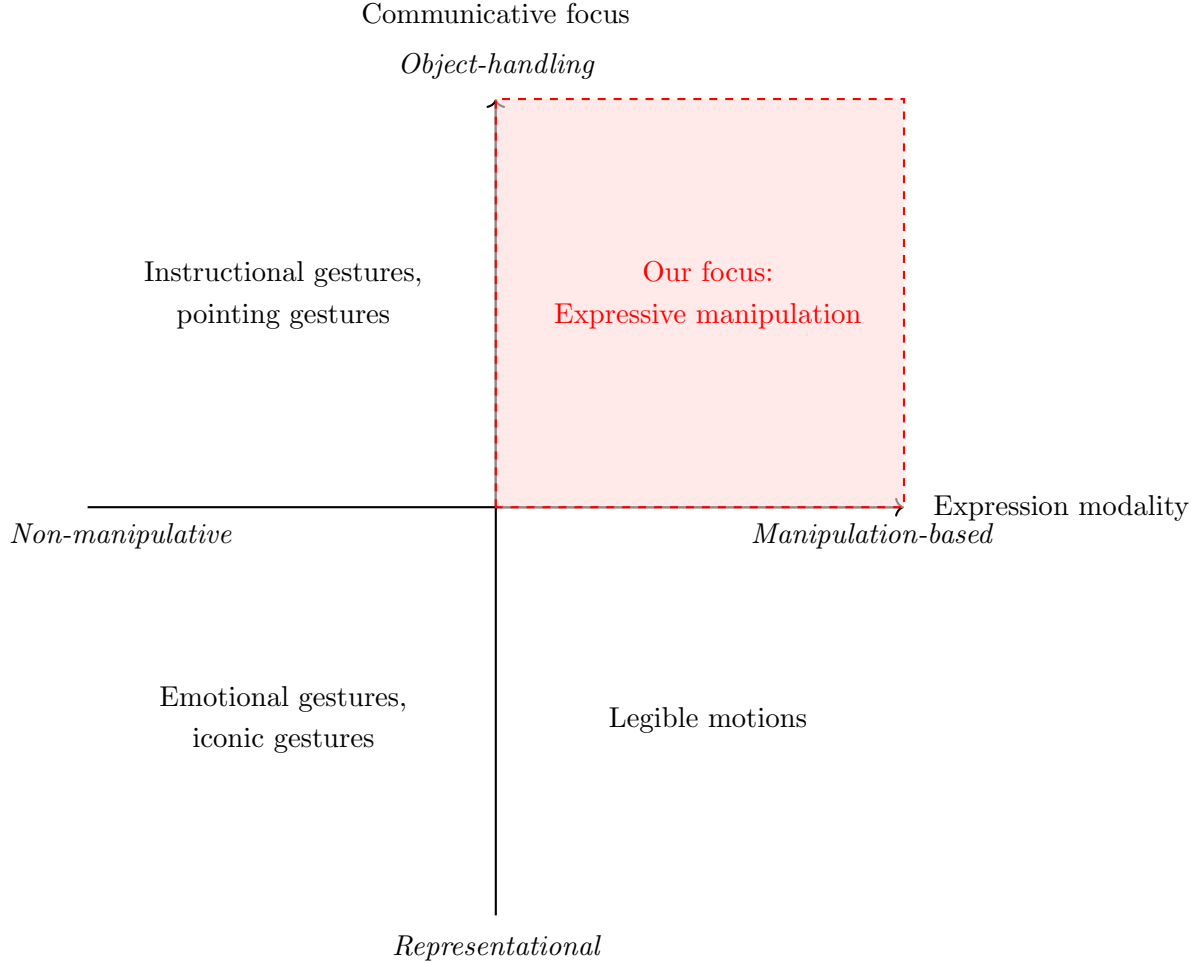


Figure 1.1: A conceptual framework illustrating the relationship between expression modality and communicative focus in non-verbal robot communication through motion. This thesis focuses on the upper-right quadrant, where robots convey object-handling meaning through expressive manipulation.

There is also research that communicates object-handling meaning without performing object manipulation. For example, instructional gestures guide humans on how to move or operate on objects in collaborative tasks [85], and pointing gestures indicate where an object should be grasped or placed [76]. While these studies do convey how objects should be handled, they rely on they express this information through gestures rather than through manipulation behaviors.

In contrast, this thesis focuses on manipulation as a communication channel

because object-handling meaning is inherently tied to the way objects are physically handled. Humans can intuitively infer how an object is handled through motion dynamics such as how many hands are involved, the timing and speed of movement, or the placement of the hands on the object. By leveraging object manipulation itself as the medium for communication, this research explores a space that has remained largely underexplored.

To illustrate this landscape, Fig. 1.1 organizes existing work along these two dimensions of expression modality and communicative focus. Within this framework, existing studies can be grouped into four quadrants:

- **Lower-left: Non-manipulative, representational communication.** Robots express internal states such as emotion or personality [47, 56, 58], or represent visible object properties through gestures and body movements [32, 33, 81]. These cues improve social transparency but do not convey meaning related to how objects are treated.
- **Upper-left: Manipulation-based, representational communication.** Legible motion makes the robot’s goal or intention clear to humans through object manipulation [22, 23], but does not express the meaning behind how the object is being handled.
- **Lower-right: Non-manipulative, object-handling communication.** Instructional gestures and pointing gestures communicate how or when to grasp or handle an object [76, 85], but do not express this meaning through the manipulation itself.
- **Upper-right: Manipulation-based, object-handling communication (our focus).** This thesis examines how robots can express object-handling meaning through manipulation itself, allowing humans to understand the implied information behind the handling directly from the robot’s physical handling behaviors.

Accordingly, this thesis addresses two key research questions:

- **RQ1:** How do humans interpret object-handling meaning through robot manipulation behaviors?
- **RQ2:** How can robots generate manipulation behaviors that are physically feasible while effectively communicating object-handling meaning?

1.3 Key Challenges

Communicating object-handling meaning through manipulation introduces challenges that extend beyond conventional motion control or gesture design. Unlike representational cues that express what an object is or what the robot intends, object-handling meaning concerns how an object should be handled in different contexts. For example, fragile objects may be handled carefully, heavy objects may be grasped with two hands or carried low, and objects sharing a functional property may be moved in a coordinated manner. Communicating such meaning requires understanding how humans interpret these subtle motion cues, and developing systematic methods that allow robots to express such meaning through manipulation.

This thesis therefore addresses two central challenges:

- **Challenge 1: Limited understanding of human interpretation of object-handling meaning.** When robots manipulate objects, multiple motion cues emerge, such as the speed of movement, spatial configuration, and placement on the objects. However, it remains unclear which motion cues reliably convey specific object-handling meanings, and which may be ambiguous or misleading. The first challenge is to identify the perceptual mechanisms by which humans infer how an object should be handled based on robot manipulation.
- **Challenge 2: Lack of methods for generating physically feasible yet interpretable manipulation behaviors.** Even with an understanding of human interpretation, most manipulation frameworks optimize for stability, efficiency, or collision avoidance and do not explicitly encode communicative intent. There is currently no systematic mapping between object-handling meaning and feasible motion parameters. The second challenge is therefore to develop a manipulation framework that links object-handling meaning to feasible motion dynamics, enabling robots to physically express object-handling meaning while maintaining reliable control.

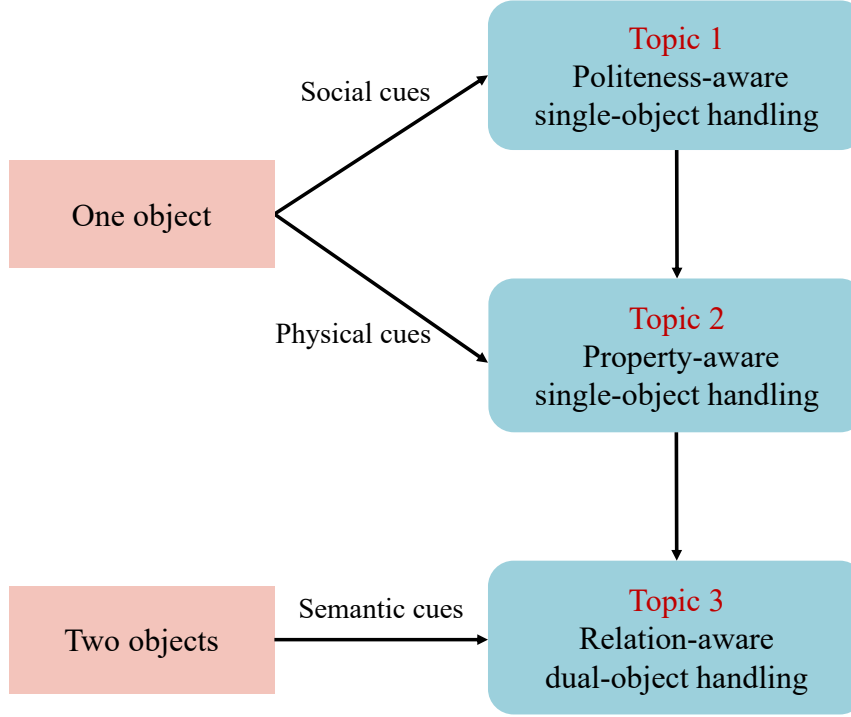


Figure 1.2: Overview of research structure.

1.4 Research Contributions

This thesis investigates how robot manipulation behaviors can communicate object-handling meaning to humans through three interrelated studies. The research progresses from examining whether a dual-arm robot should use one hand or two when handling a single object, to determining how a robot should adapt its handling for objects with different properties, and finally to extending these handling behaviors to scenarios involving two objects with semantic relations. Fig. 1.2 illustrates this conceptual progression.

- **Study 1 (Chapter 2): Modeling support behavior for the robot’s idle hand.** This study examines a fundamental question of whether a dual-arm robot should use both hands when only one is needed to handle a single object in social contexts. Through observation of human shopkeepers, this study identifies support behaviors performed by the idle hand when handling an object. A model of this support behavior is implemented on a dual-arm robot. User studies show that two-handed handling improves

perceived politeness and competence, suggesting that handling behavior carries social meaning beyond task execution.

- **Study 2 (Chapter 3): Communicating physical properties through robot object manipulation.** Building upon politeness-aware single-object handling, this study examines how manipulation behaviors can convey object properties such as weight, temperature, and fragility. We develop a property-based motion planner that adjusts two-handed manipulation to express the intended object properties. User studies show that humans can reliably infer these properties solely from the robot’s movements, demonstrating that physical handling serves as an effective channel for communicating object-related physical information.
- **Study 3 (Chapter 4): Communicating object relations through robot gestures.** Extending to dual-object scenarios, this study examines how a robot can express relations between two objects (e.g., similar, different) through coordinated bimanual behaviors. We introduce a large language model (LLM)-based system that generates manipulation-based gestures to express the intended semantic relations between objects. User studies reveal that humans can infer these relations purely from movement, and that relational handling enhances the robot’s perceived competence, sociability, and animacy, showing that object-handling meaning generalizes beyond single-object settings to dual-object interactions.

To summarize, these studies make three primary contributions. First, the thesis establishes a conceptual foundation that frames robot manipulation as a form of expressive non-verbal communication. Second, it provides empirical evidence through a series of user studies demonstrating that people can infer the meanings expressed through robots’ manipulation behaviors across social, physical, and semantic contexts. Third, it introduces computational methodologies that enable robots to convey how objects should be handled through their manipulation.

1.5 Thesis Structure

The remainder of the thesis is organized as follows. Chapter 2 models the support behavior for a robot’s idle hand and examines how such behavior influences

human perceptions of robots in interaction. Chapter 3 develops a property-based motion planner to enable robots to convey physical properties through object manipulation, and evaluates whether people can accurately infer these properties from robot manipulation. Chapter 4 introduces an LLM-based gesture generation system to enable robots to express semantic relations between two objects, and investigates how these gestures affect human interpretation and the robot’s social acceptance. Chapter 5 presents a general discussion of the main findings. Finally, Chapter 6 summarizes the main findings and outlines future research directions.

CHAPTER 2

MODELING SUPPORT BEHAVIOR FOR ROBOT’S IDLE HAND

2.1 Introduction

In recent years, the field of robotics has witnessed significant advances in human-robot interaction, with social robots increasingly being deployed in museums [27, 34], retail stores [20, 60, 68], schools[42, 86], and healthcare [1, 54]. While these technological advancements are undeniably impressive, one cannot help but notice a significant gap in the robots’ ability to convey social presence.

The subtleties of hand motion, for instance, play a crucial role in human interactions, but they are often overlooked in robotic design. Imagine walking into a shop where a robot acts as the shopkeeper (Fig. 2.1 (a)): how would you feel if the robot continually handled a box containing a diamond ring with one hand while its other hand remained idle?

In the human realm, a shopkeeper who interacts with products using only one hand is often perceived as inattentive or even impolite. A competent shopkeeper, in general, actively coordinates both hands in a manner that not only increases efficiency but also adds a layer of social appropriateness and politeness. For instance, in the given scenario where a diamond ring is being shown, a human shopkeeper would typically hold the box with one hand while using the other to

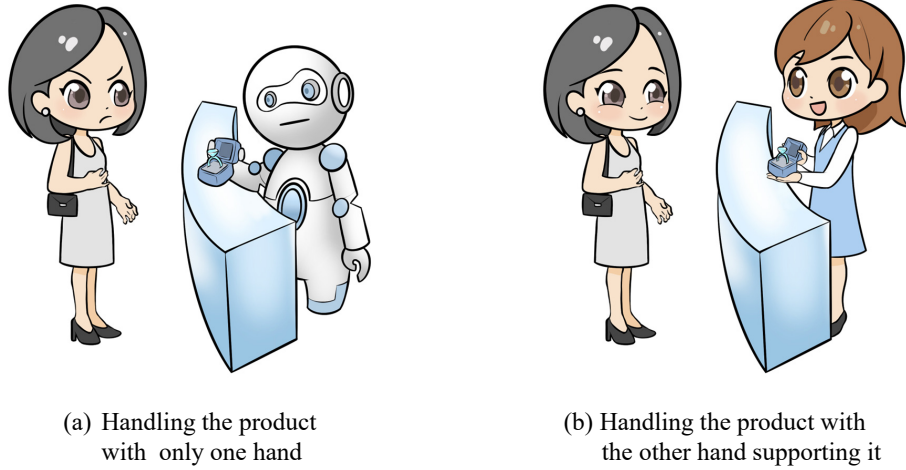


Figure 2.1: An example highlighting the importance of hand motion subtleties.

provide additional support, as illustrated in Fig. 2.1 (b). In contrast, a robot shopkeeper that persists in using only one hand could be seen as less lively, lacking the depth of a social entity we would naturally expect.

Recently, there has been a sustained research effort toward improving the acceptance of social robots as social entities. One concept related to this effort is idle motion. It can present a basic level of the “illusion of life,” which helps people regard robots as social entities [39]. Idle motions refer to “adaptor movements,” i.e., postural or other non-verbal movements often performed during idle moments [6]. Previous studies mainly explored the impact of head motion (e.g., head-scratching), face motion (e.g., eye-blinking, mouth movement), arm motion (e.g., arm-folding), and leg motion (e.g., leg-swinging) [17, 41, 49] on enhancing robots’ social presence when robots are not engaged in specific tasks.

However, our scenario presents a unique challenge: the robot is not in an idle state, but one of its hands, which is not required for the object manipulation task, remains idle. How this unused hand (hereinafter referred to as the idle hand) should move while the primary hand executes manipulative tasks remains an open question.

To the best of our knowledge, little research has been conducted on the coordination of the idle hand with the primary hand for social purposes. To this end, we introduce the novel concept of “idle hand motion,” which here refers to the coordination of the idle hand with the primary hand during ongoing manipulation. Our work makes two main contributions. First, we propose the concept

of idle hand motion and its corresponding model. Second, we explore the impact of idle hand motion on enhancing a robot’s social acceptance, particularly in the role of a shopkeeper.

2.2 Related Work

2.2.1 Gestures in Human Communications

People commonly use gestures during interactions to draw the attention of listeners [30], disambiguate unclear speech, and supplement speech with additional information [44]. All visible hand movements except for self-touching and object-manipulations are regarded as gestures [44, 64]. In general, gestures can be divided into four common types: (1) iconics, which have a close relationship to the semantic content of speech, (2) metaphors, which present an image of an abstract concept such as knowledge, (3) deictics, which are pointing movements towards concrete entities or abstract space, and (4) beats, which are typically frequent biphasic (e.g., up-and-down, back-and-forth) movements [64]. These gestures greatly support human communication, and they enhance learning and memory. Feyereisen [26] investigated that sentences with representational gestures are easier to recall for listeners than nonrepresentational ones. Cook et al. [16] found that gestures help children learn new mathematical concepts during instruction.

In our study, idle hand motion can be used to draw attention and convey politeness and competence. Therefore, idle hand motion can be considered to be a gesture.

2.2.2 Idle Motions

Idle motions in animations refer to micro-movements that occur in idle states between animation clips [25]. These subtle motions imbue agents with lifelike qualities. For example, Kocoń [50] proposed a method to synthesize idle motion for virtual persons, enhancing perceptions of friendliness. Egges et al. [25] developed a flexible idle motion engine for virtual agents in idle states.

Research on human-robot interaction has also recognized that micro-movements play an important role in making robots animated and alive in the idle state. So-

cial robots such as Robovie [41] use idle motions like “scratch the head” and “fold its arm” to indicate they are lifelike, but how these idle motions affect the human perception of robots has not been investigated. Song et al. [89] observed clerks via video ethnography, extracted several idle motions (e.g., mouth movement, arm movement), and tested whether these idle motions could be recognized when they were applied to a robot. Cuijpers and Knops [17] conducted an experiment in which participants were exposed to different levels of social verification. They found that robots with low levels of social verification that portrayed idle motions (e.g., breathing, gaze shift motion), were more anthropomorphic and emotionally expressive compared to robots lacking such motions.

To summarize, existing studies focus on the synthesis of idle motions from human behavior observations, primarily in contexts where robots are in an idle state. However, the use of the idle hand during manipulation tasks has not yet been investigated.

2.2.3 Two-handed Motions

Recent research in human-robot interaction has extensively explored two-handed gestures. Huang and Mutlu [32] analyzed how participants used gestures while describing a paper-making activity, and further investigated the use of gestures for enhancing interaction. They later modeled the coordination of speech, gaze, and gesture in narration using a dynamic bayesian network [33].

Research on two-handed manipulation has been intensively pursued over the past few decades. Bai et al. [7] presented an approach of dual-arm coordinated control for twisting manipulation, merging optimized motion planning with real-time human-led teleoperation. Vahrenkamp et al. [94] proposed a Grasp-RRT algorithm to generate collision-free dual-arm grasping motions. Bestick et al. [8] developed an approach to estimating personalized human kinematic models, and it could generate safer and more motion for two-handed robot-to-human handovers.

Existing research on two-handed gestures and manipulations has primarily concentrated on coordinating hand motions where both hands are essential for communication or task execution. By contrast, our study is unique in examining how one unused hand can be coordinated with the other hand, which sets our research apart from the existing literature.

2.2.4 Social Cues of Hand Motions

Legibility, i.e., the intuitive understanding of robots’ intentions, is crucial for coordinating joint action and positive perception [59]. Dragan et al. [22] developed a model for generating goal-directed legible arm motions and legible pointing gestures [31]. Bodden et al. [10] introduced a method for synthesizing legible manipulation motions, enhancing the robot’s communicative intent.

