Communicating Object Relations through Robot Gestures

Xiang Pan
Kyoto University
Kyoto, Japan
pan@robot.soc.i.kyoto-u.ac.jp

Malcolm Doering
Kyoto University
Kyoto, Japan
doering@robot.soc.i.kyoto-u.ac.jp

Takayuki Kanda
Kyoto University
Kyoto, Japan
kanda@i.kyoto-u.ac.jp

Abstract

We proposed a system for generating relational gestures that convey semantic relations such as similarity and difference between two objects. To understand how humans naturally express such relations, we conducted an observational study with experienced shopkeepers as they frequently compare objects using both speech and gestures. Through analysis of their interactions, we identified four common types of object relations and extracted representative gesture patterns for each. For example, similarity was often conveyed through synchronized hand movements bringing both hands closer together, accompanied by alternating gaze between two objects. Based on these findings, we developed a gesture generation system in which one large language model (LLM) infers the intended object relation from utterance text, and another LLM adapts gestures from a co-speech gesture system that aligns them with speech, integrating relational cues without disrupting this alignment. These modified gestures were automatically mapped onto a dual-arm robot. We evaluated the system through two user studies. In the first study, 20 participants were asked to identify object relations from 24 relational gestures performed by the robot without accompanying speech. They correctly identified the intended relations with an average accuracy of 89.8% across all relation types. In the second study, another 20 participants compared two robot conditions in a within-subjects design: one with relational gestures and one without. Results showed that the robot using relational gestures was perceived as more competent, sociable, and animate compared to the robot without them.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computer systems organization** → **Robotics**.

Keywords

Non-verbal communication, large language models, relational gestures, object relations

ACM Reference Format:
Xiang Pan, Malcolm Doering, and Takayuki Kanda. 2026. Communicating Object Relations through Robot Gestures. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3757279.3785554

HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3785554

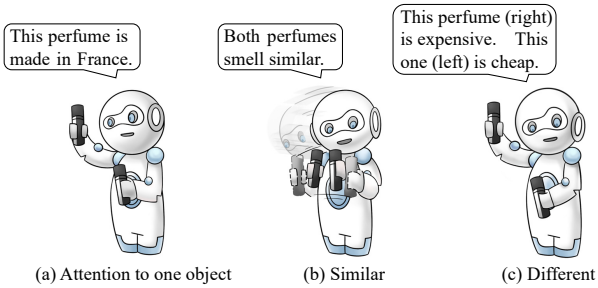


Figure 1: Robot gestures during object comparison.

1 Introduction

Gestures play a fundamental role in human communication. Across a wide range of social settings from conversation [21] and storytelling [29] to formal instruction [3, 36], people naturally accompany their speech with gestures. These gestures help draw attention, reinforce or supplement spoken content, convey semantic content, and visualize spatial or abstract concepts [2, 15, 21, 29].

Among the diverse types of gestures, one important category involves holding and moving two physical objects while describing their semantic relations, such as similarity or 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, they reinforce verbal content while visually clarifying relational 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 (HRI). For robots designed 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. 1 (a)), moving both hands closer while alternating gaze (Fig. 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. 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.

2 Related Work

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 [22]. Gestures are commonly categorized into iconic, metaphoric, deictic, and beat gestures [29], each serving distinct communicative functions [2, 15, 21, 29]. Recent work has examined whether the semantic congruency between representational gestures and lexical words is evaluated similarly when words are conveyed through text versus speech [18]. Existing work has also shown that gestures can represent conceptual distinctions, for example by maintaining spatial separation to express contrast or using shared spatial locations to convey similarity [22, 46]. Such gestures use concrete physical actions to represent abstract concepts and fall within the broader scope of metaphoric gestures. While these studies demonstrate that gestures can refer to objects or encode conceptual relations, the use of gestures to communicate semantic relations between objects remains underexplored.

2.2 Gestures in Human-Robot Interaction

Gestures serve as an essential non-verbal modality in HRI, encompassing hand movements, facial expressions, gaze shifts, head motions, and full-body movements [35]. Prior research has widely recognized the importance of gestures in making robot communication more legible, expressive, and socially appropriate [11, 12, 14, 25, 26]. Recent research also focuses on conveying object properties through physical interactions with one object [33, 37]. Furthermore, supportive gestures toward a single object can increase perceived politeness and competence in service contexts [32]. While these studies provide valuable insights into gesture interaction with one 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[41]. Yet, the communication of semantic relations between objects through robot gestures remains underexplored.