Polite hand motions are another way to non-verbally signal intent to humans. Politeness can be manifested by etiquette that makes use of verbal, physical, and gestural modes of interaction [66]. A representative example of this is the practice of passing objects with both hands to convey politeness, a gesture deeply rooted in traditions like the Japanese tea ceremony [91]. Central to politeness is Brown and Levinson [11]’s Politeness Theory, emphasizing the importance of maintenance of “face” in social interactions. This theory has been widely applied in the field of HRI. Kumar et al. [53] noted that a robot’s polite behavior influenced user perceptions. Ritschel et al. [79] found gender-based preferences for a companion robot: male participants favored polite robotic feedback, while females opted for more direct commands. Rea et al. [77] found that although an impolite robot was not preferred during physical exercises, it pushed participants to perform better than a polite one. Lee et al. [57] explored how perceived politeness levels in a robot’s speech and gestures can influence a user’s intention to comply.

While existing research predominantly focuses on verbal communication, our study aims to address the gap in understanding politeness through non-verbal cues, particularly hand motions.

2.3 Model of Support Hand Behavior

2.3.1 Observation of Human Shopkeeper Behaviors

Scenario and Motivations

To understand strategies for idle hand usage during object-based interactions, we initially observed shopkeepers’ hand behaviors in ten different retail stores informally, with a particular focus on how the idle hand was used. We noticed that for expensive items like watches and jewelry shopkeepers often used their idle hand to either directly touch or maintain a slight distance from the bottom

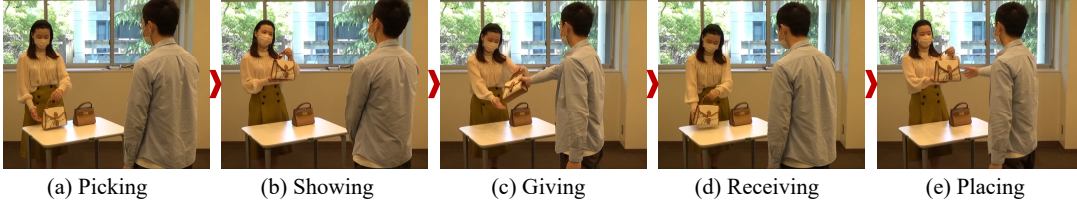


Figure 2.2: Situations where the support hand behavior occurs during shopkeeper-customer interaction.

of them as a support. By contrast, this behavior was less observed with cheap items like snacks and daily necessities.

However, due to limitations in video recording, we could not quantify this hand behavior. Consequently, we conducted an observation of object-based interaction between a shopkeeper and a customer in a simulated handbag store, where two handbags were placed on a table. We selected expensive handbags with the expectation that more interesting interactions would occur during their presentation. We then invited four participants with shopkeeper work experience (2f, 2m) to conduct role-plays where one acted as a shopkeeper and the other as a customer. We focused on how the handbags were presented and handled during the sales process, without specific mention of idle hand usage.

Results

From observation of the shopkeepers, we found that one behavior, the shopkeeper’s “support” hand behavior, consistently reoccurs across multiple interactions as expected. In other words, when the primary hand is tasked with manipulating an object, the other hand, which could actually remain in an idle state, is no longer idle but positions itself to support the object.

We observed that this behavior usually took place in five situations during shopkeeper-customer interactions: picking, showing, giving, receiving, and placing for the primary hand (Fig. 2.2). In Fig. 2.2 (a), a female shopkeeper grasped a handbag with her left hand while her right hand provided support from beneath. In Fig. 2.2 (b), the shopkeeper held the handbag with the left hand while the other hand supported it. In Fig. 2.2 (c), the shopkeeper used the left hand to pass the handbag to the male customer with the right hand supporting it. In Fig. 2.2 (d), the shopkeeper retrieved the handbag by grasping it with the left



Figure 2.3: An example of how a human shopkeeper returns to support an object with the idle hand.

hand while the other hand supported it. In Fig. 2.2 (e), the shopkeeper placed the handbag on the table while keeping the right hand under it.

One representative example of the support hand behavior is shown in Fig. 2.3. In Fig. 2.3 (a), a female shopkeeper held one handbag with the right hand supporting it. When a male customer inquired about the difference between the handbag she was holding and another on the table, she then used her right hand to pick up the second handbag for comparison. Afterwards, she placed the second handbag on the table as shown in Fig. ?? (b)–(d). Interestingly, after this brief task, her right hand returned to its support hand position beneath the first handbag, as shown in Fig. 2.3 (e). This behavior may be surprising because, even after shifting focus to another task, the shopkeeper returns to the initial support hand position, highlighting the importance of this motion.

2.3.2 Analysis of Support Hand Behavior

To quantify how the idle hand behaves, we examined the hand behaviors for the idle hand and classified them into four patterns: supporting an object (supporting), other idle hand motions such as putting the arm down and hanging the arm (others), various gestures such as pointing somewhere on the object (gestures), and different manipulation types such as picking and placing another handbag (manipulations). Tab. 2.1 shows the frequency of these patterns in different situations.

From this table, it is clear that the support behavior is by far the most commonly used behavior in all situations compared to other idle hand motions. Therefore, we chose to focus on the support behavior for the idle hand.

Table 2.1: Observed hand behaviors

Primary Hand	Secondary Hand			
	Idle State		Busy State	
	supporting	others	gestures	manipulations
picking	6	1	0	0
showing	16	3	9	11
giving	4	0	0	0
receiving	5	0	0	0
placing	6	2	0	1
total	37	6	9	12

2.3.3 Reasons for Support Hand

After the role-plays, we conducted interviews with the participants to explore the underlying motivations for their frequent use of the support hand behavior while role-playing as shopkeepers. These interviews revealed three primary reasons for this behavior:

- To show politeness to customers, since being polite is an expected attitude and even a requirement for shopkeepers.
- To prevent the products from accidentally falling as it would make customers feel uncomfortable. Furthermore, this behavior would make customers appreciate the value, fragility, and other attributes of the product, which cannot be conveyed by casual handling such as one-handed behaviors.
- To draw customers’ attention to the item being showcased.

2.3.4 Modeling

From the observation of shopkeepers’ behaviors and subsequent interviews with the participants, we found that the support hand behavior is the main one performed by the idle hand, hence we developed a model for this behavior consisting of computing a support point and positioning the support hand.

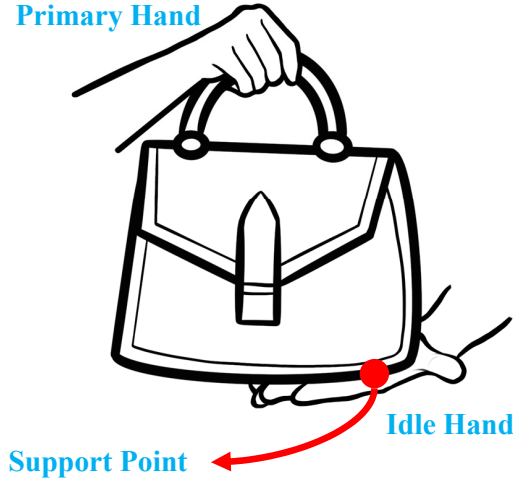


Figure 2.4: Schematic diagram of support hand behavior.

Computing Support Point

The support point serves as a crucial component in our model. As illustrated in Fig. 2.4, the primary hand holds the object while the idle hand supports it beneath the “support point.” Here, the support point is essential to the model of idle hand motion, and the idle hand basically moves under the support point. While for this study we manually selected a point at the bottom of the held object to be the support point, let us explain how the support point could be computed.

In theory, this point can be computed through physical simulations that take into account geometric shape, weight distribution, and a set of specific criteria defining what makes a support point optimal in the given context (e.g., most stable, most visually human-like). The steps for the computation are as follows. First, calculate the center of mass of the object using the available information on the geometric shape and weight distribution of the object. Second, sample candidate points along the edge of the object that are located below the center of mass. Third, assess and rank these candidate points based on the provided criteria. Finally, choose the optimal point that best meets the defined criteria. The point we manually chose satisfies the idea included in these computation steps.

Positioning Support Hand

Upon determining the optimal support point, the next step is to accurately position the idle hand. This involves not only computing the position but also the orientation of the idle hand. In observations of the human shopkeeper, the idle hand was frequently situated just below and in direct contact with the handbag, with the palm facing upwards and fingers pointing sideways and forward, to naturally support the object being held by the primary hand as illustrated in Fig. 2.4.

However, the support hand behavior does not always imply direct contact with the object. Physical support is only necessary when the object is too heavy to be effectively handled by one hand. In the case of lighter objects, it is sufficient to place the support hand below, but not touching, the object. Placing the support hand in such a way accomplishes the goal of emphasizing the object and signaling politeness, while also making it possible for the primary hand to showcase the object by rotating it and showing different angles. For instance, human shopkeepers showcasing a handbag might rotate it repeatedly to display its shape and finer details while maintaining a slight gap between the object and the support hand. Thus, the robot should emulate this behavior.

2.4 Implementation

We implemented an autonomous robot system that reproduces the support hand behavior observed in Section 2.3.

2.4.1 Robot Hardware

We used the TIAGo++ robot * from PAL Robotics. The robot features a mobile base, a lifting torso, a head, and two arms. Both arms have 7 degrees of freedom (DoF) ending in grippers, and the torso has a stroke of 35 cm so that the height of the robot can be adjusted between 110 and 145 cm. Its eyes are equipped with an RGB-D camera, and there are speakers inside the base.

*<https://blog.pal-robotics.com/tiago-bi-manual-robot-research/>

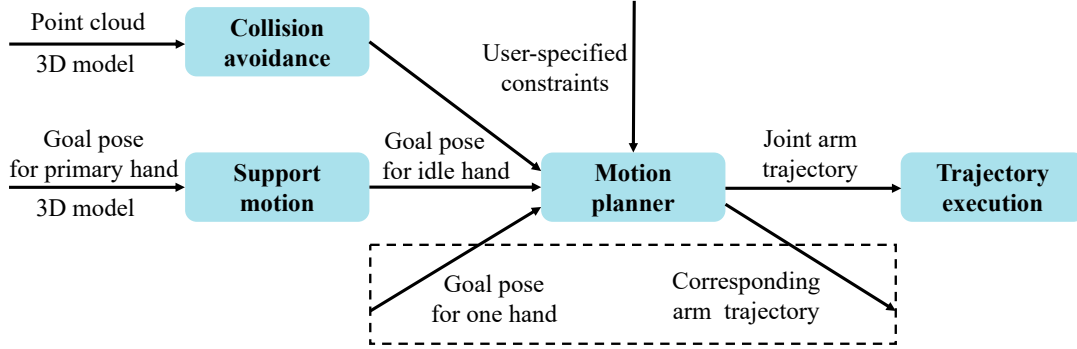


Figure 2.5: Diagram of implementation of proposed model.

2.4.2 Architecture

Fig. 2.5 illustrates the architecture of the proposed model. The point cloud from the RGB-D camera is passed to the collision avoidance module, which represents the operating environment for the robot. Meanwhile, to address the potential occlusion of parts of the object by the robot's gripper, a 3D model of the object is also fed into this module. The support motion module takes the goal pose for the primary hand and the 3D model of the object being held as input, and it outputs the joint goal poses for both hands. To generate an optimal joint arm trajectory for both hands, the motion planner module takes the goal poses of both hands as input, subject to meeting the user-specified constraints and avoiding collisions. When the idle hand is engaged in specific tasks, its goal pose is fed to the motion planner module, which generates an optimal arm trajectory as shown in the dotted box in this figure.

2.4.3 Collision Avoidance

To generate hand motions without collision, we need to build a 3D representation of the environment around the robot. This is achieved by the collision avoidance module, which uses point cloud data from the RGB-D camera to create a 3D occupancy grid map called Octomap. This Octomap provides information on where obstacles are located in the robot's environment, and it is directly passed to the flexible collision library (FCL) for collision checking.

2.4.4 Support Motion

The support motion module generates the goal pose for the idle hand. As described in Section 2.3.4, human shopkeepers leave a slight distance between the bottom of the object and the idle hand. Therefore, we decided to keep such a distance. We empirically adjusted the distance between the support hand and the object to 10 centimeters, aiming to reproduce this distance impression as well as the stability and safety of the arm motion (too close a distance sometimes causes planning failure, as the carried object becomes an obstacle for the planner).

Meanwhile, we found that the palm of the idle hand is parallel to the ground from our observations of how human shopkeepers used the support hand behavior. Consequently, the roll angle was set to 90 degrees, which makes a wider surface of the gripper to support the object, the pitch and yaw angles were set to zero and 45 degrees respectively.

2.4.5 Motion Planner

Given the goal poses of both hands, the motion planner module computes the joint arm trajectories, i.e., time-series of all joints of both arms. Here, we assume that the goal pose is within short straight lines from the current poses. In cases where longer distance travel is needed, a series of sub-goals for both arms need to be pre-computed and provided to this module to ensure synchrony between both arms. Thus, we use the support motion module in conjunction with the motion planner module to plan trajectories from one sub-goal to the next.

We select the RRT-Connect motion planner [52] for our system because of its ability to quickly find a feasible arm trajectory between the start pose and the goal pose, and its suitability for user-specified constraints. The motion planner module takes the goal poses for both hands as input, and outputs a synchronized joint arm trajectory, enabling both arms to reach their corresponding pose at the same time.

However, due to the characteristics of RRT-Connect, the generated arm trajectory is often not optimal, as it is excessively long and visually strange to humans. To counteract this, we sample many candidate arm trajectories, and select the one with the shortest length as optimal, since a good arm trajectory should avoid long detours between the start pose and the goal pose. The length of the trajectory can be summed by the coordinates of both grippers computed by forward

Algorithm 1 Generation of optimal joint arm trajectory

Input: start pose $P_{\text{start_p}}$, start pose $P_{\text{start_s}}$, goal pose $P_{\text{goal_p}}$, object O being held, number of candidate trajectories n , user-specified constraints C_{user}

Output: optimal joint arm trajectory T^*

1. Compute Goal Pose for Support Hand:

$P_{\text{goal_s}} \leftarrow \text{SupportHandModel}(P_{\text{goal_p}}, O)$

2. Generate Candidate Arm trajectories:

for $i = 1$ **to** n

$t_i \leftarrow \text{MotionPlan}(P_{\text{goal_p}}, P_{\text{goal_s}}, C_{\text{user}})$

end for

List candidate joint arm trajectories $T = \{t_1, t_2, \dots, t_n\}$

3. Select Optimal Arm Trajectory:

$T^* = \arg \min_{t \in T} L(t) \quad \text{subject to} \quad C_{\text{user}}(t)$

kinematics and basic transformation for points in the planned arm trajectories. Moreover, to make the trajectories natural and human-like, we set user-specified constraints on joint angles that conform to human shoulder and elbow positions.

The entire process of deriving the optimal joint arm trajectory for the support hand behavior is encapsulated in Algorithm 1, where $L(t)$ represents the length of the joint arm trajectory for both arms.

2.4.6 Trajectory Execution

Our robot system is based on the MoveIt framework[†]. The trajectory execution module is responsible for taking a planned trajectory and executing it with the physical robot automatically.

2.5 Evaluation

2.5.1 Hypotheses

Drawing from the insights in Section 2.2.4, handling an object with two hands is generally considered more polite than using just one. Supporting this, Tagai et al. [91] found in their fMRI study that observing an action involving two hands was

[†]<https://moveit.ros.org/>

associated with higher politeness compared to the use of one hand. Additionally, our observations of polite and competent shopkeepers revealed that they often use the support hand behavior when picking, placing, showing, and passing items to customers or receiving them back, and thus we made the following two hypotheses:

- **H1:** Participants perceive robots with the support hand behavior as **more polite** than those without it
- **H2:** Participants perceive robots with the support hand behavior as **less rude** than those without it.

Meanwhile, Cuijpers and Knops [17] reported that robots performing meaningful motions are seen as more socially competent and skilled than those engaging in idle or no motion. In our study, we regard the support hand behavior as one such meaningful motion. Considering that demonstrating politeness is a crucial aspect of a shopkeeper’s role, we expect robots behaving in a polite manner to be perceived as competent in their role as a shopkeeper. We thus developed the following hypothesis:

- **H3:** Participants perceive robots with the support hand behavior as **more competent** than those without it.

2.5.2 Participants

We recruited a total of 20 participants from a part-time job recruiting website in the age range of 18 to 60 years ($M = 35.750$, $SD = 14.790$). Ten of them self-identified as male, and ten as female, with a similar number for different age groups. All participants who took part in the experiment were compensated with 3000 JPY.

2.5.3 Conditions

The robot’s performance was compared under the following two conditions:

- **Proposed:** The robot runs in the support hand mode as described in Section 2.4.
- **Baseline:** The support motion module is excluded from the architecture of the proposed model. In this case, the goal pose of either hand is directly

fed to the motion planner module. It then outputs the arm trajectory for the corresponding hand, as shown in the dotted box in Fig. 2.4.

We chose a within-subjects design to allow participants to experience different conditions since even in Japan, where many robots are being deployed in public spaces, people remain unfamiliar with robots and lack consensus on the standards for politeness in their behaviors. Meanwhile, the order of conditions was counter-balanced.

2.5.4 Procedure

When participants arrived at the experiment site, they were first introduced briefly to the study and procedure. They were then asked to sign a consent form. To make participants comfortable around the robot, we showed several demos of speech and hand movements by the robot, and then we asked them to control the robot through the remote controller. Afterwards, each participant was instructed to play the role of a customer who showed interest in one of the handbags. The robot shopkeeper would show the handbag with synchronized speech and pass it to and from the customer. In response, the customer needed to receive the handbag from the customer and then give it back to the shopkeeper after trying it on. After each condition, the participant was asked to complete a questionnaire. Finally, after interacting with the robot in both conditions, a semi-structured interview was conducted. The experiment was approved as ethical by the Institutional Review Board. The experiments and interviews were conducted in Japanese.

2.5.5 Measurement

Dependent variables of politeness, rudeness, and competence were measured through a questionnaire that is composed of several 1-to-7 point Likert items.

• Politeness and Rudeness

We consulted the latest review on politeness, compiled by Ribino [78], but did not find any validated scales or commonly-used questionnaires for measuring politeness and rudeness. Therefore, we decided to adopt the items used in Salem et al. [82, 83], to assess politeness with the item “polite” and rudeness with the item “rude.”

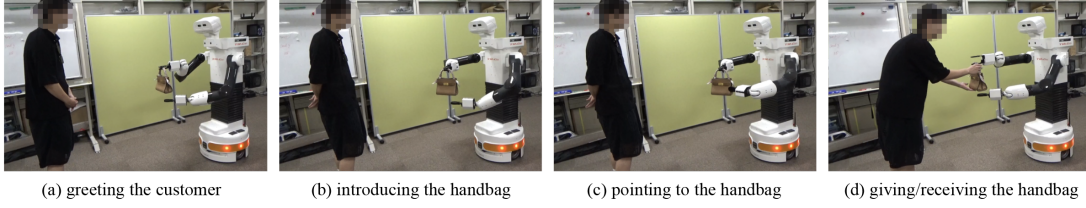


Figure 2.6: Robot in proposed condition.

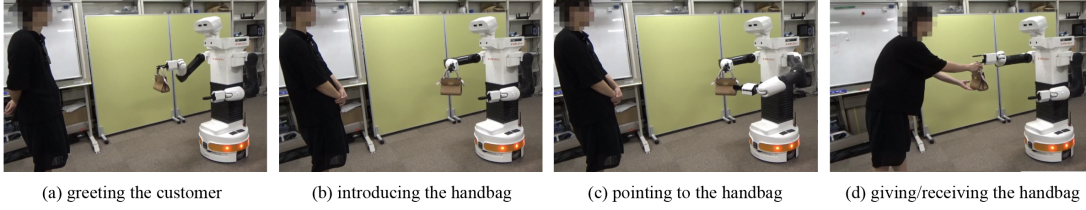


Figure 2.7: Robot in baseline condition.

- **Competence**

We adopted all six items for measuring competence from the RoSAS, a validated scale frequently used in the HRI community [12]. These items are “capable,” “responsive,” “interactive,” “reliable,” “competent,” and “knowledgeable.”

2.6 Results

2.6.1 Observation

Fig. 2.6 and 2.7 show typical interactions in the proposed condition and the baseline condition respectively. The robot first greets the customer and then introduces the handbag while interleaving the pointing gesture. It gives the handbag to the customer and receives it back after the customer examines the handbag. Participants generally behaved with the robot as they would behave with human shopkeepers, nodding to the robot’s speech. This behavior was consistent between conditions.

Interestingly, we did observe some social behaviors toward the robot. In both conditions, many participants nodded to the robot when it greeted them. Moreover, a large proportion of participants who rated the robot with the proposed model as more polite tended to bow to the robot before or after they passed

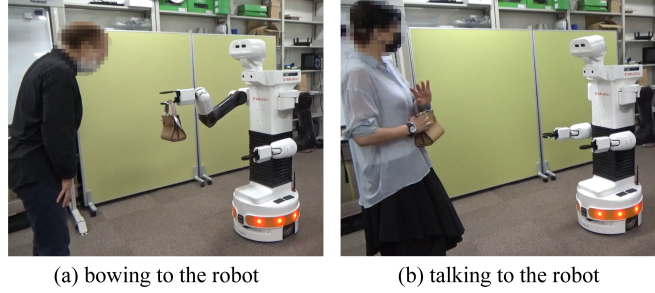


Figure 2.8: Social behaviors toward the robot.

the handbag to the robot (Fig. 2.8 (a)). In addition, one participant attempted to talk to the robot even though she knew it could not respond (Fig. 2.8 (b)).

2.6.2 Verification of Hypotheses

Politeness and Rudeness

First, we conducted the Shapiro-Wilk test to assess the normality of the data. For the item “polite,” both the baseline condition ($W = .810$, $p = .001$) and the proposed condition ($W = .708$, $p < .001$) showed significant departures from normality. Similarly, for the item “rude,” both the baseline condition ($W = .642$, $p < .001$) and the proposed condition ($W = .495$, $p < .001$) were also significantly non-normal. Such non-normality breaks the assumption for using the t-test. Therefore, we decided to use Wilcoxon’s rank sum test (also known as the Mann-Whitney U test), a non-parametric test with no strict requirement for data distribution and variance.

As shown in the first set of bars in Fig. 2.9, there is a significant difference between the proposed model ($M = 6.450$, $SD = 0.759$) and the baseline model ($M = 5.200$, $SD = 1.609$) for the item “polite” ($U = 96$, $p = .002$). This result supports our first hypothesis: A robot using the proposed model is perceived as more polite than the robot using the baseline model.

In the second set of bars, the difference between the proposed model ($M = 1.200$, $SD = 0.410$) and the baseline model ($M = 1.850$, $SD = 1.461$) for the item “rude” was not statistically significant ($U = 246$, $p = .063$). Although this result does not support our second hypothesis, it suggests its approaching significance.

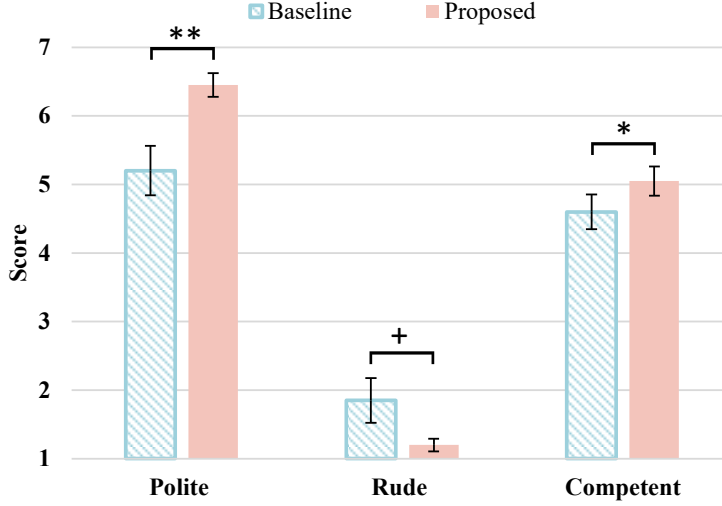


Figure 2.9: Results. The error bars show the standard error for the mean at $\pm 1 SE$. (+ : $.05 < p < .1$, * : $p < .05$, ** : $p < .01$)

Competence

To obtain the data for the item “competence,” we averaged the above-mentioned six items for the scale from the RoSAS. As done above, we ran the Shapiro-Wilk test on the data of the item “competence.” The test results indicate that the data of both the baseline condition ($W = .976$, $p = .865$) and the proposed condition ($W = .972$, $p = .792$) followed a normal distribution. Consequently, a paired t-test was used to evaluate the statistical significance of the differences in competence between conditions.

As illustrated in the third set of bars in Fig. 2.9, the difference between the proposed model ($M = 5.050$, $SD = 0.949$) and the baseline model ($M = 4.600$, $SD = 1.140$) was significant (paired t-test $t = 2.193$, $p = .020$), and the effect size was medium (Cohen’s $d = 0.490$). This result supports our third hypothesis: A robot using the proposed model is regarded as more competent than a robot using the baseline model.

2.6.3 Interview Results

When asked about the reason for their judgments on politeness, nine participants attributed their positive judgments to the robot’s support hand behavior, noting its resemblance to human shopkeepers’ respectful service with both hands. Two

participants emphasized that the robot’s careful handling of products, as indicated by its support hand behavior, suggested a dedication to providing superior service. Another two recognized differences in robot behavior across conditions but perceived equal levels of politeness in both. Six participants reported that they did not find differences between the conditions. Interestingly, one female participant perceived the robot exhibiting the support hand behavior as less polite, interpreting this as the robot exerting sales pressure, which made her uncomfortable.

When it came to the robot’s competence, three participants reported that a robot demonstrating great service etiquette implied greater competence. Two participants observed that the support hand behavior mirrored that of human shopkeepers, enhancing perceptions of the robot’s competence. Two others linked the robot’s meticulous product handling to competence. Additionally, one participant remarked that the robot’s enthusiasm in presenting the product was indicative of competence. The remaining participants based their assessments on subjective feelings, without providing specific reasons for their viewpoints.

2.7 Discussion

2.7.1 Design Implications

As robots become integral to our daily lives, understanding the sociological aspects of their interactions, especially concerning object manipulation, gains paramount importance. One facet of this is the concept of ‘legibility’ of robot motions. Recent studies [10, 22] focus on developing robot motions with clear intentions to enhance their legibility. This principle emphasizes designing robot motions that humans can quickly and easily interpret, fostering smoother and more predictable interactions. Drawing parallels to the research on feeding support [69], we see a similar emphasis on the sociological dimensions. Robot feeding behavior, as outlined in their work, must be subtle and not draw undue attention, suggesting that robots should not only perform tasks but also align with the unspoken social codes governing those tasks.

The support hand behavior serves as an illustrative example of this merging between the physical and the sociological, offering insights into designing more socially-aware robots.

2.7.2 Should Robots Exhibit Support Hand Behavior?

It is arguable whether robot shopkeepers need to exhibit the support hand behavior. While operating both hands consumes more energy than one, potentially requiring more frequent recharging of the battery-operated robots, we believe that stores would prefer to use the support hand behavior simply to give an impression of politeness or competence. If retail stores’ target shoppers value a more “human-like” shopping experience, robots with the support hand behavior might be considered justified. Furthermore, given labor shortages, robots will very likely be shopkeepers in the future. Such robots are expected to provide interactions that are not just transactional but also socially-appropriate or socially desirable. Expressing politeness and competence through the support hand behavior can be a vital component of this development, greatly contributing to making the robots more acceptable to users.

2.7.3 Which Occasions Should Robots Exhibit Support Hand Behavior?

The use of the support hand behavior by robots should align with cultural norms, customer expectations, and specific service contexts. In luxury retail, this behavior not only underscores the item’s value but enhances customer engagement. Similarly, in cultural rituals like tea ceremonies, the dual-handed presentation is synonymous with respect, warranting robots to emulate this to uphold the ceremony’s essence. In hierarchical settings, using both hands to present items conveys deference, emphasizing the importance of recognizing social norms in robot behaviors. Conversely, in casual contexts like convenience stores, where efficiency is prioritized, such formalities may be superfluous. Finally, the support hand behavior can be used for small, lightweight objects that do not require two hands to hold. For large, heavy objects a two-handed grip is required, which is beyond the scope of our study.

2.7.4 Generalizability across Cultures

It remains an open question whether the support hand behavior can be applied to other cultures. However, our study provides preliminary insights and serves as a starting point for an extensive investigation into the use of idle hand motions

in different cultures.

While this research was conducted in Japan, we believe that the support hand behavior shows potential for use in different cultural contexts. The use of both hands for handling objects signifies politeness, and this practice is prevalent in East Asian countries such as China and Korea [91]. Additionally, we asked several members of our lab with international backgrounds about their impressions of the support hand behavior. They responded that they had also observed the support hand behavior overseas (including North America and Europe). They felt discomfort when observing shopkeepers handle valuable items with a single hand, but did not mind in the case of inexpensive items.

2.7.5 Limitations

While revealing the importance of the support hand behavior, there are still several limitations to this work. First, we chose only handbags in the observation and evaluation; it is uncertain whether people would rate the robot differently with other objects regarding politeness and competence. Second, as the observation of shopkeepers’ hand behavior and the experiment were conducted in a laboratory setting, the situation would be a bit different in a real-world environment. However, we believe our experiment design effectively captured the important factors of hand motions in object-based interactions, and the setup allowed us to focus on the hand motions and not other factors such as verbal interaction.

2.8 Conclusion

This study was motivated by the phenomenon in which robots keeping one hand fully idle during object manipulation gives the impression of weirdness, particularly in roles demanding social cues such as that of shopkeepers. Accordingly, we first observed how human shopkeepers served customers with both hands, identifying the supportive role of the idle hand during object-based interactions. We then developed a computational model to replicate the support behavior for the idle hand of robot shopkeepers. The experimental results indicate that robots implemented with support behaviors for the idle hand are perceived as markedly more polite and competent than those without such behaviors. We believe that the incorporation of the support hand behavior can drastically enhance social

acceptance of robots, especially in roles like shopkeepers.

CHAPTER 3

COMMUNICATING PHYSICAL PROPERTIES THROUGH ROBOT OBJECT MANIPULATION

3.1 Introduction

Effective non-verbal communication between robots and humans is crucial in dynamic and unpredictable environments. Imagine a scenario in a noisy restaurant kitchen where a service robot needs to hand over a freshly prepared dish to a waiter. The plate is extremely hot, and while a verbal warning could suffice in a quiet setting, the loud clamor of the kitchen makes it difficult for the waiter to hear. Furthermore, the heat of the plate cannot be reliably perceived by simply observing its appearance. In such situations, a robot capable of conveying *hot* through intuitive hand motions—such as quickly fanning the air around the object (Fig. 3.1 (a)), rapidly shaking the hand (Fig. 3.1 (b)), or frequently switching the object between hands (Fig. 3.1 (c))—could effectively communicate the potential danger without relying on verbal cues.

Similarly, imagine a robot assisting with the curation and transport of delicate artifacts in a museum. If the robot encounters a biological specimen that emits a strong and unpleasant odor, it needs to alert the staff discreetly as speaking

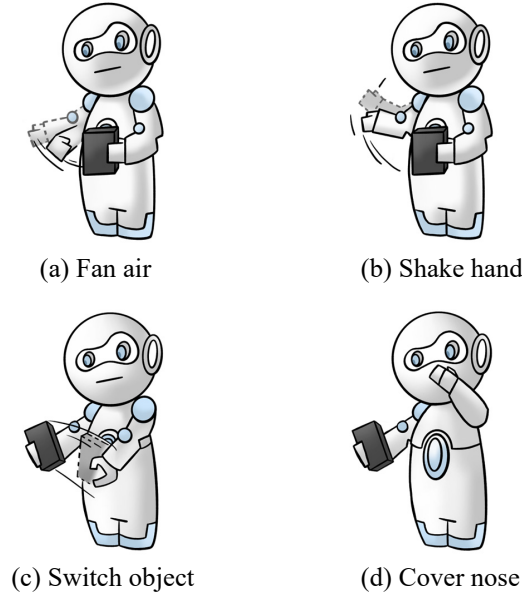


Figure 3.1: Communication of object properties during handovers.

aloud is not appropriate in such a quiet and respectful environment. In this case, covering its nose with one hand while carefully holding the artifact with the other (Fig. 3.1 (d)) allows the robot to signal *smelly* to the nearby staff without disturbing visitors. These examples highlight the importance of non-verbal communication of object properties, especially when verbal communication is limited.

Extensive research has focused on using robot gestures for non-verbal communication, covering intentions [28, 31], emotions [56, 58], and personality [5, 47]. However, the communication of object properties, particularly through robot manipulation, remains significantly underexplored. To bridge this gap, our study makes three main contributions. First, we identify eight categories of object properties that are commonly communicated during handover interactions. Second, we propose a property-based motion planner capable of generating optimal hand action sequences to convey these properties. Third, we demonstrate the feasibility of our approach through human studies, showing that people can accurately infer object properties from robot object manipulation.

3.2 Related Work

3.2.1 Gestures for Communications

Gestures play a crucial role in human communication by enhancing engagement, clarifying speech, and supporting cognitive processes such as learning and memory [21, 26, 44, 64]. Beyond their role in human interaction, gestures are also critical for non-verbal communication between robots and humans. In human-robot interaction, robot gestures have been shown to improve the quality of interactions during narrations [32]. Additionally, robots use the support hand gesture to enhance perceptions of politeness and competence, particularly in service roles [72].

The existing literature offers substantial insights into how gestures facilitate communication. However, there is limited research on conveying object properties through robot gestures, particularly via robot manipulation. This gap motivates our investigation into leveraging robot object manipulation to effectively communicate a broader range of object properties.

3.2.2 Legibility

Legibility, i.e., the intuitive understanding of robots’ intentions, is crucial for coordinating joint action and positive perception [59]. Numerous models and methods have been developed to enhance the legibility of arm motions [22, 23], ranging from gestures [28, 31] to handovers [63, 87]. Additionally, a viewpoint-based model was proposed to generate legible motions [70], and a method for synthesizing legible manipulation motions was developed to effectively communicate a robot’s intent to human collaborators [10]. Legible motions have also been extensively applied in character animation, where animation techniques have been employed to express forethought, enhancing the legibility of robot behaviors and shaping human perceptions [92].

While prior research has primarily focused on enhancing legibility through direct robot motions to signal intent, our study shifts the focus to communicating object properties through robot object manipulation, distinguishing it from existing work.

3.2.3 Communication of Object Properties

Research has demonstrated that people can infer object properties by observing human behaviors. For instance, children can accurately judge an object’s weight based on how others lift and transport it [40]. Similarly, individuals can deduce properties such as slipperiness and temperature from the way objects are handled [84].

This ability to infer object properties also applies to robot gestures and manipulations. Robots can use iconic gestures to convey the shape and size of objects [81]. Lifting behaviors have been employed to effectively communicate an object’s weight [84], while teleoperated robots can convey weight through visuo-proprioceptive cues, where the robot’s movements differ from those of the operator [74]. Furthermore, robots have been designed to express carefulness when handling full or empty cups, allowing observers to infer their contents [55].

While these studies focus on communicating specific object properties through robot gestures or manipulations, the communication of a broader range of object properties by robots remains underexplored.

3.3 How Do People Communicate Properties through Object Manipulation?

To investigate how people communicate object properties through manipulation, we selected specific object properties for this study. Subsequently, we recruited actors to convey these properties by their hand behaviors. From their role-plays, we extracted fundamental yet expressive hand behaviors, which we termed *key elements*.

3.3.1 Selection of Object Properties

To identify object properties commonly conveyed during interactions, we adopted two related sets of words and applied criteria for the selection. The selected properties were subsequently grouped based on their similarities.

We began by reviewing the literature on cognitive science, taxonomy, and materials, adopting a widely recognized set of semantic features for conceptual representation of objects [65]. From this set, we extracted 220 unique adjectives,

as adjectives are directly relevant for describing object properties. In addition, recognizing the critical role of tactile properties in object handovers, we incorporated a list of 262 adjectives specifically designed to capture the nuances of tactile sensations [29]. Combining these two sets yielded a total of 342 distinct adjectives.

To select suitable object properties for this study, we established the following four key criteria:

- *Commonly communicated in handover contexts.* Knowing object properties in advance helps recipients avoid potential hazards, discomfort, or object damage during handovers.
- *Not directly audibly or visibly perceivable.* Properties that are easily perceived through hearing or sight do not require communication.
- *Descriptive adjectives for non-living objects.* While living beings possess properties, they are rarely manipulated during handovers. Therefore, we focused on non-living objects that are typically manipulable.
- *Associated with physical characteristics.* Properties related to functionality or emotions are often subjective. Hence, we prioritized physical ones that can be directly perceived during object interactions.

By applying these criteria, we identified 39 object properties and clustered them into eight categories based on their similar characteristics: *hot*, *cold*, *heavy*, *light*, *slippery*, *sticky*, *fragile*, and *smelly*, as shown in Tab. 3.1.

3.3.2 Observations of Human Hand Behaviors

To investigate how people communicate object properties through hand behaviors during handover interactions, we recruited two Japanese actors (male, 61; female, 48). Both actors have over 10 years of professional acting experience. Neither had a background in robotics nor had they previously participated in robot-related experiments.

Procedures

Upon arrival, the actors were provided with a brief overview of the study and asked to sign a consent form. To familiarize them with the robot’s capabilities,

Table 3.1: Eight categories of object properties

Hot	Cold	Heavy	Light	Slippery	Sticky	Fragile	Smelly
burning	cold	dense	light	damp	gelatinous	breakable	musty
hot	cool	heavy	lightweight	drenched	gooey	brittle	smelly
overheated	freezing			greasy	goopy	crumbly	strong-smelling
scalding	frigid			moist	gummy	delicate	
scorching	frosty			oily	icky	fragile	
warm	icy			slippery	slimy		
				wet	sticky		
					viscous		

we demonstrated basic hand motions, including actions such as picking up an object from a table with one or both hands, switching an object between hands, and rotating an object.

Afterward, we instructed the actors to interact with a generic box in a manner that would communicate the eight specified object properties solely through hand behaviors. They were tasked with emulating the dual-armed robot within five different assigned time limits, beginning before picking up the object and ending with moving it to the goal position. For each time limit, the actors performed two trials. After each trial, they were asked to provide brief explanations of the behaviors they performed.

Identification of Human Key Elements

Key elements play a crucial role in helping people recognize object properties. To extract such elements from actors’ role-plays, the first step was to define a resolution. What constitutes a good resolution? We propose that it should meet two criteria: (1) it should be minimal, avoiding unnecessary complexity, and (2) it should enable people to associate actions with a set of candidate object properties. For instance, the action of briefly touching an object and quickly withdrawing the hands did not individually convey *hot*; however, their combination evoked an association with the property and was thus regarded as a key element.