2.3 Co-speech Gesture Generation

Co-speech gestures are visible actions produced while speaking [42], and their generation therefore focuses on producing human body movements aligned with speech input [7]. Existing approaches can be broadly divided into rule-based and data-driven methods [6, 24, 31]. However, deterministic generative models often yield oversmoothed gestures [4, 44] due to their inability to handle many-to-many mappings. Recent advances therefore adopt probabilistic generative models, such as diffusion-based gesture synthesis [7, 27]. In parallel, LLM-driven methods incorporate explicit semantic reasoning, either via gesture retrieval from a self-built gesture library [47] or prompt-based gesture description [34]. Despite these advances, the encoding of semantic relations between objects through gestures remains underexplored.

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 [30, 40] to high-level planning [1, 17]. VLMs are also leveraged to interpret social context, enabling a mobile robot to plan socially aware paths [38] or generate task plans for robots [43]. Building on this, VLAs leverage robotics data to enable direct robot control [13, 23, 48]. In addition, LLMs are explored for robot gesture generation such as head gestures [28] and hand gestures [16]. While these studies address robot control or gesture generation in isolation, the generation of robot gestures that convey object relations remains underexplored.

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.

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.

3.1.1 ***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.

3.1.2 ***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.

3.1.3 ***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.

3.1.4 ***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 variety 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**: Expresses 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.

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. 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.

3.2.1 ***Hand Pattern.*** This feature captures how both hands are coordinated when conveying object relations:

- **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.

Table 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 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

- **Synchronized**: Both hands move simultaneously 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. 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.

3.2.2 ***Hand Direction.*** This feature captures the dominant axis along which the hands move:

- **Vertical**: Upward or downward movements. For example, the shopkeeper raises the left hand holding a handbag.
- **Horizontal**: Side-to-side or inward–outward movements. For example, the shopkeeper swings both hands holding handbags from left to right.
- **Other**: All other cases.

Similarly, Tab. 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.

3.2.3 ***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.
- **Both-target**: The head is oriented to encompass both hands simultaneously. For example, the shopkeeper slightly tilts

Table 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

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 a smartwatch.
- **Other**: All other cases.

Likewise, as shown in Tab. 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).

3.2.4 ***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. 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).
- **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 hand (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 Relational Gesture Generation System

4.1 Overview

We propose an LLM-based system that enables robots to compare objects through co-speech relational gestures. Fig. 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 relational gesture plan with the co-speech gestures, integrating relational cues while preserving the synchrony between speech and gestures.

(a) One-object (b) No-relation (c) Similar (d) Different

Figure 2: Representative gestures for each object relation.

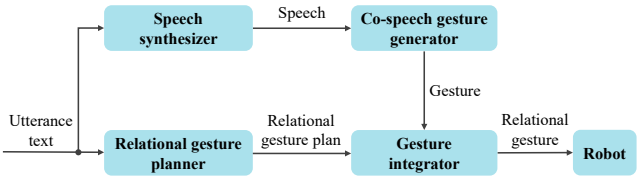


Figure 3: Architecture of the proposed system.

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).

4.2 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* [7] to instantiate this module, as it represents a state-of-the-art model for producing full-body gestures synchronized with speech.