Applying this resolution, we analyzed video recordings of actors interacting

3. Communicating Physical Properties Through Robot Object Manipulation

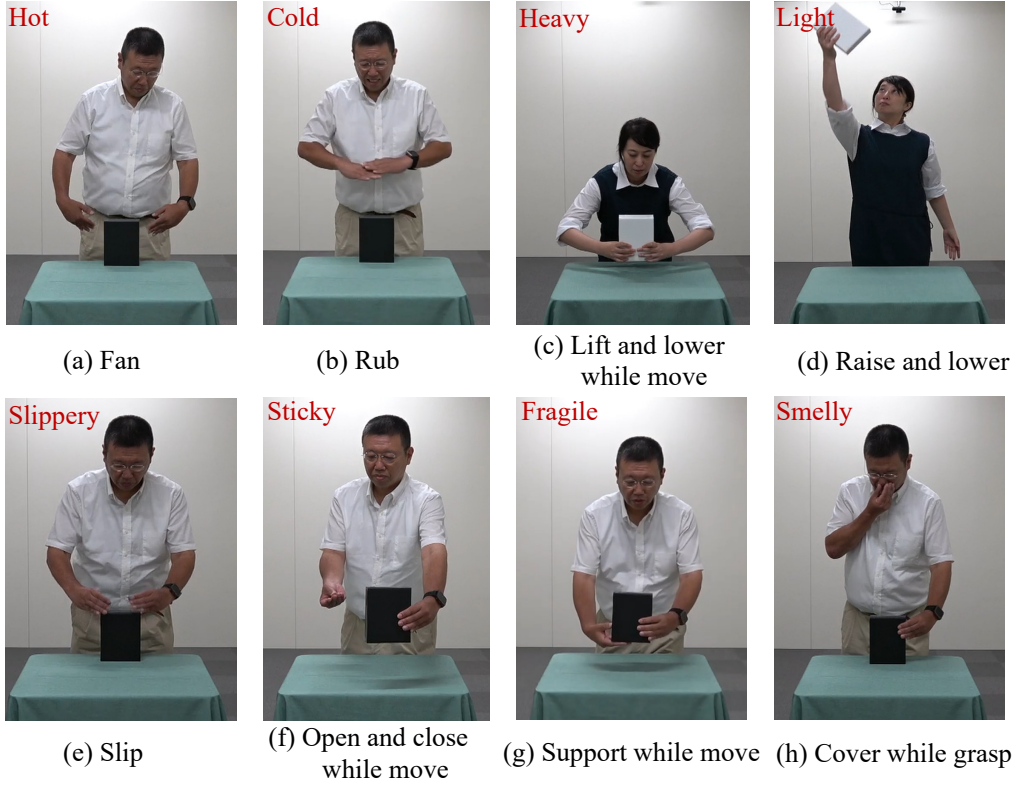


Figure 3.2: Examples of human key elements for communicating object properties.

with objects and identified a set of key elements, integrating the actors' explanations. These elements are listed in Tab. 3.2, with examples corresponding to specific object properties illustrated in Fig. 3.2.

Finally, we divided these key elements into four sequential interaction stages:

- *Prepare*: Actions performed before an object is successfully grasped.
- *Grasp*: Actions performed during the successful grasping of an object.
- *Handle*: Actions performed while holding an object, but before moving it towards the intended goal position.
- *Transport*: Actions performed while moving an object towards the intended goal position.

3. Communicating Physical Properties Through Robot Object Manipulation

Table 3.2: Key elements for object properties

Properties	Elements	Stages	Descriptions	Utility
Default	grasp from middle	grasp	grasp an object from the middle with both hands	0
	move	transport	move an object to the goal with both hands	0
Hot	touch and withdraw	prepare	briefly touch an object, then quickly withdraw both hands	1.85; 2.1; 1.45; 1.1
	fan	prepare	fan away the air around an object with both hands	0.6; 1.1; 0.7; 0.65
	shake	prepare	shake with two hands	-0.1; 0.25; 0; 0.05
	wipe	prepare	wipe the sweat from the forehead with one hand	-0.1; -0.15; -0.45; -0.7
	grasp from top	grasp	grasp an object from the top with both hands	0.3
	fan while grasp	grasp	fan away the air around an object with one hand while grasping it with the other	-0.1
	shake while hold	handle	quickly shake with one hand while holding an object with the other	0; -0.2; -0.15; -0.4
	switch while move	transport	quickly switch an object between both hands while moving it to the goal	1.7
Cold	touch and withdraw with pause	prepare	touch an object for seconds, then withdraw both hands	0.7; 1.15; 0.6; 0.15
	rub	prepare	rub with two hands	0.75; 0.95; 0.5; 0.15
	grasp from top	grasp	grasp an object from the top with both hands	0.1
	rub while hold	handle	rub skins with one hand while holding an object with the other	-0.1; -0.05; -0.1; -0.2
	breathe while hold	handle	breathe one hand while holding an object with the other	0.35; 0.45; 0.2; -0.15

3. Communicating Physical Properties Through Robot Object Manipulation

Properties	Elements	Stages	Descriptions	Utility
Heavy	fail	prepare	slightly lift an object from the top with both hands, then immediately place it down	2.5; 2.55; 2.05; 1.95
	grasp from bottom	grasp	lift an object from the bottom by extending both hands	0.1
	jitter while move	transport	jitter with both hands while moving the object to the goal	0.4
	lift and lower while move	transport	lift and lower an object with both hands while moving it to the goal	2.2
Light	grasp with one hand	grasp	grasp an object with one hand	1.3
	raise and lower	handle	raise and lower an object over the head with one hand	2.3; 2.75; 2.45; 2.35
	rotate	handle	rotate an object with one hand	0.95; 0.9; 0.75; 0.6
	raise and lower while rotate	handle	raise and lower an object over the head while rotating it with one hand	1.85; 1.9; 1.5; 1.3
Slippery	slip	prepare	slip along an object's surface with both hands	2.05; 2.55; 2.5; 2.3
	grasp from bottom	grasp	grasp an object from the bottom with both hands	0.1
	slip while hold	handle	slide along an object's surface with one hand while holding it with the other	0.1; 0.2; 0; -0.2
	slip while move	transport	slide along an object's surface with one hand while moving it to the goal	0.15
	support while move	transport	support an object from the bottom while moving it to the goal	-0.35
Sticky	touch and withdraw with pause	prepare	touch an object for seconds, then withdraw both hands	0.1; 0; -0.15; -0.4
	open and close	prepare	quickly open and slowly close both palms	1.85; 2.15; 1.5; 1.05

3. Communicating Physical Properties Through Robot Object Manipulation

Properties	Elements	Stages	Descriptions	Utility
Sticky	grasp from top	grasp	grasp an object from the top with both hands	0.15
	open and close while hold	handle	quickly open and slowly close one palm while holding an object with the other	2.4; 2.3; 1.65; 1.45
	open and grasp while hold	handle	quickly open and slowly grasp an object with one hand while holding it with the other	1.05; 1.2; 0.5; 0.25
	open and close while move	transport	quickly open and slowly close one palm while moving an object to the goal	-0.3
Fragile	grasp from bottom	grasp	grasp an object from the bottom with both hands	0.35
	cradle	handle	slowly cradle an object to the chest with both hands	0.15; -0.4; -0.95; -0.95
	tap	handle	slightly tap an object with one hand while holding it with the other	-1.5; -1.75; -1.8; -1.85
	support while move	transport	support an object from the bottom while moving it to the goal	2.4
	turn over while move	transport	turn an object over while moving it to the goal	0.15
Smelly	fan	prepare	fan away the odor around an object with both hands	-0.1; -0.1; -0.15; -0.2
	fan while cover	prepare	fan away the odor around an object with one hand while covering nose with the other	2.5; 2.8; 2.9; 2.75
	touch and sniff	prepare	touch an object with both hands, then sniff the hands	-0.55; -0.7; -1.05; -1.1
	grasp with one hand	grasp	grasp an object with one hand	0.15
	cover while grasp	grasp	cover nose with one hand while grasping an object with the other	2.6
	fan while grasp	grasp	fan away the odor around an object while grasping it with the other	-0.45

Properties	Elements	Stages	Descriptions	Utility
Smelly	fan while move	transport	fan away the odor around an object with one hand while moving it to the goal	-0.4
	cover while move	transport	cover nose with one hand while moving an object to the goal	2.8

We found that key elements from the *prepare* and *handle* stages were generally repeatable, whereas those from the *grasp* and *transport* stages were not. For example, actors raised and lowered an object repeatedly to indicate *light*, but supported an object from the bottom to the goal position only once to indicate *fragile*.

3.4 Property-based Planner for Communicating Object Properties

This section proposes a property-based motion planner designed to enable robots to communicate object properties. To this end, we began by replicating human key elements on a robot. Subsequently, we conducted a user study to evaluate these robot hand motions (i.e., robot key elements). Finally, we designed and implemented the proposed planner on a robot.

3.4.1 Architecture

The architecture of the property-based motion planner is illustrated in Fig. 3.3. The proposed planner consists of two modules: the *action sequence generator* and the *motion planner*. The action sequence generator takes as input the object property, time ratio, initial pose of the object, and goal pose, along with the information about robot key elements, to generate a sequence of sub-goal poses. These sub-goal poses, combined with user-specified constraints, are then processed by the motion planner module to produce human-like arm trajectories that effectively communicate the specified object property.

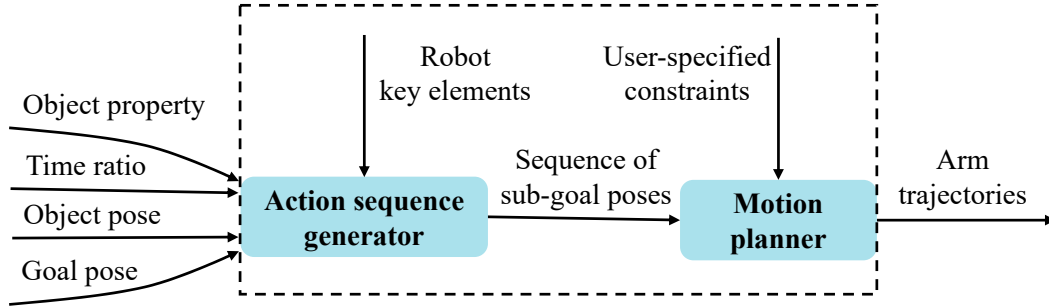


Figure 3.3: The Diagram of proposed planner.

3.4.2 Robot Key Elements

Human key elements are essential for conveying object properties. To evaluate their effectiveness, which we termed as *utility*, when implemented on robots, the first step was to represent the corresponding robot key elements. We identified three critical factors characterizing these robot key elements: *sub-goal poses*, *execution time*, and *utility*.

We selected a dual-arm robot for this study as it can closely mimic human behaviors, facilitating a better understanding of the robot’s actions and intentions. Based on observations of human behaviors and the robot’s physical capabilities, we manually defined sub-goal poses corresponding to all the human key elements in Tab. 3.2. These elements were then implemented on the robot, with examples of robot key elements shown in Fig. 3.4.

3.4.3 Utility

To quantify the utility of robot key elements, we conducted a user study.

Procedures

We first developed multiple robot demonstrations showcasing all the human key elements in Tab. 3.2, with the frequency of each key element varying depending on the stage. Specifically, we created one robot demonstration for the key elements of the *grasp* and *transport* stages. For the *prepare* and *handle* stages, four robot demonstrations were created for each key element, with the frequency of each

3. Communicating Physical Properties Through Robot Object Manipulation

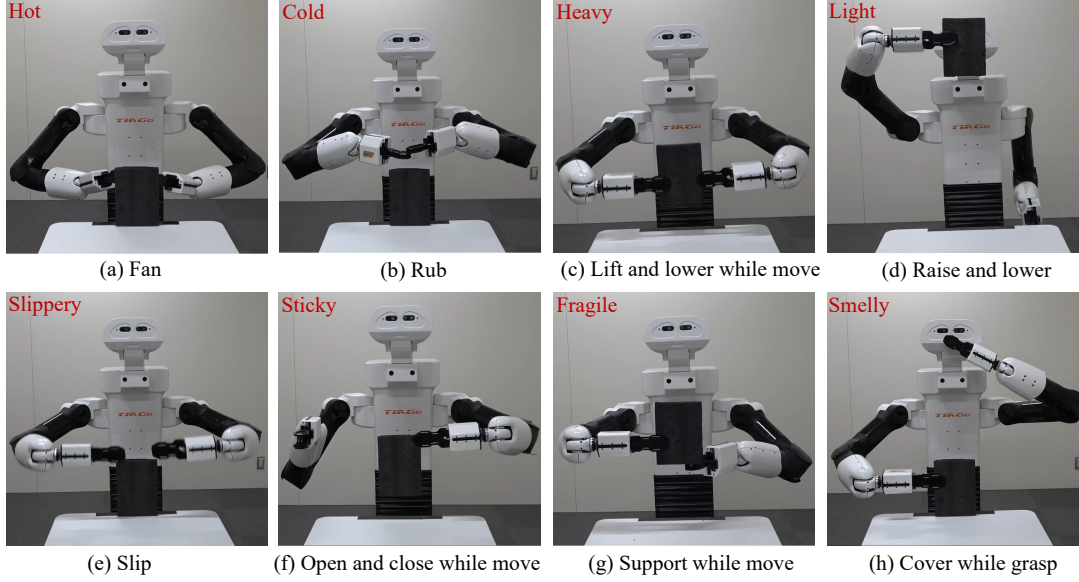


Figure 3.4: Robot key elements corresponding to the human key elements illustrated in Fig. 3.2.

element ranging from one to four. All demonstrations involved grasping an object from a table and moving it to the goal position.

Subsequently, we invited 20 participants (10f, 10m) to evaluate the utility of all the robot key elements for the corresponding object properties. To minimize potential order effects, we employed a partial Latin square design to determine the order in which participants viewed and evaluated the videos. Participants rated each video on a scale from -3 to 3, where a score of 3 indicated that the key element helped to recognize the object property, while a score of -3 indicated that the key element hindered recognition of the object property. A score of 0 represented neutrality, indicating that the key element neither helped nor hindered the recognition. The utility reflects the effectiveness of key elements in conveying object properties, with higher utility indicating greater expressiveness.

Results

We calculated the utility for all the robot key elements and their repetitions by averaging participants' ratings, as shown in the last column of *utility* in Tab. 3.2.

The results revealed that the utility of repeatable key elements generally peaked

at two repetitions when sufficient time was available. In contrast, the utility of unrepeatable ones was generally lower, except for those associated with the properties of *light* and *smelly*. Additionally, there were slight variations in utility on the grasping location (e.g., “grasp from bottom” versus “grasp from top”).

3.4.4 Action Sequence Generator

Our observations of role-plays revealed that the allotted time significantly influences the variety and expressiveness of hand motions used to convey object properties. For instance, when given the shortest time, actors typically grasped the object with one hand and moved it to the goal position to represent *light*. However, with more time, they introduced additional actions (e.g., rotating an object) beyond these two basic actions.

To reproduce these adaptive behaviors, we introduced the concept of *time ratio*. The time ratio refers to a multiple of the time required to grasp the middle of an object and move it from the table to the goal position. A time ratio of one corresponds to the duration of these two default actions.

Building on this concept, we developed an action sequence generator. As shown in Fig. 3.3, this module takes two categories of inputs: (1) object property, time ratio, the object’s initial pose, and goal pose; and (2) information about robot key elements including utility, execution time, and sub-goal poses. It then generates an optimal sequence of sub-goal poses corresponding to the action sequence of robot key elements with the highest utility for the given time ratio. Specifically, depth-first search is first employed to generate all possible action sequences with durations under the time limit dictated by the time ratio. Action sequence durations are calculated with pre-computed key element execution times obtained from tests on the robot. Finally, the action sequence with the highest utility is selected as optimal and mapped to the corresponding sequence of sub-goal poses associated with its key elements.

3.4.5 Motion Planner

Once the sequence of sub-goal poses is determined, the motion planner module computes the corresponding sequence of sub-goal arm trajectories.

3.4.6 Implementation

We implemented the property-based motion planner on a dual-arm robot system.

Robot Hardware

We used a dual-arm robot named TIAGo++^{*} from PAL Robotics. The robot features a mobile base, a lifting torso, a head, and two arms. Both arms have 7 degrees of freedom (DoF) ending in grippers, and the torso has a stroke of 35 cm so that the height of the robot can be adjusted between 110 and 145 cm. Its eyes are equipped with an RGB-D camera, and there are speakers inside the base.

Trajectory Execution

Our robot system is built on the MoveIt framework[†], which automatically takes the arm trajectories as input and executes them on the physical robot.

3.4.7 Examples of Planning Results

We evaluated the functionality of the proposed planner. As shown in Fig. 3.5, at the lowest time ratio (i.e., 1.0), the planner generates an action sequence consisting of two key elements, one from the *grasp* stage and the other from the *transport* stage, for *light*. As the time ratio increases, the planner incorporates additional key elements, such as “raise and lower” or “rotate,” leading to higher utility. The results indicate a positive relationship between utility and time ratio: as the time ratio increases, utility also improves. Similarly, Fig. 3.6 illustrates the planning results for *hot*.

3.5 Evaluation

We designed and conducted an experiment to evaluate whether a robot could effectively communicate object properties by manipulating an object.

^{*}<https://pal-robotics.com/blog/tiago-bi-manual-robot-research/>

[†]<https://moveit.ros.org/>

3. Communicating Physical Properties Through Robot Object Manipulation



Figure 3.5: The planning results for *light* at different time ratios.

3. Communicating Physical Properties Through Robot Object Manipulation



Figure 3.6: The planning results for *hot* at different time ratios.

3.5.1 Participants

We recruited 20 participants through a part-time job recruitment website, ranging in age from 18 to 68 years ($M = 36.6$, $SD = 15.48$). Ten participants self-identified as male and ten as female, with a balanced distribution across age groups. Each participant received 3000 JPY as compensation for their participation in the experiment.

3.5.2 Conditions

The robot’s performance was estimated under five conditions. Specifically, the participants needed to infer object properties under different time ratios (i.e., 1, 1.5, 2, 2.5, and 3). A within-subjects design was employed to allow participants to experience all conditions, as even in countries like Japan, where many robots are being deployed in public spaces, people remain unfamiliar with interpreting robots’ behaviors and intentions. To minimize order effects, a partial Latin square design was used to determine the order of time ratios and object properties within each ratio.

3.5.3 Procedure

Participants were welcomed into a large room where the robot was situated, and the experimenter administered the experiment overview and consent form. To familiarize participants with the robot and to ease any potential discomfort, several demonstrations of the robot’s hand motions were shown. To prevent visual identification of the object relation, a generic box that concealed its property was selected as the experimental prop.

Prior to the experiment, participants were informed that they could attribute any of the eight specified object properties to each robot hand motion, ensuring that each interaction remained independent. Additionally, they were asked to choose the most effective property for each robot demonstration. The experimental setting is depicted in Fig. 3.7. Finally, we had semi-structured interviews with the participants. The experiment was approved as ethical by the Institutional Review Board. The experiments and interviews were conducted in Japanese.

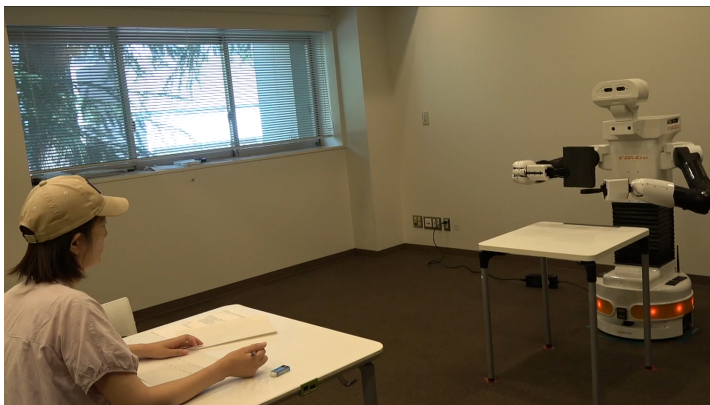


Figure 3.7: Setup for experiments.