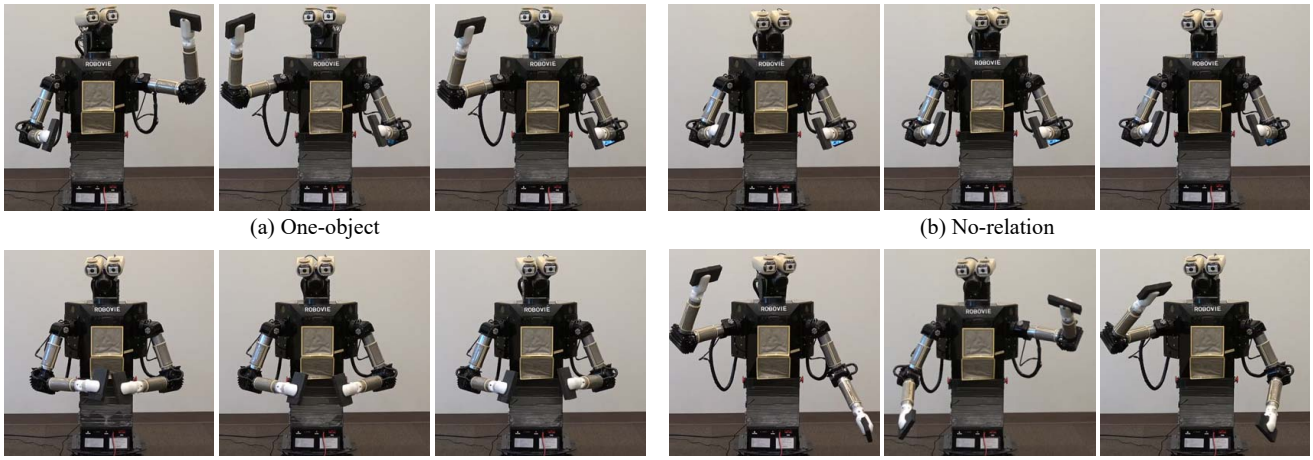
For generalizability across robotic platforms, we focus only on seven upper-body joints: the head, shoulders, elbows, and wrists. These joints are most relevant for relational gestures and ensure that our system can operate on robots with varying degrees of freedom. The extracted sequences of seven joints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

4.3 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 [1]) through prompt-based inference.

The planner operates in a single end-to-end pipeline, and the representative gesture descriptions identified in Sec. 3.2.4 serve as the main knowledge source for the LLM prompt. Specifically, the

[1]https://platform.openai.com/docs/models/gpt-4.1



(a) One-object (b) No-relation

(c) Similar (d) Different

Figure 4: Examples of robot relational gestures.

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. 3.1.4);
- 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."

4.4 Gesture Integrator

Once a relational gesture plan is obtained, the gesture integrator integrates it with the full-body co-speech gestures. 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 needed 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 [16]. 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[2] for this module.

4.5 Robot

We used Robovie [19], 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 head and arm orientations and automatically map these joint angles to the robot.

5 Study 1: 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.

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 were compensated with 4000 JPY.

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. 3.1.2. For each object pair, our system generated

[2]https://platform.openai.com/docs/models/o3

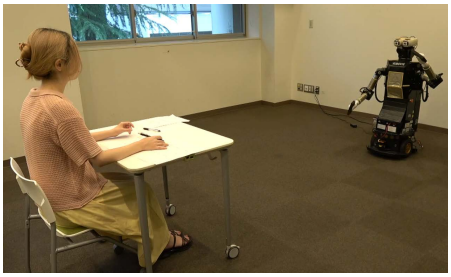


Figure 5: Setup for the experiment in Study 1.

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.

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. 5. Each participant completed the task individually. They observed 24 robot-performed gestures without speech (Sec. 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. 3.1.4) 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**: Expresses the similarity between the two objects.
- **Different**: Expresses the difference between the two objects.

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

5.4 Results

5.4.1 ***Recognition Results.*** Tab. 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.

Table 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**

5.4.2 ***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.

6 Study 2: 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 [7]. 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.

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 [9, 32]. In our study, relational gestures convey semantic relations between objects and can therefore be considered meaningful motion. This leads to 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 verbal or non-verbal modalities [20]. Given that Study 1 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 [10]. Moreover, robots exhibiting a variety of congruent gestures are often perceived as more human-like compared to those with limited or inconsistent gestures [45]. Since relational gestures are modeled from shopkeeper behaviors and offer greater diversity than co-speech gestures. Accordingly, we made the following prediction:

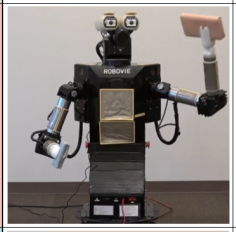
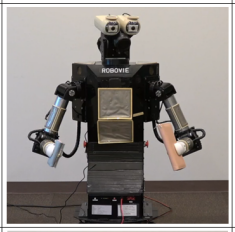
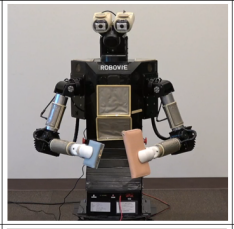
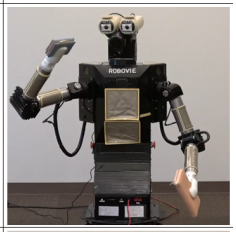
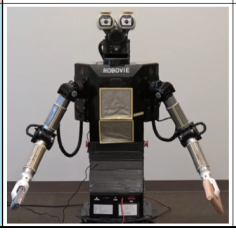
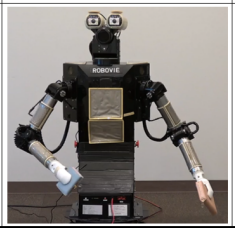
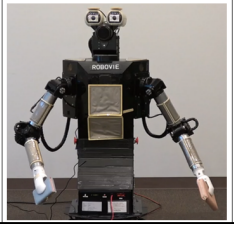
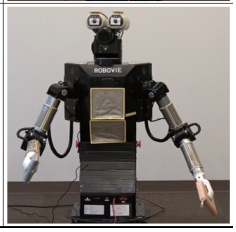
Relation	One-object	No-relation	Similar	Different
Utterance	In my left hand is a long pink wallet.	I would like to introduce two classical wallets.	Both wallets are made in France.	The long wallet is heavy. The short one is light.
Proposed				
Baseline				

Figure 6: Examples of the robot's behavior in Study 2.

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

6.2 Participants

We recruited 20 participants based on a priori power analysis (Cohen's d = 0.8) [8], 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 were compensated with 4000 JPY.

6.3 Conditions

The robot's performance was compared under two conditions:

- **Proposed**: The robot operates with the gestures generated from the proposed system described in Section 4.
- **Baseline**: The module of the relational gesture planner is excluded from the architecture of the proposed system (Fig. 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.

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. 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.

6.5 Measurement

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

- **Competence**: Measured using six RoSAS items[5] - *capable*, *responsive*, *interactive*, *reliable*, *competent*, and *knowledgeable*.
- **Sociability**: Measured using four HRIES items[39] - *warm*, *likeable*, *trustworthy*, and *friendly*.
- **Animacy**: Measured using four HRIES items - *alive*, *natural*, *real*, and *human-like*.

6.6 Results

6.6.1 *Verification of Predictions.*

- **Competence**: As illustrated in the first set of bars in Fig. 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**.

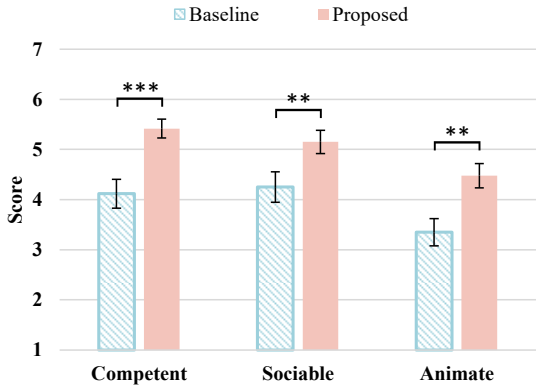


Figure 7: Results from our quantitative measures. The error bars show the standard error for the mean at ±1 *SE*. ($** : p < .01, *** : p < .001$)

- **Sociability**: As shown in the second set of bars of Fig. 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. 7, animacy scores were averaged across the four corresponding items. Normality was also confirmed ($W_{baseline}$ = .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**.

6.6.2 ***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 expressed 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, expressing a preference for robots that are "not too clever," as this maintained a clearer boundary between humans and robots.

7 Discussion

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.

7.2 Limitations and Future Works

While our results demonstrate the effectiveness of relational gestures, several limitations should be acknowledged. First, our data collection focused on retail scenarios with professional shopkeepers comparing hand-held objects. Although the participants had extensive sales experience and the selected object pairs covered common everyday items with varied features, the observed gestures may not fully represent relational gestures used in other domains such as education or collaborative work. Second, our system prioritizes the most frequently observed gestures, which supports reliable evaluation but reduces the variability and expressiveness typical of natural human gesturing. Finally, the system is not yet optimized for fully real-time use, as LLM-based gesture integration can introduce latency.