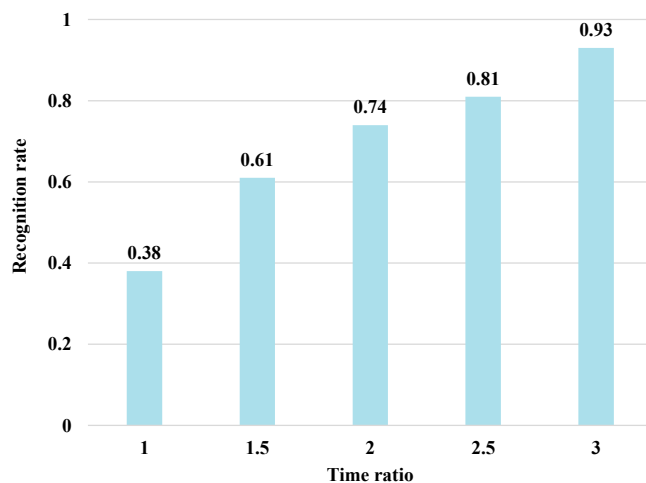


Figure 3.8: Relationship between recognition rate and time ratio.

3.6 Results

3.6.1 Recognition Results

We evaluated the recognition accuracy for each object property at various time ratios, as illustrated in Fig. 3.8. The results indicate a positive relationship between the recognition rate and the time ratio: as the time ratio increases, the recognition rate improves, reaching a peak of 0.93 at the time ratio of 3. This suggests that longer interaction time helps participants accurately identify object properties.

To further explore participants’ responses to action sequences of key elements across different time ratios, confusion matrices were employed and presented in Tab. 3.3, where each column corresponds to a specific robot action sequence, while each row reflects participant ratings. The confusion matrices reveal not only the correct recognition patterns but also the types of misclassifications that occurred, showing how participants perceived and confused certain object properties.

Notably, some object properties achieved high recognition rates early on, particularly those with short but highly expressive key elements, such as “cover nose” for *smelly*, “jitter while move” for *heavy*, “raise and lower” for *light*, and “slip” for *slippery*. In contrast, others such as *hot* and *sticky* required more key elements presented over longer time ratios to reach similar levels of accuracy.

3.6.2 Interview Results

When asked about the ease of inferring properties from the robot manipulating objects, all participants found it straightforward to recognize *smelly*, *heavy*, *light*, and *slippery*, with *smelly* being the easiest to identify. Sixteen participants reported that it was relatively easier to infer *fragile*, as the support hand behaviors conveyed a sense of cautious handling, which aligns with the view in [72].

Seven participants mentioned that while they could infer *hot* from the element of “touch and withdraw,” they still found it challenging to distinguish between *hot* and *cold*. Three participants suggested that rubbing hands could indicate either *hot* or *cold*. Similarly, two participants linked the element of “rub” with *sticky*, as it resembled behaviors used to remove stickiness. Furthermore, four participants associated the element of “open and close” either with both *hot* and *cold*, as it mimicked behaviors for cooling or warming hands. However, they noted that the element of “open and close while hold” was highly expressive and even humorous, as it simulated the robot’s hands being stuck together, helping them confirm their judgments.

Interestingly, two participants expressed that adding more hand behaviors did not necessarily improve the clarity of conveyed properties. For example, they noted that additional behaviors, such as the element of “rotate” for *light*, caused confusion and made them hesitant in their judgments.

3. Communicating Physical Properties Through Robot Object Manipulation

Table 3.3: The recognition results for object properties at different time ratios

	Hot	Cold	Heavy	Light	Slippery	Sticky	Fragile	Smelly
Time Ratio: 1								
Hot	1	2	0	0	0	2	0	0
Cold	0	0	0	0	1	1	0	0
Heavy	1	1	17	0	6	0	2	0
Light	11	12	0	20	9	14	12	0
Slippery	2	2	1	0	1	0	4	0
Sticky	0	1	0	0	0	1	1	0
Fragile	5	1	2	0	2	2	1	0
Smelly	0	0	0	0	1	0	0	20
Time Ratio: 1.5								
Hot	15	0	1	0	0	1	0	0
Cold	1	1	1	0	0	3	0	0
Heavy	0	3	16	0	0	0	2	0
Light	2	14	0	20	2	16	5	0
Slippery	1	1	1	0	18	0	0	0
Sticky	1	1	0	0	0	0	5	0
Fragile	0	0	1	0	0	0	7	0
Smelly	0	0	0	0	0	0	1	20
Time Ratio: 2								
Hot	13	0	0	0	0	14	2	0
Cold	2	11	1	0	0	2	2	0
Heavy	0	2	19	0	0	0	0	0
Light	0	2	0	20	0	0	2	0
Slippery	2	2	0	0	20	0	1	0
Sticky	1	2	0	0	0	2	0	0
Fragile	2	1	0	0	0	2	13	0
Smelly	0	0	0	0	0	0	0	20

3. Communicating Physical Properties Through Robot Object Manipulation

	Hot	Cold	Heavy	Light	Slippery	Sticky	Fragile	Smelly
Time Ratio: 2.5								
Hot	14	3	0	0	0	7	0	0
Cold	3	12	0	0	1	4	1	0
Heavy	0	0	20	0	0	0	0	0
Light	0	0	0	20	0	1	0	0
Slippery	0	0	0	0	19	1	1	0
Sticky	3	5	0	0	0	7	0	0
Fragile	0	0	0	0	0	0	18	0
Smelly	0	0	0	0	0	0	0	20
Time Ratio: 3								
Hot	19	1	0	0	1	1	0	0
Cold	0	15	0	0	0	2	0	0
Heavy	0	0	20	0	0	0	0	0
Light	0	0	0	20	0	0	1	0
Slippery	0	0	0	0	19	0	1	0
Sticky	1	4	0	0	0	17	0	0
Fragile	0	0	0	0	0	0	18	0
Smelly	0	0	0	0	0	0	0	20

3.7 Discussion

3.7.1 Generalizability

Humans naturally use a combination of cues (e.g., facial expressions, body movements, posture) to communicate information. However, not all robots are equipped to replicate such multimodal communication. For example, robots without legs might move on wheels, some may lack arms, and others may have only a head or a simple manipulator. To address this variability, we assume robots using our proposed technique must possess human-like qualities, focusing on hand motions as a primary means of communication. This minimal assumption ensures flexibility, making the approach applicable across different robot platforms. Although

combining hand motions with other modalities (e.g., head or body movements) could enhance communication (e.g., a robot rotating its head or tilting its body while presenting a smelly object), our approach deliberately avoids multimodal dependencies to maximize adaptability.

To enhance generalizability across cultures, our study benefits from distinguishing between *motor-control motions* and *communicative motions*. Motor-control motions are universal, arising from innate responses to physical interaction (e.g., adjusting lifting speed for weight or withdrawing hand(s) quickly from a hot surface). In contrast, communicative motions are socially motivated and explicitly designed to convey information (e.g., rotating an object to indicate lightness or supporting an object to suggest fragility). These culturally influenced hand motions vary in interpretation, depending on shared conventions and experiences. By highlighting these distinctions, robots can be designed to interact naturally with diverse user groups, ensuring effective communication across different cultural contexts.

3.7.2 Optimization for System Performance

Enhancing system performance involves addressing multiple aspects, including adaptability to time-critical scenarios, personalization, and distinguishing similar key elements across object properties.

First, in time-critical scenarios, our system can be further improved by reducing the resolution of conveyed properties. For instance, similar properties such as *hot* and *cold* could be merged into a generalized cautionary category. Although users may not differentiate the exact property, they will still recognize the need for careful handling. This approach prioritizes safety and efficiency in environments where detailed communication is impractical.

Second, personalization can enhance performance by accounting for variability in individual and cultural preferences. Currently, utility metrics are derived from aggregated participant ratings, which may overlook nuanced differences. Incorporating personalization methods, such as tailoring utility scores for users with similar preferences or cultural contexts, could significantly improve effectiveness. Additionally, the system can benefit from co-designing gestures with end-users in specific environments, such as noisy or silent settings, to identify a set of hand motions that are intuitive, easily understood, and broadly applicable across diverse

contexts.

Finally, distinguishing similar key elements across object properties is crucial for reducing ambiguity. Some properties, such as *cold* and *sticky*, share overlapping elements like “touch and withdraw with pause,” while *hot* features a slight variation of “touch and withdraw.” To address this challenge, the system can prioritize emphasizing unique key elements specific to each property, making the motions more distinct and easier to interpret.

3.7.3 Limitations

While our system demonstrates its effectiveness in generating expressive and property-specific hand motions, it presents a scalability limitation, particularly when integrating new object properties. Each new property requires the identification and extraction of key elements from actors’ role-plays, the development of robot key elements, and the evaluation of utility for these elements. Although this approach ensures the generation of expressive and property-specific actions, it is inherently time-intensive and resource-demanding.

Despite this challenge, we believe our system design provides a strong foundation for designing robot hand motions that effectively communicate object properties, paving the way for further advancements in this area.

3.8 Conclusion

In this paper, we investigated how robots communicate properties through object manipulation. We first identified eight categories of object properties and extracted a set of key elements from role-plays. These elements were implemented on a dual-arm robot and evaluated for their utility based on participant feedback. To effectively convey object properties within time constraints, we developed a property-based motion planner that optimally balances utility and time ratio. Our within-subjects study with 20 participants demonstrated that individuals could accurately interpret these properties through robot object manipulation.

CHAPTER 4

COMMUNICATING OBJECT RELATIONS THROUGH ROBOT GESTURES

4.1 Introduction

Gestures play a fundamental role in human communication. Across a wide range of social settings from conversation [44] and storytelling [64] to formal instruction [4, 80], people naturally accompany their speech with gestures. These gestures help draw attention [30], reinforce or supplement spoken content, convey semantic content [44], and visualize spatial or abstract concepts [3, 64].

Among the diverse types of gestures, one particularly expressive category involves holding and moving two physical objects while describing their semantic relations (e.g., similarity, difference). We refer to this category as *relational gesture*. Relational gestures are especially common in product demonstrations, educational settings, and collaborative decision-making. When well-synchronized with speech, such gestures not only reinforce verbal content but also visually clarify comparative meanings, making communication more intuitive and engaging.

Despite its prevalence in everyday human interaction, relational gestures remain underexplored in the field of Human-Robot Interaction. For robots designed

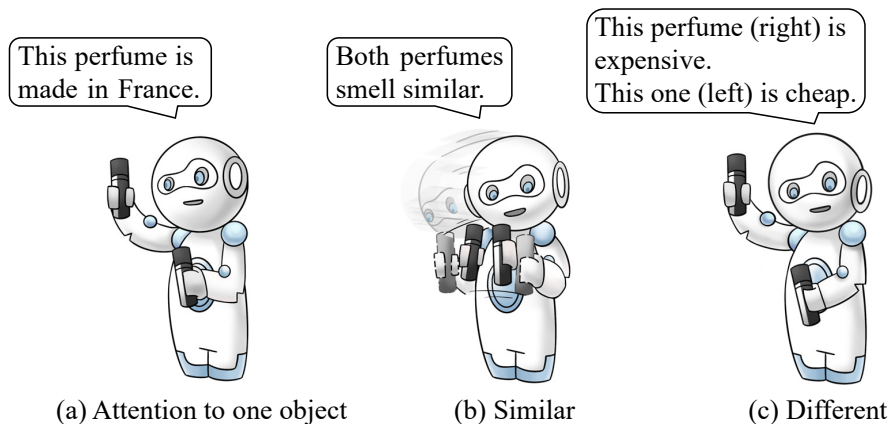


Figure 4.1: Robot gestures during object comparison.

to engage in multimodal communication, the ability to perform such gestures is essential. Imagine a robot holds two objects but remains motionless. Even with accurate verbal descriptions, its communicative intent may appear ambiguous, potentially leading to listeners’ confusion or disengagement. In contrast, expressive gestures such as raising one hand while looking towards the referred object ((Fig. 4.1 (a))), moving both hands closer while alternating gaze (Fig. 4.1 (b)) to indicate similarity, or sequentially raising one hand while orienting the head toward the referred object, then lowering the other hand while shifting gaze to the other object (Fig. 4.1 (c)) to emphasize difference, can significantly enhance clarity and effectiveness of the communication, especially when verbal cues are limited or ambiguous.

This raises a key challenge: How can robots effectively perform relational gestures in coordination with speech? To address this gap, we propose an LLM-based co-speech gesture generation system inspired by how humans naturally synchronize relational gestures with speech. Our work makes four primary contributions. First, we identify four object relations commonly conveyed in demonstrative comparisons. Second, we propose an LLM-based system that produces relational gestures synchronized with speech to convey these relations. Third, we demonstrate the effectiveness of our system through human studies, showing that people can accurately identify object relations from robot gestures. Fourth, we explore the impact of relational gestures on enhancing a robot’s social acceptance, particularly in the role of a shopkeeper.

4.2 Related Work

4.2.1 Gestures in Human Communications

Gesture refers to a visible action of any body part, when used as an utterance, or as part of an utterance[45]. Gestures are commonly categorized into iconic, metaphoric, deictic, and beat gestures [64], each serving distinct communicative functions [3, 30, 44, 64]. Recent work has examined the semantic congruency between representational gestures and accompanying speech [37]. However, while such studies focus on how gestures refer to objects or spatial information, the use of gestures to communicate semantic relations between objects remains underexplored.

4.2.2 Gestures in Human-Robot Interaction

In HRI, gestures serve as an essential nonverbal modality, encompassing hand movements, facial expressions, gaze shifts, head motions, and full-body movements [75]. Prior research has widely recognized the importance of gestures in making robot communication more legible, expressive, and socially appropriate [28, 31, 56, 58]. Recent research also focuses on conveying object properties through physical interactions with one object [73, 84]. Furthermore, supportive gestures toward a single object can increase perceived politeness and competence in service contexts [72]. While these studies provide valuable insights into gesture interaction with a single object, more recent work has begun to explore multi-object scenarios. For example, multiple gesture types were combined to perform actions on two objects (e.g., swap the positions of two objects) with gesture sentences[95]. Yet, the communication of semantic relations between objects through robot gestures remains underexplored.

4.2.3 Co-speech Gesture Generation

Co-speech gestures are visible actions produced while speaking [96]. Co-speech gesture generation, therefore, focuses on generating human body movements aligned with speech input [14]. Existing approaches can be broadly divided into rule-based and data-driven methods [71]. Early rule-based approaches relied on linguistic rules to translate speech into sequences of predefined gesture segments

[13, 51]. Recent advances have shifted to data-driven approaches. However, traditional deterministic generative models tend to produce overly smooth gestures [9, 100] due to their inability to handle many-to-many mappings. To address this, modern generative models like Diffusion [19] have been utilized such as SynTalker[14] and GestureLSM[61]. While these approaches are capable of producing full-body, speech-aligned gestures, how to encode semantic relations between objects through these gestures remains underexplored.

4.2.4 Large Foundation Models in Robotics

Recent advancements in Large Foundation Models (LFMs), including LLMs, Vision-Language Models (VLMs), and Vision-Language-Action Models (VLAs), show significant potential in robotics. LLMs demonstrate capabilities from low-level planning [67, 93] to high-level planning [2, 36]. VLMs are also leveraged to interpret social context, enabling a mobile robot to plan socially aware paths [88] or generate task plans for robots [97]. Building on this, VLAs leverage robotics data to enable direct robot control [24, 48, 101]. In addition, LLMs are explored for robot gesture generation such as head gestures [62] and hand gestures [35]. While these studies address robot control or gesture generation in isolation, how to generate robot gestures synchronized with speech remains underexplored.

4.3 Relational Gestures During Object Comparisons

During object comparisons, people’s body movements naturally accompany their speech, helping to highlight similarities or emphasize differences between objects. Understanding these naturally occurring gestural patterns is essential for enabling robots to communicate in a more human-like and intuitive manner. In this section, we observe how shopkeepers naturally use speech and gestures to compare objects, and analyze the patterns that emerge.

4.3.1 Data Collection from Human Demonstrations

To collect expressive relational gestures, we recruited shopkeepers to compare objects. From their speech, we identified four types of object relations.

Participant Selection

We recruited two experienced female shopkeepers (aged 30 and 60), each with over ten years of retail experience. Both participants reported rich experience in comparing products using both speech and gestures to highlight features, explain similarities and differences, and assist customers in making well-informed decisions.

Selection of Objects

Our study included eight object pairs for comparison: two generic items (bottle and box), and six commonly used items (smartphone, earphones, watch, handbag, wallet, and pen). These objects were used based on two criteria: (1) objects should be handheld with multiple contrastive features; and (2) objects should commonly be used in everyday or demonstrative communication. Together, they represent both general and context-specific scenarios.

Procedures

Upon arrival, shopkeepers received a brief overview and signed informed consent. They were then asked to compare the eight object pairs using natural speech and gestures, just as they would in real-world customer interactions. For each pair, we provided several candidate comparison features (e.g., weight, material) and encouraged participants to introduce additional features as needed.

The study consisted of two rounds across all pairs. In the first round, participants compared each of the eight pairs while holding each object in a designated hand. In the second round, the same sequence of pairs was repeated, with the two objects swapped between hands. The entire session lasted approximately three hours.

Identification of Object Relations

We video-recorded all sessions and transcribed the shopkeepers' natural speech. From these transcripts, we focused only on utterances that directly referred to the hand-held objects. We then annotated these sentences with the type of semantic relation they expressed. Because the shopkeepers compared a diverse set of object pairs in natural sales interactions, these utterances provided a wide vari-

ety of object descriptions. Based on their semantic content, we identified three frequently occurring relation types:

- **No-relation:** Mentions both objects without expressing any relation. For example, “Let’s look at the screen size of these two phones.”
- **Similar:** Expresses similarity between the two objects. For example, “The size of these two phones is similar.”
- **Different:** Emphasizes difference between the two objects.

For *no-relation* and *similar*, identifying the object relations conveyed in utterances was straightforward, as the relations were typically expressed within a single sentence. However, for *different*, contrasts were often conveyed across two consecutive utterances. For example, “The iPhone is thinner. The Android is thicker.” To ensure consistent annotation, we applied minimal coding criteria to determine whether a pair of utterances represented *different*: (1) each utterance refers to a different object; (2) both concern the same semantic topic (e.g., size, price); and (3) they collectively express an explicit contrast.

In addition to these three relation types, we found that most utterances referred to only one object without making a comparison. While these utterances did not express a semantic relation, they coincided with important and meaningful gestures that draw attention to one object. We thus included them as a separate category:

- **One-object:** Refers to a single object. For example, “The color of the iPhone is red.”

In total, we annotated 370 *one-object*, 51 *no-relation*, 83 *similar*, and 29 *different* utterances.

4.3.2 Gesture Analysis

To understand how gestures convey object relations in coordination with speech, we conducted an utterance–gesture analysis. For each utterance collected in Sec. 4.3.1, we examined a range of gestural features, including the number of hands involved, the direction of hand movements, head movements, the amplitude of head and hand movements, and the presence of repeated movements.

Through this analysis, we identified three key features that effectively distinguish between different object relation types: *hand pattern*, *hand direction*, and *focus of attention*. In addition, we identified the most representative gesture associated with each relation type.

Hand Pattern

This feature captures how both hands are coordinated when conveying object relations. We identified five recurrent patterns:

- **One-hand:** Only one hand moves while the other remains still. For example, the shopkeeper raises the left hand holding a long wallet while keeping the right hand holding a short one at rest.
- **Synchronized:** Both hands move at the same time with similar or symmetrical trajectories. For example, the shopkeeper moves both hands holding wallets upward and forward.
- **Sequential:** The two hands move one after the other in an alternating manner. For example, the shopkeeper raises the left hand holding a long wallet, and then lowers the right hand holding a short one.
- **Still:** Both hands remain motionless.
- **Other:** All other cases.

We analyzed the distribution of hand patterns across four relation types. Tab. 4.1 presents the frequency of each hand pattern (rows) observed within each relation type (columns). For example, among the 370 one-object utterances, 335 were accompanied by one-hand gestures, suggesting a strong association between one-object and one-hand. In contrast, most similar utterances (62 out of 83) were expressed with synchronized hand movements, while most different utterances (22 out of 29) involved sequential hand movements.

Hand Direction

This feature captures the dominant axis along which the hands move:

- **Vertical:** Movements directed upward or downward. For example, the shopkeeper raises the left hand holding a handbag.

Table 4.1: Hand pattern

	One-object (370)	No-relation (51)	Similar (83)	Different (29)
One-hand	335	1	3	0
Synchronized	4	14	62	4
Sequential	3	10	3	22
Still	20	26	15	3
Other	8	0	0	0

Table 4.2: Hand direction

	One-object (370)	No-relation (51)	Similar (83)	Different (29)
Vertical	326	21	26	22
Horizontal	15	3	36	4
Other	29	27	21	3

- **Horizontal:** Movements directed side-to-side or inward/out-ward. For example, the shopkeeper swings both hands holding handbags from left to right.
- **Other:** All other cases.

Similarly, Tab. 4.2 also presents the distribution of hand directions across four relation types. For example, vertical hand movements dominated the *one-object* (326 out of 370) and *different* (22 out of 29). In contrast, *similar* was more often expressed through horizontal movements (36 out of 83). *No-relation* frequently fell into *other* (27 out of 51), as many involved still hands without clear directions.

Focus of Attention

This feature captures how the head was oriented toward both hands holding objects:

- **One-target:** The head is directed toward the hand holding the referred object. For example, the shopkeeper oriented the head toward the right hand holding an electronic watch while describing it.

Table 4.3: Focus of attention

	One-object (370)	No-relation (51)	Similar (83)	Different (29)
One-target	333	0	0	0
Both-target	1	3	23	2
Sequential	1	31	38	22
Other	35	17	22	5

- **Both-target:** The head is oriented to encompass both hands simultaneously. For example, the shopkeeper slightly tilts the head downward to look at the area between both hands holding watches, giving balanced attention to them.
- **Sequential:** The head alternates its gaze between both hands. For example, the shopkeeper looks at the right hand holding an electronic watch, then shifts gaze to the left hand holding an Apple Watch.
- **Other:** All other cases.

Likewise, as shown in Tab. 4.3, *one-object* was most often accompanied by *one-target* (333 out of 370). In contrast, *no-relation*, *similar*, and *different* were most frequently expressed with sequential head movements (31 out of 51, 38 out of 83, and 22 out of 29, respectively).

Summary of Representative Gestures

The analyses of hand patterns, hand directions, and focus of attention revealed consistent combinations that characterize each relation type. By combining these three features, we identified the most representative gesture for each relation (Fig. 4.2):

- **One-object:** One hand holding the referred object rises while the other remains still (319 out of 370), and the head orients toward the hand holding the raised object (333 out of 370).
- **No-relation:** Both hands holding objects remain still (26 out of 51), and the head alternates its gaze between the hands (31 out of 51).

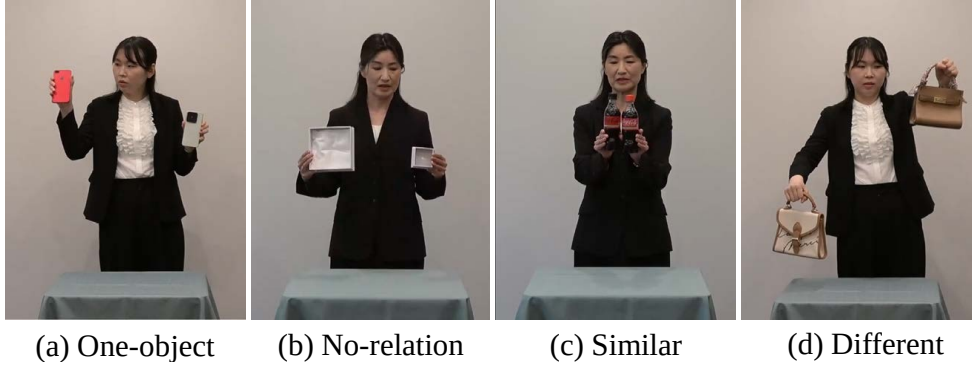


Figure 4.2: Representative gestures for each object relation.

- **Similar:** Both hands holding objects move horizontally toward each other in synchrony (34 out of 83), and the head alternates its gaze between the hands (38 out of 83).
- **Different:** One hand holding the referred object rises or lowers as the gaze follows it, then the other hand moves in the opposite direction (13 out of 29) and the head shifts to the other object (22 out of 29).

Overall, these findings suggest that object relations are conveyed not by isolated features, but by coordinated patterns of hand and head movements. These representative gestures provide the basis for designing relational gestures.

4.4 Relational Gesture Generation System

4.4.1 Overview

We propose an LLM-based system that enables robots to compare objects through co-speech relational gestures. Fig. 4.3 illustrates the architecture of our proposed system, which consists of three main modules: a *co-speech gesture generator*, a *relational gesture planner*, and the *gesture integrator*. The system takes utterance text used for comparing objects as input. The speech synthesizer generates speech audio, which is then fed into the co-speech gesture generator to produce gestures synchronized with speech. Simultaneously, the same text is processed by the relational gesture planner to generate the relational gesture plan that reflects the semantic object relations. The gesture integrator then integrates the rela-

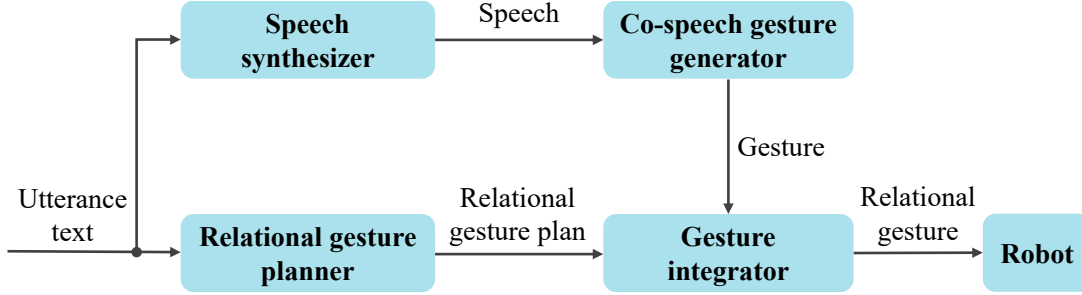


Figure 4.3: The diagram of proposed system.

tional gesture plan with the co-speech gestures, integrating relational cues while preserving the synchrony between speech and gestures.

We adopt a modular architecture for two key reasons. First, it allows precise control over relational gestures especially since our goal is to investigate how relational gestures contribute to communication. Second, by designing the relational gesture planner as a stand-alone module, it can be reused and adapted across different tasks and robotic platforms. To maintain generalizability, our planner does not rely on handcrafted gesture labels (e.g., annotating which sentence should trigger which hand and head movements). Instead, the planner only requires the utterance text and minimal object–hand assignments (which object is in each hand).

4.4.2 Co-Speech Gesture Generator

Co-speech gestures serve as the foundation onto which our system integrates relational gestures. To generate these base gestures, we use a co-speech gesture generator that takes speech audio as input and outputs full-body skeleton trajectories. We adopt *SynTalker* [14] to instantiate this module, as it represents a state-of-the-art model for producing full-body gestures synchronized with speech.

For generalizability across robotic platforms, we focus only on seven upper-body joints: the head, shoulders, elbows, and wrists. These joints are most relevant for relational gestures and ensure that our system can operate on robots with varying degrees of freedom. The extracted sequences of seven joints are then passed to the gesture integrator for integration with relational gestures.

4.4.3 Relational Gesture Planner

The relational gesture planner is responsible for determining which gestures should be performed to express semantic object relations described in the utterance text. To accomplish this, we implement the planner using a large language model (GPT-4.1 *) through prompt-based inference.

The planner operates in a single end-to-end pipeline, and the representative gesture descriptions identified in Sec. 4.3.2 serve as the main knowledge source for the LLM prompt. Specifically, the behavioral patterns we extracted from human demonstrations are directly passed to the model to guide its reasoning. To make these descriptions more general, we slightly modified them by removing explicit references to specific objects, and converting them into bullet-point format. For example, the original description for similar (“Both hands holding objects move horizontally toward each other in synchrony, and the head alternates its gaze between the hands”) was rephrased as: “– Both hands move horizontally toward each other in synchrony. – The head alternates its gaze from one hand to the other hand.” The full prompt therefore includes:

- Definitions of the four object relation types (Sec. 4.3.1);
- The modified representative gesture descriptions for each relation type (as above);
- The utterance text for comparing objects;
- Object–hand assignments (e.g., “The leather handbag is in the left hand. The canvas one is in the right hand”).

These inputs are used to instruct the model to (1) infer the relation type described in the utterance, and (2) generate a gesture plan specifying a high-level textual description of what type of relational gesture should occur, which will be used by the gesture integrator to integrate gestures. For example, given the utterance “Both handbags are made in France,” the planner generates the plan “– Both hands move horizontally toward each other in synchrony. – The head alternates its gaze from the left hand to the right hand.”

*<https://platform.openai.com/docs/models/gpt-4.1>

4.4.4 Gesture Integrator

Once a relational gesture plan is obtained, the gesture integrator integrates it with the full-body co-speech gesture. While it may seem intuitive to generate the relational gestures first and then integrate them with co-speech gestures, this approach is impractical for two key reasons: (1) LLMs do not know the number of skeleton frames required to represent relational gestures, making temporal alignment difficult; and (2) the skeletons generated by LLMs may not conform to the kinematic structure used by the co-speech gesture generator.

Instead, we treat the gestures from co-speech gesture generator as a base layer, and apply relational gestures as an overlay, modifying only the relevant joints (head and arms) as specified by the plan from the relational gesture planner. To support LLM interpretation, we preprocess the skeleton trajectories of co-speech gestures by converting the coordinate system and downsampling the sequence to reduce input length [35]. We apply integration rules based on the plan: (1) if only one hand is used, co-speech gestures are applied only to that hand; (2) for synchronized gestures, co-speech gestures are mirrored across both hands; (3) when both hands move independently, gestures are applied separately, and (4) co-speech gestures are applied to the head consistently. Finally, the integrated gestures are upsampled back to restore frame alignment with the speech audio. We use the o3 model[†] for this module.

4.4.5 Robot

We used Robovie [41], a mid-sized humanoid robot (1.2 m tall) equipped with a wheel-based mobile base and human-like arm and head motions. Both arms have 4 degrees of freedom (DoF), and the head has 3 DoF, enabling expressive hand and head movements.

Given a relational gesture, we compute the orientations of the head and arms and automatically map these joint angles to the robot.

[†]<https://platform.openai.com/docs/models/o3>

4.5 Study I: Relational Gestures Recognition

The goal of this study is to evaluate how effectively robot-performed relational gestures communicate object relations without speech. Specifically, we measure *recognition accuracy*, defined as how accurately participants can identify the intended object relation based solely on the robot’s gestures.

4.5.1 Participants

We recruited 20 participants through a part-time job recruitment website, ranging in age from 18 to 60 years ($M = 26.900$, $SD = 12.981$). Ten participants self-identified as male and ten as female. All participants received compensation for their participation.

4.5.2 Stimuli

We used three representative object pairs as stimuli for the experiment: *smartphones*, *handbags*, and *watches*, selected from the set described in Sec. 4.3.1. For each object pair, our system generated two gesture variations per relation type using different input utterances. This resulted in a total of 24 robot-performed relational gestures. Examples of these gestures are shown in Fig. 4.4.

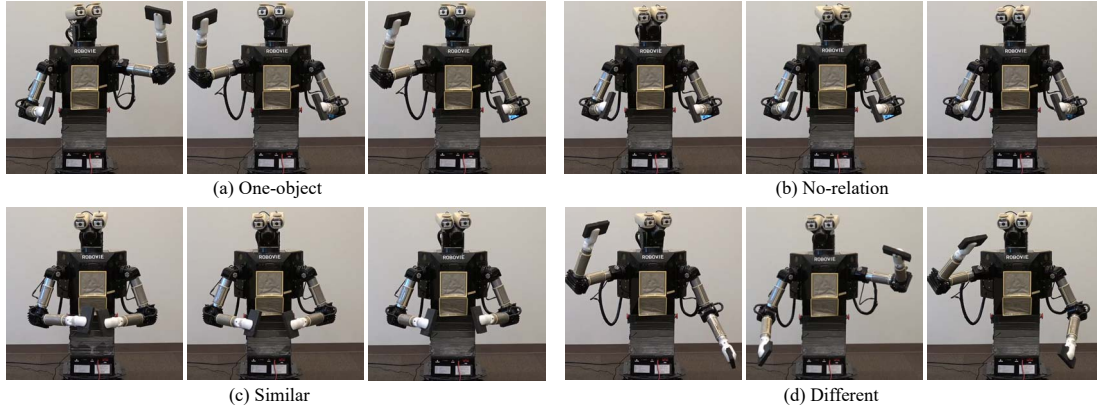


Figure 4.4: Examples of robot relational gestures.



Figure 4.5: Setup for the experiment in Study I.

4.5.3 Procedure

Participants were welcomed into a large room where the robot was situated and were given an overview of the study before signing a consent form. Two identical black boxes were used as props to eliminate any influence from object appearance.

The setup of the experiment is shown in Fig. 4.5. Each participant completed the task individually. They observed 24 robot-performed gestures without speech (Sec. 4.5.2), each corresponding to one of four object relation types. After each gesture, participants selected the relation they thought the robot intended to convey. This setup allowed us to assess recognition accuracy across relation types. To ensure each interaction was independent, participants were informed that any of the four relations could be attributed to each gesture. To reduce potential order effects, a partial Latin square design was used to counterbalance the gesture order. To reflect the gesture-only nature of the task, we revised the original definitions of the object relations (Sec. 4.3.1) to focus solely on the relations conveyed through gestures, and presented them to the participants:

- **One-object:** Attention is directed toward a single object.
- **No-relation:** Attention is directed toward both objects, but no specific relation is conveyed.
- **Similar:** The gesture expresses the similarity between the two objects.
- **Different:** The gesture emphasizes the difference between the two objects.

Table 4.4: The recognition accuracy for object relations

	One-object (120)	No-relation (120)	Similar (120)	Different (120)	Average (480)
One-object	119	2	1	1	
No-relation	0	105	12	9	
Similar	0	5	100	3	
Different	1	8	7	107	
Accuracy	0.992	0.875	0.833	0.892	0.898

Finally, participants took part in a semi-structured interview. The study was approved by the Institutional Review Board.

4.5.4 Results

Recognition Results

Tab. 4.4 shows the confusion matrix of participants’ responses, along with the recognition accuracy for each object relation. Each column represents each gesture type performed by the robot, and each row corresponds to the relation recognized by participants. Overall, participants correctly identified the intended object relations in 89.8% of the trials. The recognition accuracy was 99.2% for *one-object*, 87.5% for *no-relation*, 83.3% for *similar*, and 89.2% for *different*.

The most common misclassifications involved *no-relation*. Specifically, *similar* was misclassified as *no-relation* in 12 instances, and *different* was misclassified as *no-relation* in 9 instances. These results suggest that, in the absence of speech, the visual distinctions among no-relation, similar, and different may be less distinct, making them more difficult to differentiate.

Interview Results

The interview results were consistent with the recognition results. All participants reported that *one-object* was the easiest to recognize. Seven participants mentioned that the robot’s hand and head movements made the gestures vivid and easy to recognize. Four participants noted that some gestures, particularly similar (e.g., bringing both hands closer together), resembled those commonly used by human shopkeepers. In contrast, four participants expressed difficulty

distinguishing *no-relation* from *similar* and *different*, citing a lack of clear visual contrast.

4.6 Study II: Co-Speech Relational Gesture Evaluation

We investigate how relational gestures affect people’s impressions of a robot during object comparison. Specifically, we examine the effect of our proposed system, which integrates relational gestures into co-speech gestures, by comparing it against a baseline that uses only co-speech gestures generated by SynTalker [14]. We selected SynTalker as the baseline because it represents a state-of-the-art model in recent co-speech gesture generation research. Since our proposed system also incorporates SynTalker for generating base gestures, this comparison allows us to isolate and assess the added value of relational gestures in shaping perceptions of the robot.

4.6.1 Hypotheses and Predictions

Prior research has shown that robots exhibiting meaningful motions are perceived as more competent and skilled than those without such motions [17, 72]. In our study, relational gestures convey semantic relations between objects, positioning them as a form of meaningful motion. Based on this, we made the following prediction:

- **P1:** Robots performing co-speech relational gestures will be perceived as more **competent** than those performing only co-speech gestures.

Robots are perceived as more sociable when they convey information clearly through either verbal or non-verbal modalities [43]. Given that Study I showed people could correctly identify object relations from relational gestures, we made the following prediction:

- **P2:** Robots performing co-speech relational gestures will be perceived as more **sociable** than those performing only co-speech gestures.

Robot gestures derived from human demonstrations tend to appear more human-like [18]. Moreover, robots exhibiting a variety of congruent gestures are often perceived as more human-like compared to those with limited or inconsistent gestures [99]. Since relational gestures are modeled from shopkeeper behaviors and offer greater diversity than co-speech gestures. Accordingly, we made the following prediction:

- **P3:** Robots performing co-speech relational gestures will be perceived as more **animate** than those performing only co-speech gestures.

4.6.2 Participants

We recruited 20 participants based on a priori power analysis (Cohen’s $d = 0.8$) [15], which indicated that a minimum of 15 participants was required to achieve 80% power at a 95% confidence level. Participants were recruited through a part-time job recruitment website, and ranged in age from 18 to 59 years ($M = 32.150$, $SD = 16.178$). Ten participants self-identified as male and ten as female. All participants received compensation for their participation.

4.6.3 Conditions

The robot’s performance was compared under the following two conditions:

- **Proposed:** The robot operates with the gestures generated from the proposed system described in Section 4.4.
- **Baseline:** The module of the relational gesture planner is excluded from the architecture of the proposed system (Fig. 4.3). In this case, the robot operates with the gestures generated from the co-speech gesture generator without being modified by the gesture integrator.

We employed a within-subjects design, with the order of conditions counter-balanced. The spoken utterances were identical across conditions; the only difference lay in the robot’s gestures.

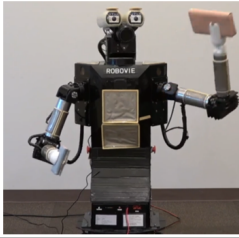
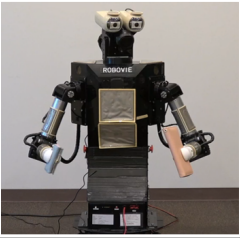
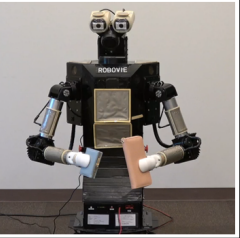
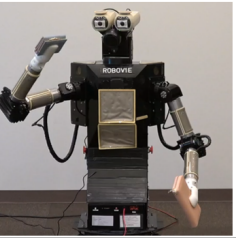
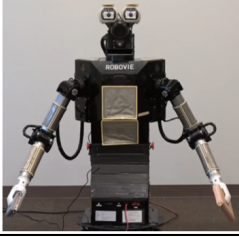
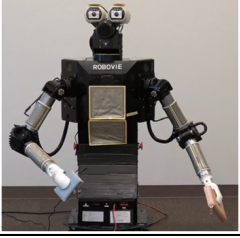
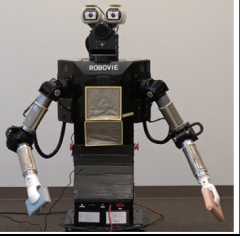
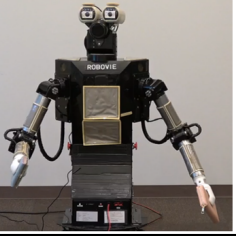
Relation	One-object	No-relation	Similar	Different
Utterance	In my left hand is a long pink wallet.	I would like to introduce two classical wallets.	Both wallets are made in France.	The long wallet is heavy. The short one is light.
Proposed				
Baseline				

Figure 4.6: Examples of the robot’s behavior in Study II.