Accordingly, future work should extend data collection to more diverse participant groups, domains, and interaction scenarios, and explore richer gesture variants. In addition, system responsiveness should be improved through more efficient gesture integration and advances in lightweight LLMs.

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.

Acknowledgments

This work was supported by JST Moonshot R and D under Grant Number JPMJMS2011, Japan, and JSPS KAKENHI under Grant Number 24H00722, Japan.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691* (2022). doi:10.48550/arXiv.2204.01691

[2] Martha W Alibali. 2005. Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial cognition and computation* 5, 4 (2005), 307–331. doi:10.1207/s15427633scc0504_2

[3] Martha W Alibali and Mitchell J Nathan. 2014. Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. In *Video research in the learning sciences*. Routledge, 349–365.

[4] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036. doi:10.1145/3474085.3475223

[5] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262. doi:10.1145/2909824.3020208

[6] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. 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*. 413–420. doi:10.1145/192161.192272

[7] Bohong Chen, Yumeng Li, Yao-Xiang Ding, Tianjia Shao, and Kun Zhou. 2024. Enabling synergistic full-body control in prompt-based co-speech motion generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 6774–6783. doi:10.1145/3664647.3680847

[8] Jacob Cohen. 1992. Quantitative methods in psychology: A power primer. *Psychol. Bull.* 112 (1992), 1155–1159. doi:10.1037/0033-2909.112.1.155

[9] Raymond H Cuijpers and Marco AMH Knops. 2015. Motions of robots matter! the social effects of idle and meaningful motions. In *International Conference on Social Robotics*. Springer, 174–183. doi:10.1007/978-3-319-25554-5_18

[10] Jan De Wit, Paul Vogt, and Emiel Krahmer. 2023. The design and observed effects of robot-performed manual gestures: A systematic review. *ACM Transactions on Human-Robot Interaction* 12, 1 (2023), 1–62. doi:10.1145/3549530

[11] Anca D Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S Srinivasa. 2015. Effects of robot motion on human-robot collaboration. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 51–58. doi:10.1145/2696454.2696473

[12] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308. doi:10.1109/HRI.2013.6483603

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*.

[14] Michael J Gielniak and Andrea L Thomaz. 2011. Generating anticipation in robot motion. In *2011 RO-MAN*. IEEE, 449–454. doi:10.1109/ROMAN.2011.6005255

[15] Christian Heath. 1992. Gesture's discreet tasks: Multiple relevancies in visual conduct and in the contextualisation of language. (1992). doi:10.1075/pbns.22.08hea

[16] Peide Huang, Yuhan Hu, Nataliya Nechyporenko, Daehwa Kim, Walter Talbott, and Jian Zhang. 2025. Emotion: Expressive motion sequence generation for humanoid robots with in-context learning. *IEEE Robotics and Automation Letters* (2025). doi:10.1109/LRA.2025.3575983

[17] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*.

[18] Sarah S Hughes-Berheim, Laura M Morett, and Raymond Bulger. 2020. Semantic relationships between representational gestures and their lexical affiliates are evaluated similarly for speech and text. *Frontiers in psychology* 11 (2020), 575991. doi:10.3389/fpsyg.2020.575991

[19] Takayuki Kanda, Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, and Ryohei Nakatsu. 2002. Development and evaluation of an interactive humanoid robot" Robovie". In *Proceedings 2002 IEEE international conference on robotics and automation (Cat. No. 02CH37292)*, Vol. 2. IEEE, 1848–1855. doi:10.1109/ROBOT.2002.1014810

[20] Dahyun Kang, Sonya S Kwak, Hanbyeol Lee, Eun Ho Kim, and JongSuk Choi. 2020. 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)*. IEEE, 11383–11390. doi:10.1109/IROS45743.2020.9341067

[21] Adam Kendon. 1994. Do gestures communicate? A review. *Research on language and social interaction* 27, 3 (1994), 175–200. doi:10.1207/s15327973rlsi2703_2

[22] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press. doi:10.1017/CBO9780511807572

[23] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246* (2024). doi:10.48550/arXiv.2406.09246