4.6.4 Procedure

Participants were welcomed into a large room where the robot was situated and were given an overview of the study before signing a consent form. In both conditions, the interaction followed a one-way format. The robot acted as a shopkeeper introducing and comparing two wallets. The participant, assigned the role of a customer, was instructed to listen and observe the robot, without talking to it. Each session lasted for five minutes. Examples of the robot’s behavior in both conditions are shown in Fig. 4.6. For example, when speaking the utterance “Both are made in France,” the robot in the proposed condition brought both hands closer together and smoothly shifted its head from one hand to the other. In contrast, the robot in the baseline condition exhibited a larger movement with its right hand, a smaller motion with its left hand, and unclear head movement without expression of the intended relation. After each condition, participants completed a questionnaire evaluating their impressions of the robot.

Finally, we had semi-structured interviews with the participants. This study was also approved by the Institutional Review Board.

4.6.5 Measurement

We evaluated competence, sociability, and animacy by using 1-to-7 point Likert-scale questionnaires composed of validated items from the HRI literature.

- **Competence:** Measured using six items from the RoSAS [12] - *capable, responsive, interactive, reliable, competent, and knowledgeable*.
- **Sociability:** Measured using four items from the HRIES [90] - *warm, likeable, trustworthy, and friendly*.
- **Animacy:** Measured using four items from the HRIES - *alive, natural, real, and human-like*.

4.6.6 Results

Verification of Predictions

- **Competence:** As illustrated in the first set of bars in Fig. 4.7, competence scores were averaged across the six corresponding items. A Shapiro-Wilk test confirmed normality for both conditions ($W_{baseline} = .957, p = .481$; $W_{proposed} = .977, p = .885$). Consequently, a paired t-test was employed. There was a significant difference between the proposed condition ($M = 5.417, SD = 0.844$) and the baseline ($M = 4.117, SD = 1.279$), $t(19) = 6.264, p < .001$, Cohen's $d = 1.401$. This supports P1: **the robot with co-speech relational gestures was perceived as more competent**.
- **Sociability:** As shown in the second set of bars of Fig. 4.7, sociability scores were averaged across the four corresponding items. Normality was confirmed ($W_{baseline} = .948, p = .342$; $W_{proposed} = .975, p = .863$). A paired t-test showed a significant increase in sociability for the proposed condition ($M = 5.150, SD = 1.043$) compared to the baseline ($M = 4.250, SD = 1.360$), $t(19) = 3.105, p = .006$, Cohen's $d = 0.694$. This supports P2: **the robot with co-speech relational gestures was perceived as more sociable**.
- **Animacy:** As shown in the third set of bars of Fig. 4.7, animacy scores were averaged across the four corresponding items. Normality was also confirmed

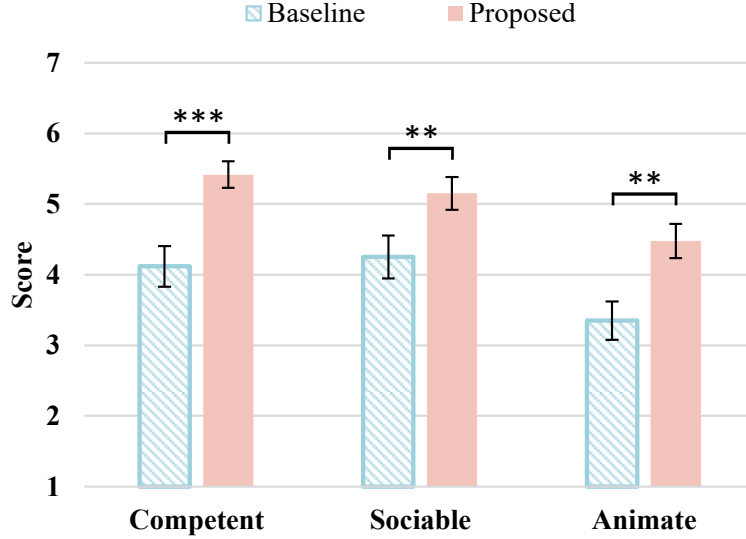


Figure 4.7: Results from our quantitative measures. The error bars show the standard error for the mean at $\pm 1 SE$. (** : $p < .01$, *** : $p < .001$)

($W_{baseline} = .944$, $p = .281$; $W_{proposed} = .964$, $p = .628$). The proposed condition ($M = 4.475$, $SD = 1.085$) was rated significantly higher than the baseline ($M = 3.350$, $SD = 1.215$), $t(19) = 3.135$, $p = .005$, Cohen’s $d = 0.701$. This supports P3: **the robot with co-speech relational gestures was perceived as more animate.**

Interview Results

When asked to explain their judgments regarding competence, 19 participants remarked that the robot’s relational gestures were well-aligned with its speech, making the explanations easier to follow. Six participants noted that these gestures closely resembled those of human shopkeepers, which enhanced the robot’s perceived competence. In contrast, four participants described the robot without relational gestures as mechanical, as if it were simply playing a recording or broadcasting information. One participant even compared its behavior to a child waving a toy around, which she found annoying.

In terms of *sociability* and *animacy*, 12 participants favored the robot with relational gestures, explaining that its demonstrated competence made it appear better suited for social interaction. Two of them also noted that this perceived competence increased their trust in the robot. However, four participants ex-

pressed a contrasting view: they felt that natural human behavior often includes minor or awkward movements, which contribute to a sense of charm or lifelikeness. As a result, they found the robot without relational gestures more appealing. Interestingly, one participant reported discomfort with how closely the robot mimicked human shopkeeper behavior, stating a preference for robots that are “not too clever,” as it maintained a clearer boundary between humans and robots.

4.7 Discussion

4.7.1 Design Implications

Our system demonstrates that LLMs can effectively generate relational gestures for object comparison tasks. This highlights the value of integrating empirical findings on human behaviors into generative AI systems to produce expressive and communicative robot behaviors with relational cues.

Beyond this specific application, our approach outlines a generalizable pipeline for behavior generation: (1) identify a target communicative function; (2) collect natural human interaction data that exemplifies this function; (3) analyze the data to extract behavior patterns (e.g., gestures, object manipulations); and (4) use large foundation models (e.g., LLMs, VLMs) to generate behaviors guided by identified patterns. This framework bridges empirical human studies with generative AI, and can be used broadly in domains such as human–robot interaction and virtual agents.

4.7.2 Limitations

While our findings demonstrate the effectiveness of relational gestures in enhancing communication and social perception, several limitations should be noted. First, our data collection and evaluation focused on retail settings with shopkeepers comparing objects. It may not fully capture the diversity of relational gestures in broader contexts such as education and collaborative work. Second, our system prioritizes gestures that are most frequently observed and easily interpretable. While this design enhances clarity and recognition, it reduces the variability seen in human gestures. Finally, our system is not yet optimized for

real-time use. The *gesture integrator* relies on LLM-based inference, which introduces latency and may limit responsiveness in interactive dialogue scenarios.

4.8 Conclusion

We explored how robots can communicate object relations through relational gestures. We began by collecting behavioral data from shopkeepers comparing objects. Through analysis of their speech and gestures, we identified four common types of object relations, along with representative gestures for each. Building on these insights, we developed an LLM-based system that can generate relational gestures for comparing objects. We evaluated the system through two user studies. The first study with 20 participants showed that participants were able to accurately identify the intended object relations from robot relational gestures. The second study with another 20 participants revealed that robots using relational gestures were perceived as more competent, sociable, and animate than those without them.

CHAPTER 5

GENERAL DISCUSSION

5.1 Unifying Framework

While the three studies in this thesis investigate different types of cues and employ different generation methods, they are all driven by the same research objective: to enable robots to communicate how objects should be handled through manipulation behaviors. This direction differs from traditional robot manipulation research, which mainly focuses on achieving task goals, as well as from prior expressive motion studies that primarily convey representational information such as a robot’s intention, emotion, or the identity of an object. Here, the emphasis is placed on object-handling meaning that implicates how objects should be handled or interacted with, allowing humans to understand appropriate handling strategies.

In all three studies, the robot’s manipulation behaviors serve as the communication channel. Instead of relying on gaze, gestures, or facial movements, the motions used to interact with objects are designed to expose informative cues that humans can observe and interpret. Furthermore, the expressive behaviors are tightly integrated into functional manipulation. This combination of requirements ensures that the behavior is both communicatively expressive and operationally valid within collaborative tasks.

Altogether, these studies provide a cohesive foundation showing that expressive

manipulation can reduce uncertainty, help humans anticipate the robot’s handling strategy, and support more coordinated and smooth collaboration.

5.2 Technical Differences

Although the three studies all aim to communicate object-handling meaning through manipulation, they differ substantially in their technical approaches. These differences arise not from arbitrary design choices, but from the fundamentally different types of problems each study addresses.

In Study 1, polite handling is expressed by positioning the idle hand beneath the object. Given the primary hand pose as input, this social meaning can be conveyed through appropriate placement of the idle hand. Because the task is the computation of the poses for the idle hand and does not require reasoning over physical constraints or task-level feasibility, a simple modeling approach is sufficient to generate the desired sub-goal poses for the idle hand.

In Study 2, the robot aims to convey physical properties such as temperature or weight. The input consists of the utility of key expressive elements and the time ratio available for exposure, requiring the robot to balance communicative effectiveness with task feasibility. This formulation leads to a constraint-aware optimization problem, for which a planner is essential, as the task requires multi-objective planning under physical and temporal constraints.

In Study 3, the robot communicates semantic relations between objects, such as similarity or difference. The input is an utterance describing the objects, and the system must infer the underlying relational meaning and map it to the corresponding manipulation behaviors. The core challenge is not optimization, but the identification of gesture patterns expressed in language. Large language models are suitable for this role, as they can process textual descriptions and generate action patterns that reflect the intended semantic relations.

As a result, their methods are not interchangeable, but are instead tailored to the nature of the information being communicated and the type of inference required.

5.3 Design Principles for Scalable Expressive Manipulation

A natural question arises regarding how many expressive gestures are needed for a social or collaborative robot. This concern highlights a fundamental challenge in expressive manipulation research: if each communicative function required its own handcrafted gesture, the design space would quickly become unmanageable, making it difficult for robots to generalize beyond the specific scenarios for which these behaviors were designed.

The aim of this thesis is therefore not to enumerate a collection of isolated behaviors, but to identify generalizable principles for how object manipulation can convey meaningful information about handling. Although the three studies investigate different object-handling meanings, they share a common methodological structure: observing human handling behaviors, analyzing and extracting the key motions that encode meaning, and developing systems that enable robots to reproduce these expressive patterns. These studies serve as representative cases that reveal design principles for expressive manipulation, showing that meaningful handling behaviors can be constructed through meaning-to-motion mappings, rather than through ad hoc gesture design. This perspective enables future systems to generalize across meanings, tasks, and objects, moving toward unified and flexible frameworks for non-verbal communication through object manipulation.

5.4 Is Manipulation Necessary for Conveying Information?

While manipulation offers a powerful and intuitive communication channel, it should not be assumed to be universally necessary or always appropriate. Expressive manipulation inherently couples communication with task execution. This coupling is what makes it effective, as users can infer information directly from the way the robot handles the object. However, it also introduces important limitations that must be carefully considered.

First, excessive expressiveness may disrupt task fluency. Manipulation behaviors often carry both instrumental and communicative functions. If an exaggerated motion, pause, or adjustment becomes too pronounced, it may slow down the

task, confuse the user about the robot’s true intent, or even reduce efficiency in collaborative actions. For example, an overly cautious motion designed to express fragility may prevent smooth handovers, while a highly exaggerated trajectory built to show relational constraints may appear unnatural or unpredictable.

Second, manipulation is not always the most efficient channel for all types of information. Some meanings, such as emotional state, high-level task goals, or referential pointing, can be more naturally expressed through gesture, gaze, or other non-manipulative cues. Attempting to force all forms of communication through manipulation risks overloading the motion itself and mixing multiple signals, leading to ambiguity. The studies in this thesis focus specifically on object-handling meaning, where manipulation is uniquely well suited; however, not all representational meanings require or benefit from object interaction.

Third, manipulation-based communication imposes physical constraints. The information that can be conveyed is bounded by the object’s shape, affordances, mass distribution, and safety requirements. For example, a robot cannot easily express “fragile” through manipulation if the object is large and rigid, nor can it signal “light” if it must maintain strict stability during transport. In such cases, supplementary non-manipulative cues may be necessary to avoid compromising safety.

5.5 Cognitive Foundations of Interpreting Expressive Manipulation

The interpretability of robot actions relies on human perceptual and cognitive mechanisms that extract meaningful information from observed motion.

Prior work shows that the more similar an agent is to a human in appearance or movement, the more likely it is to be anthropomorphized [98]. A robot with more humanlike appearance tends to elicit mind attribution of internal experience, even if robots do not possess genuine internal states. In addition, other research indicates that when observing another agent’s movement, humans generate anticipatory estimates of how the action will unfold, even when the agent is a robot rather than a human [46]. As a result, subtle differences in robot kinematics such as timing, velocity, and hand configuration enable humans to anticipate how the object should be handled next and prepare their own actions

accordingly.

5.6 Generalizability

The generalizability of the findings in this thesis concerns whether expressive manipulation is interpreted by humans in a consistent manner across different user groups, robot embodiments, and cultural contexts, as well as whether the systems remain effective beyond the specific experimental settings evaluated in the three studies.

First, all three studies were evaluated with participants between approximately 20 and 60 years of age, and most of them are not familiar with robots. Within this range, expressive manipulation consistently influenced human interpretation. However, extending these findings to other user groups requires caution. Differences in cognitive development, perceptual sensitivity, and prior experience with robots may influence how expressive manipulation is perceived. For instance, children may interpret expressive motions more literally or anthropomorphically, attributing genuine emotion or physical sensation to the robot. In contrast, older adults may exhibit reduced sensitivity to subtle kinematic cues. Since these factors were not systematically examined, it remains uncertain how expressive manipulation would be interpreted outside the tested user groups. This uncertainty does not imply that the approach would fail, but rather that its effectiveness may depend on how strongly the expressive cues align with the perceptual and cognitive tendencies of the target user group.

Second, each study was conducted using a single robot platform, specifically dual-arm humanoid robots (TIAGo++ and Robovie) with humanlike arm placement and reachable workspaces. Under these conditions, expressive manipulation was shown to affect human interpretation. However, prior work suggests that the attribution of meaning to robot motion depends on embodied resemblance [98]. This implies that the same motion patterns may be interpreted differently on robots with substantially different morphologies. Importantly, even within humanoid platforms, differences in appearance, motion smoothness, or perceived autonomy may influence how expressive cues are perceived. Therefore, while expressive manipulation may plausibly transfer to other robots with comparable bimanual configurations, its interpretation cannot be assumed to be invariant across embodiments and requires explicit cross-platform investigation.

Third, cultural generalizability involves two distinct questions: whether certain non-verbal behaviors exist in human interactions across different cultures, and whether humans interpret these behaviors in the same way when they are performed by robots. Prior work indicates that support hand behaviors appear in other East Asian cultural contexts [91], and they were also observed in Europe and North America, suggesting that these behaviors are not unique to a single cultural context. However, the existence of a behavior in human interaction does not necessarily imply that its robotic counterpart will be interpreted in the same way. Cultural norms influence not only how actions are performed, but also how agency and intention are attributed to robots. In addition, some expressive motions grounded in physical interaction, such as withdrawing the hand rapidly from a hot object, may be more robust across cultures due to shared human motor-control responses. Even in such cases, whether users interpret the robot’s behavior as communicative rather than purely reactive remains an open question. Therefore, cultural generalizability depends both on the prevalence of the corresponding human behaviors and on users’ willingness, across different cultural backgrounds, to map those behaviors onto robotic actions.

Finally, generalizability must be examined at the system level, where the three studies differ substantially in their processing mechanisms. In Study 1, expressive motions were generated using a relatively simple modeling approach that focuses on support-hand placement, without explicitly modeling object shape or variability. As a result, the system’s behavior may change when object geometry or handling conventions differ, and its generalizability beyond the tested objects remains uncertain. In Study 2, expressive manipulation relies on a planner that balances utility and time ratio for specific physical properties. The generalizability of this approach depends on whether the utility remains effective for other objects, object properties, or task contexts. In Study 3, expressive motions are generated through an LLM-based system, and the system performance depends on both the complexity of textual inputs and the reasoning capabilities of the large language models. The reliability and interpretability of the generated behaviors remain uncertain, which influences the generalizability of this approach. Overall, the system-level generalizability of each study differs and thus should be examined separately.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusions

This thesis investigates how robot object manipulation can be designed not only to achieve task execution but also to serve as an expressive channel through which humans can understand how objects should be handled. Departing from prior work that primarily focuses on conveying robot intention, emotion, or object identity, this research explores expressive manipulation as a means of communicating object-handling meaning that implicates appropriate, safe, or socially situated ways of interacting with objects.

Across three studies, the thesis examines how different types of object-handling meaning can be communicated through manipulation behaviors. Study 1 investigates whether a dual-arm robot should use both hands in social contexts when a single hand is sufficient to complete a single-object task. By modulating supportive behaviors for the idle hand during interactions, the robot is able to convey politeness. The results demonstrate that subtle differences in how an object is handled can shape human perceptions of robots in social contexts.

Study 2 focuses on physical object properties that are not directly perceivable, such as weight and temperature. By applying subtle motion cues such as

movement speed, hand involvement, and object or hand placement, the robot expresses these properties through its handling behaviors. The findings show that humans can reliably infer such properties from the ways in which a robot handles an object.

Extending expressive manipulation from single-object handling to dual-object handling, Study 3 examines how a robot can express relations between two objects through manipulation behaviors. The results show that people are able to infer semantic relations between objects from how the robot handles both objects, and that these handling behaviors also significantly improve people’s perceptions of the robot.

Taken together, these three studies show that manipulation is not merely a task-execution mechanism but also a form of expressive communication. This thesis establishes a link between physical interaction with objects and human interpretation: the way a robot handles objects influences how people understand its behavior and may shape their expectations and strategies for collaboration.

6.2 Future Directions

While this thesis demonstrates the feasibility and value of expressive manipulation for communicating object-handling meaning, several research directions remain to further advance this space.

6.2.1 Multimodal and Context-Rich Expression

Future work could expand beyond manipulation alone by combining tactile, visual, and auditory cues to create richer and more robust communication. For instance, slowing the motion when handing over a hot object, together with subtle light or sound indicators, may strengthen the user’s perception of caution. Investigating how different modalities reinforce or interfere with one another would enable robots to form coherent multimodal expressions that better align with human expectations and situational context.

6.2.2 Cognitive Modeling of Human Interpretation

A deeper understanding of how humans infer meaning from motion dynamics is needed. Future research could build computational or psychological models

explaining how observers map manipulation cues to physical and social meanings. Such models may draw on action perception, motor simulation, social cognition, and psychophysics, ultimately enabling robots to predict human interpretation and adjust behaviors accordingly.

6.2.3 Human-Robot-Object Triadic Interaction Framework

Finally, future research may further conceptualize the role of objects as active mediators of meaning. Rather than functioning only as task targets, objects shape the flow of information and expectations between human and robot. A formalized triadic framework would clarify how handling behaviors influence human attitudes, coordination, and trust, and how object characteristics constrain and transform communication. Such a framework could help generalize design principles across different tasks, embodiments, and social settings.

Together, these directions extend the foundational contributions of this work toward a more comprehensive understanding of expressive manipulation in human-robot interaction, enabling robots to act not only as capable manipulators but also as transparent and socially aligned partners in shared environments.

ACKNOWLEDGEMENTS

I have no idea how far I have traveled to reach this point. I can still vaguely remember when I was four or five, my grandfather would carry me on his shoulder as we climbed a mountain overgrown with weeds and ringed by prickly fir trees, heading to the school on the other side. I would sit in the corner of the classroom, receiving my very first lessons. Looking back now, leaving those mountains behind, studying in Japan, and completing my doctoral degree at Kyoto University have been a remarkable journey of growth.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Takayuki Kanda. Throughout my doctoral studies, he provided meticulous guidance in every step of my research: from topic selection, methodology, experimental design, academic writing to conference presentations. This dedication is reflected in the nearly one thousand emails and the hundreds of insightful comments he provided for each paper, all of which carried his patience and commitment. His inquiry-driven teaching style encouraged me to think independently, helping me accumulate strength through challenges and complete each research project successfully. His rigorous academic standards will continue to inspire me as I move forward in my future career. He represents, in my heart, the truest image of a scientist.

I would like to express my sincere appreciation to Professor Takayuki Ito and Professor Michinori Hatayama, members of my thesis committee, for their profound and constructive suggestions during critical phases of my research. I am also grateful to Dr. Malcolm Doering for his continuous guidance and support throughout my doctoral work. I am equally grateful to the faculty members in our laboratory: Professor Dražen Brščić, Professor Jani Even, Professor Stela Hanbyeol Seo, and Professor Kanae Kochigami for their invaluable help. I wish to

Acknowledgements

extend my appreciation to our lab secretaries: Ms. Yoko Kubota, Ms. Yuka Okamoto, Ms. Yukari Hasegawa, Ms. Hiromi Kobayashi, and Ms. Tomoe Weedall, whose assistance in both research and daily life has always brought warmth to my journey.

My sincere thanks go to Ms. Yuan-Chia Chang, Ms. Rui Chen, Ms. Yuxin Zhao, Mr. Yuyi Liu, Ms. Wa Gao, Mr. Kanghui Du, Mr. Yongqiang Jiang, Mr. Yifan Shao, Mr. Lei He, and Ms. MungTim Chung. Your support and encouragement during difficult times have helped me keep moving forward. I would also like to thank Mr. Yuto Ushijima, Mr. Joshua Ravishankar, Ms. Sachi Natasha Edirisinghe, and Mr. Zulkafil Abbas for their collaboration and companionship in research.

I greatly appreciate the MEXT scholarship, which provided essential support for my study and life in Japan.

I am deeply grateful to my family, who have given me unconditional love and unwavering support since the moment I came into this world. My sincere thanks also go to my girlfriend, who has always been by my side these past years. With her, daily life becomes steadier and our shared experiences become more fulfilling. Your support in my research journey has always been a source of great courage.

Finally, I would like to thank myself for choosing perseverance and effort even when faced with doubt and uncertainty. May I continue to move forward with confidence and pursue the path I truly aspire to take.

Xiang PAN, January 2026

BIBLIOGRAPHY

- [1] Roxana M Agrigoroaie and Adriana Tapus. Developing a healthcare robot with personalized behaviors and social skills for the elderly. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 589–590. IEEE, 2016.
- [2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- [3] Martha W Alibali. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial cognition and computation*, 5(4):307–331, 2005.
- [4] Martha W Alibali and Mitchell J Nathan. Teachers’ gestures as a means of scaffolding students’ understanding: Evidence from an early algebra lesson. In *Video research in the learning sciences*, pages 349–365. Routledge, 2014.
- [5] Amir Aly and Adriana Tapus. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 325–332. IEEE, 2013.
- [6] Thibault Asselborn, Wafa Johal, and Pierre Dillenbourg. Keep on moving! exploring anthropomorphic effects of motion during idle moments. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 897–902. IEEE, 2017.

BIBLIOGRAPHY

- [7] Weibang Bai, Ningshan Zhang, Baoru Huang, Ziwei Wang, Francesco Cursi, Ya-Yen Tsai, Bo Xiao, and Eric M Yeatman. Dual-arm coordinated manipulation for object twisting with human intelligence. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 902–908. IEEE, 2021.
- [8] Aaron M Bestick, Samuel A Burden, Giorgia Willits, Nikhil Naikal, S Shankar Sastry, and Ruzena Bajcsy. Personalized kinematics for human-robot collaborative manipulation. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1037–1044. IEEE, 2015.
- [9] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2027–2036, 2021.
- [10] Christopher Bodden, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. A flexible optimization-based method for synthesizing intent-expressive robot arm motion. *The International Journal of Robotics Research*, 37(11):1376–1394, 2018.
- [11] Penelope Brown and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.
- [12] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. The robotic social attributes scale (rosas) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, pages 254–262, 2017.
- [13] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Drouville, Scott Prevost, and Matthew Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420, 1994.
- [14] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. Enabling synergistic full-body control in prompt-based co-speech motion

BIBLIOGRAPHY

- generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6774–6783, 2024.
- [15] Jacob Cohen. Quantitative methods in psychology: A power primer. *Psychol. Bull.*, 112:1155–1159, 1992.
- [16] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. Gesturing makes learning last. *Cognition*, 106(2):1047–1058, 2008.
- [17] Raymond H Cuijpers and Marco AMH Knops. Motions of robots matter! the social effects of idle and meaningful motions. In *International Conference on Social Robotics*, pages 174–183. Springer, 2015.
- [18] Jan De Wit, Paul Vogt, and Emiel Krahmer. The design and observed effects of robot-performed manual gestures: A systematic review. *ACM Transactions on Human-Robot Interaction*, 12(1):1–62, 2023.
- [19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794, 2021.
- [20] Malcolm Doering, Takayuki Kanda, and Hiroshi Ishiguro. Neural-network-based memory for a social robot: Learning a memory model of human behavior from data. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(4):1–27, 2019.
- [21] Xiao Dou, Chih-Fu Wu, Kai-Chieh Lin, and Tzu-Min Tseng. The effects of robot voice and gesture types on the perceived robot personalities. In *Human-Computer Interaction. Perspectives on Design: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part I 21*, pages 299–309. Springer, 2019.
- [22] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE, 2013.
- [23] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. Effects of robot motion on human-robot collaboration. In *Proceedings of*

BIBLIOGRAPHY

- the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 51–58, 2015.
- [24] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023.
- [25] Arjan Egges, Tom Molet, and Nadia Magnenat-Thalmann. Personalised real-time idle motion synthesis. In *12th Pacific Conference on Computer Graphics and Applications, 2004. PG 2004. Proceedings.*, pages 121–130. IEEE, 2004.
- [26] Pierre Feyereisen. Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, 18(2): 185–205, 2006.
- [27] Raphaela Gehle, Karola Pitsch, Timo Dankert, and Sebastian Wrede. How to open an interaction between robot and museum visitor? strategies to establish a focused encounter in hri. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 187–195, 2017.
- [28] Michael J Gielniak and Andrea L Thomaz. Generating anticipation in robot motion. In *2011 RO-MAN*, pages 449–454. IEEE, 2011.
- [29] Steve Guest, Jean Marc Dessirier, Anahit Mehrabyan, Francis McGlone, Greg Essick, George Gescheider, Anne Fontana, Rui Xiong, Rochelle Acklerley, and Kevin Blot. The development and validation of sensory and emotional scales of touch perception. *Attention, Perception, & Psychophysics*, 73:531–550, 2011.
- [30] Christian Heath. *Gesture’s discreet tasks: Multiple relevancies in visual conduct and in the contextualisation of language*. 1992.
- [31] Rachel M Holladay, Anca D Dragan, and Siddhartha S Srinivasa. Legible robot pointing. In *The 23rd IEEE International Symposium on robot and human interactive communication*, pages 217–223. IEEE, 2014.

BIBLIOGRAPHY

- [32] Chien-Ming Huang and Bilge Mutlu. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems*, volume 2. Citeseer, 2013.
- [33] Chien-Ming Huang and Bilge Mutlu. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 57–64, 2014.
- [34] Chien-Ming Huang, Takamasa Iio, Satoru Satake, and Takayuki Kanda. Modeling and controlling friendliness for an interactive museum robot. In *Robotics: science and systems*, pages 12–16, 2014.
- [35] Peide Huang, Yuhan Hu, Nataliya Nechyporenko, Daehwa Kim, Walter Talbott, and Jian Zhang. Emotion: Expressive motion sequence generation for humanoid robots with in-context learning. *IEEE Robotics and Automation Letters*, 2025.
- [36] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.
- [37] Sarah S Hughes-Berheim, Laura M Morett, and Raymond Bulger. Semantic relationships between representational gestures and their lexical affiliates are evaluated similarly for speech and text. *Frontiers in psychology*, 11: 575991, 2020.
- [38] Shunichiro Ito, Kanae Kochigami, and Takayuki Kanda. A robot dynamically asking questions in university classes. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 839–848. IEEE, 2025.
- [39] Jinyung Jung, Takayuki Kanda, and Myung-Suk Kim. Guidelines for contextual motion design of a humanoid robot. *International Journal of Social Robotics*, 5(2):153–169, 2013.

BIBLIOGRAPHY

- [40] Mary Kister Kaiser and Dennis R Proffitt. The development of sensitivity to causally relevant dynamic information. *Child Development*, pages 1614–1624, 1984.
- [41] Takayuki Kanda, Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, and Ryohei Nakatsu. Development and evaluation of an interactive humanoid robot” robovie”. In *Proceedings 2002 IEEE international conference on robotics and automation (Cat. No. 02CH37292)*, volume 2, pages 1848–1855. IEEE, 2002.
- [42] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1-2):61–84, 2004.
- [43] Dahyun Kang, Sonya S Kwak, Hanbyeol Lee, Eun Ho Kim, and JongSuk Choi. This or that: the effect of robot’s deictic expression on user’s perception. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11383–11390. IEEE, 2020.
- [44] Adam Kendon. Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200, 1994.
- [45] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.
- [46] James M Kilner. More than one pathway to action understanding. *Trends in cognitive sciences*, 15(8):352–357, 2011.
- [47] Heeyoung Kim, Sonya S Kwak, and Myungsuk Kim. Personality design of sociable robots by control of gesture design factors. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 494–499. IEEE, 2008.
- [48] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [49] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

- [50] Maja Kocoń. Idle motion synthesis of human head and face in virtual reality environment. In *Serious Games Development and Applications: 4th International Conference, SGDA 2013, Trondheim, Norway, September 25-27, 2013. Proceedings 4*, pages 299–306. Springer, 2013.
- [51] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*, pages 205–217. Springer, 2006.
- [52] James J Kuffner and Steven M LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 2, pages 995–1001. IEEE, 2000.
- [53] Shikhar Kumar, Eliran Itzhak, Yael Edan, Galit Nimrod, Vardit Sarne-Fleischmann, and Noam Tractinsky. Politeness in human–robot interaction: a multi-experiment study with non-humanoid robots. *International Journal of Social Robotics*, 14(8):1805–1820, 2022.
- [54] I Han Kuo, Joel Marcus Rabindran, Elizabeth Broadbent, Yong In Lee, Ngaire Kerse, Rebecca MQ Stafford, and Bruce A MacDonald. Age and gender factors in user acceptance of healthcare robots. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 214–219. IEEE, 2009.
- [55] Linda Lastrico, Nuno Ferreira Duarte, Alessandro Carfí, Francesco Rea, Alessandra Sciutti, Fulvio Mastrogiovanni, and José Santos-Victor. Expressing and inferring action carefulness in human-to-robot handovers. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9824–9831. IEEE, 2023.
- [56] Quoc Anh Le, Souheil Hanoune, and Catherine Pelachaud. Design and implementation of an expressive gesture model for a humanoid robot. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pages 134–140. IEEE, 2011.

BIBLIOGRAPHY

- [57] Namyoon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. The influence of politeness behavior on user compliance with social robots in a healthcare service setting. *International Journal of Social Robotics*, 9:727–743, 2017.
- [58] Jamy Li and Mark Chignell. Communication of emotion in social robots through simple head and arm movements. *International Journal of Social Robotics*, 3(2):125–142, 2011.
- [59] Christina Lichtenthäler and Alexandra Kirsch. Legibility of robot behavior: A literature review. 2016.
- [60] Phoebe Liu, Dylan F Glas, Takayuki Kanda, and Hiroshi Ishiguro. Data-driven hri: Learning social behaviors by example from human–human interaction. *IEEE Transactions on Robotics*, 32(4):988–1008, 2016.
- [61] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. *arXiv preprint arXiv:2501.18898*, 2025.
- [62] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. Generative expressive robot behaviors using large language models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 482–491, 2024.
- [63] Jim Mainprice, E Akin Sisbot, Thierry Siméon, and Rachid Alami. Planning safe and legible hand-over motions for human-robot interaction. In *IARP workshop on technical challenges for dependable robots in human environments*, volume 2, page 7, 2010.
- [64] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.
- [65] Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559, 2005.

BIBLIOGRAPHY

- [66] Christopher A Miller, Peggy Wu, and Harry B Funk. A computational approach to etiquette: Operationalizing brown and levinson’s politeness model. *IEEE Intelligent Systems*, 23(4):28–35, 2008.
- [67] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.
- [68] Amal Nanavati, Malcolm Doering, Dražen Bršćić, and Takayuki Kanda. Autonomously learning one-to-many social interaction logic from human-human interaction data. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 419–427, 2020.
- [69] Amal Nanavati, Patricia Alves-Oliveira, Tyler Schrenk, Ethan K Gordon, Maya Cakmak, and Siddhartha S Srinivasa. Design principles for robot-assisted feeding in social contexts. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 24–33, 2023.
- [70] Stefanos Nikolaidis, Anca Dragan, and Siddhartha Srinivasa. based legibility optimization. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 271–278. IEEE, 2016.
- [71] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, volume 42, pages 569–596. Wiley Online Library, 2023.
- [72] Xiang Pan, Malcolm Doering, and Takayuki Kanda. What is your other hand doing, robot? a model of behavior for shopkeeper robot’s idle hand. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 552–560, 2024.
- [73] Xiang Pan, Malcolm Doering, Stela Hanbyeol Seo, and Takayuki Kanda. Communicating physical properties through robot object manipulation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 738–746. IEEE, 2025.

BIBLIOGRAPHY

- [74] Pragathi Praveena, Daniel Rakita, Bilge Mutlu, and Michael Gleicher. Supporting perception of weight through motion-induced sensory conflicts in robot teleoperation. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-robot Interaction*, pages 509–517, 2020.
- [75] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(3):171–193, 2002.
- [76] Camilo Perez Quintero, Romeo Tatsambon, Mona Gridseth, and Martin Jägersand. Visual pointing gestures for bi-directional human robot interaction in a pick-and-place task. In *2015 24th IEEE International symposium on robot and human interactive communication (RO-MAN)*, pages 349–354. IEEE, 2015.
- [77] Daniel J Rea, Sebastian Schneider, and Takayuki Kanda. ” is this all you can do? harder!” the effects of (im) polite robot encouragement on exercise effort. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 225–233, 2021.
- [78] Patrizia Ribino. The role of politeness in human–machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, pages 1–38, 2023.
- [79] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In *Proceedings of the 12th ACM international conference on PErvasive technologies related to assistive environments*, pages 247–255, 2019.
- [80] Wolff-Michael Roth. Gestures: Their role in teaching and learning. *Review of educational research*, 71(3):365–392, 2001.
- [81] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4:201–217, 2012.

BIBLIOGRAPHY

- [82] Maha Salem, Micheline Ziadee, and Majd Sakr. Effects of politeness and interaction context on perception and experience of hri. In *Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5*, pages 531–541. Springer, 2013.
- [83] Maha Salem, Micheline Ziadee, and Majd Sakr. Marhaba, how may i help you? effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 74–81, 2014.
- [84] Alessandra Sciutti, Laura Patane, Francesco Nori, and Giulio Sandini. Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development*, 6(2):80–92, 2014.
- [85] Sara Sheikholeslami, AJung Moon, and Elizabeth A Croft. Cooperative gestures for industry: Exploring the efficacy of robot hand configurations in expression of instructional gestures for human–robot interaction. *The International Journal of Robotics Research*, 36(5-7):699–720, 2017.
- [86] Masahiro Shiomi, Takayuki Kanda, Iris Howley, Kotaro Hayashi, and Norihiro Hagita. Can a social robot stimulate science curiosity in classrooms? *International Journal of Social Robotics*, 7:641–652, 2015.
- [87] Emrah Akin Sisbot and Rachid Alami. A human-aware manipulation planner. *IEEE Transactions on Robotics*, 28(5):1045–1057, 2012.
- [88] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters*, 2024.
- [89] Hyunsoo Song, Min Joong Kim, Sang-Hoon Jeong, Hyen-Jeong Suk, and Dong-Soo Kwon. Design of idle motions for service robot via video ethnography. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 195–199. IEEE, 2009.
- [90] Nicolas Spatola, Barbara Kühnlenz, and Gordon Cheng. Perception and evaluation in human–robot interaction: The human–robot interaction eval-

BIBLIOGRAPHY

- uation scale (hries)—a multicomponent approach of anthropomorphism. *International Journal of Social Robotics*, 13(7):1517–1539, 2021.
- [91] Keiko Tagai, Hitomi Shimakura, Sadaki Takata, Masayoshi Nagai, Katsumi Watanabe, Kazuhisa Niki, Sunao Iwaki, and Takatsune Kumada. Differential brain response to one-or two-hand handling action: an fmri study. *Neuroscience and Neuroeconomics*, pages 21–32, 2013.
- [92] Leila Takayama, Doug Dooley, and Wendy Ju. Expressing thought: improving robot readability with animation principles. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 69–76, 2011.
- [93] Yujin Tang, Wenhao Yu, Jie Tan, Heiga Zen, Aleksandra Faust, and Tatsuya Harada. Saytap: Language to quadrupedal locomotion. *arXiv preprint arXiv:2306.07580*, 2023.
- [94] Nikolaus Vahrenkamp, Tamim Asfour, and Rudiger Dillmann. Simultaneous grasp and motion planning: Humanoid robot armar-iii. *IEEE Robotics & Automation Magazine*, 19(2):43–57, 2012.
- [95] Petr Vanc, Jan Kristof Behrens, Karla Stepanova, and Vaclav Hlavac. Communicating human intent to a robotic companion by multi-type gesture sentences. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9839–9845. IEEE, 2023.
- [96] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview, 2014.
- [97] Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. Vlm see, robot do: Human demo video to robot action plan via vision language model. *arXiv preprint arXiv:2410.08792*, 2024.
- [98] Adam Waytz, Nicholas Epley, and John T Cacioppo. Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current directions in psychological science*, 19(1):58–62, 2010.
- [99] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019.

BIBLIOGRAPHY

- [100] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.
- [101] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

SELECTED LIST OF PUBLICATIONS

- **International Conferences**

- [1] Xiang Pan, Malcolm Doering, and Takayuki Kanda. What Is Your Other Hand Doing, Robot? A Model of Behavior for Shopkeeper Robot's Idle Hand. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI'24)*, 552-560. <https://doi.org/10.1145/3610977.3634986>, **Chapter 2**
- [2] Xiang Pan, Malcolm Doering, Stela H. Seo, and Takayuki Kanda. Communicating Physical Properties through Robot Object Manipulation. *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI'25)*, 738-746. <https://doi.org/10.1109/HRI61500.2025.10973989>, **Chapter 3**
- [3] Xiang Pan, Malcolm Doering, and Takayuki Kanda. Communicating Object Relations through Robot Gestures. (*Accepted by the 2026 ACM/IEEE International Conference on Human-Robot Interaction*), **Chapter 4**