[24] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217. doi:10.1007/11821830_17

[25] Quoc Anh Le, Souheil Hanoune, and Catherine Pelachaud. 2011. Design and implementation of an expressive gesture model for a humanoid robot. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 134–140. doi:10.1109/HUMANOIDS.2011.6100857

[26] Jamy Li and Mark Chignell. 2011. Communication of emotion in social robots through simple head and arm movements. *International Journal of Social Robotics* 3, 2 (2011), 125–142. doi:10.1007/s12369-010-0071-x

[27] Pinxin Liu, Luchuan Song, Junhua Huang, and Chenliang Xu. 2025. GestureLSM: Latent Shortcut based Co-Speech Gesture Generation with Spatial-Temporal Modeling. In *IEEE/CVF International Conference on Computer Vision*.

[28] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative expressive robot behaviors using large language models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 482–491. doi:10.1145/3610977.3634999

[29] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[30] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large Language Models as General Pattern Machines. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*.

[31] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A comprehensive review of data-driven co-speech gesture generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596. doi:10.1111/cgf.14776

[32] Xiang Pan, Malcolm Doering, and Takayuki Kanda. 2024. 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*. 552–560. doi:10.1145/3610977.3634986

[33] Xiang Pan, Malcolm Doering, Stela Hanbyeol Seo, and Takayuki Kanda. 2025. Communicating Physical Properties Through Robot Object Manipulation. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 738–746. doi:10.1109/HRI61500.2025.10973989

[34] Haozhou Pang, Tianwei Ding, Lanshan He, Ming Tao, Lu Zhang, and Qi Gan. 2025. LLM Gesticulator: leveraging large language models for scalable and controllable co-speech gesture synthesis. In *Eighth International Conference on Computer Graphics and Virtuality (ICCGV 2025)*, Vol. 13557. SPIE, 1355702. doi:10.1117/12.3060395

[35] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 9, 3 (2002), 171–193. doi:10.1145/568513.568514

[36] Wolff-Michael Roth. 2001. Gestures: Their role in teaching and learning. *Review of educational research* 71, 3 (2001), 365–392. doi:10.3102/00346543071003365

[37] Alessandra Sciutti, Laura Patane, Francesco Nori, and Giulio Sandini. 2014. Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development* 6, 2 (2014), 80–92. doi:10.1109/TAMD.2014.2312399

[38] Daeun Song, Jing Liang, Amirreza Payandeh, Amir Hossain Raj, Xuesu Xiao, and Dinesh Manocha. 2024. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models. *IEEE Robotics and Automation Letters* (2024). doi:10.1109/LRA.2024.3511409

[39] Nicolas Spatola, Barbara Kühnlenz, and Gordon Cheng. 2021. Perception and evaluation in human–robot interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A multicomponent approach of anthropomorphism. *International Journal of Social Robotics* 13, 7 (2021), 1517–1539. doi:10.1007/s12369-020-00667-4

[40] Yujin Tang, Wenhao Yu, Jie Tan, Heiga Zen, Aleksandra Faust, and Tatsuya Harada. 2023. Saytap: Language to quadrupedal locomotion. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*.

[41] Petr Vanc, Jan Kristof Behrens, Karla Stepanova, and Vaclav Hlavac. 2023. Communicating human intent to a robotic companion by multi-type gesture sentences. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9839–9845. doi:10.1109/IROS55552.2023.10341944

[42] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech communication* 57 (2014), 209–232. doi:10.1016/j.specom.2013.09.008

[43] Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. 2024. Vlm see, robot do: Human demo video to robot action plan via vision language model. *arXiv preprint arXiv:2410.08792* (2024). doi:10.48550/arXiv.2410.08792

[44] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16. doi:10.1145/3414685.3417838

[45] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. 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)*. IEEE, 4303–4309. doi:10.1109/ICRA.2019.8793720

[46] Icy Zhang, Tina Izad, and Erica A Cartmill. 2024. Embodying similarity and difference: The effect of listing and contrasting gestures during US political speech. *Cognitive Science* 48, 3 (2024), e13428. doi:10.1111/cogs.13428

[47] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–17. doi:10.1145/3658134

[48] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*.

Received 2025-09-30; accepted 2025-12-01