



## Review

# Humanoid dexterous hands from structure to gesture semantics for enhanced human–robot interaction: A review

Xin Li <sup>a,b</sup>, Wenfu Xu <sup>a,1</sup>, Zaiqiao Ye <sup>b</sup>, Han Yuan <sup>a,\*</sup>

<sup>a</sup> School of Robotics and Advanced Manufacture, Harbin Institute of Technology, Shenzhen 518055, China

<sup>b</sup> Future Design School, Harbin Institute of Technology, Shenzhen 518055, China

## ARTICLE INFO

## Article history:

Received 28 May 2025

Revised 28 June 2025

Accepted 28 July 2025

Available online 20 August 2025

## Keywords:

Human–robot interaction (HRI)

Dexterous hand

Large language models

Gesture

Communication

## ABSTRACT

As human–robot interaction (HRI) technology advances, dexterous robotic hands are playing a dual role—serving both as tools for manipulation and as channels for non-verbal communication. While much of the existing research emphasizes improving grasping and structural dexterity, the semantic dimension of gestures and its impact on user experience has been relatively overlooked. Studies from HRI and cognitive psychology consistently show that the naturalness and cognitive empathy of gestures significantly influence user trust, satisfaction, and engagement. This shift reflects a broader transition from mechanically driven designs toward cognitively empathic interactions—robots' ability to infer human affect, intent, and social context to generate appropriate nonverbal responses. In this paper, we argue that large language models (LLMs) enable a paradigm shift in gesture control—from rule-based execution to semantic-driven, context-aware generation. By leveraging LLMs and visual-language models, robots can interpret environmental and social cues, dynamically map emotions, and generate gestures aligned with human communication norms. We conducted a comprehensive review of research in dexterous hand mechanics, gesture semantics, and user experience evaluation, integrating insights from linguistics and cognitive science. Furthermore, we propose a closed-loop framework—"perception–cognition–generation–assessment"—to guide gesture design through iterative, multimodal feedback. This framework lays the conceptual foundation for building universal, adaptive, and emotionally intelligent gesture systems in future human–robot interaction.

© 2025 The Author(s). Published by Elsevier B.V. on behalf of Shandong University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

Current research on dexterous hands emphasizes their superior performance in grasping tasks [1]; however, the experience and interaction skills of dexterous hands are equally important. With the development of human–robot interaction (HRI) technology, dexterous hands, as the primary medium for non-verbal communication for robots, have gradually become a research hotspot in terms of their impact on user experience. People often adopt gestural behaviors to assist, enhance, or enrich semantic expressions during daily communication or speech. Some studies have shown that appropriate gestures can effectively enhance people's understanding and memory in speech and other scenes [2]. Cohen et al. found that more than 80% of gestures in communication convey one or more pieces of information that are not expressed by language, thus supplementing the information that language cannot effectively or succinctly convey,

including the speed of movement, direction, mode, size of the object, spatial relationship, etc [3]. Meanwhile, some researchers have suggested that dexterous hands' gestures can serve as a tool for expressing emotions in human–robot communication. Jamy Li et al. experimentally demonstrated that dexterous hands may express feelings, such as warmth and comfort, through particular gestures, hence enhancing the user's interactive experience [4]. Although dexterous hands currently excel at industrial tasks, we also want robots to engage with individuals in various unstructured environments, such as domestic settings. Existing research focuses on grasping, localization, and perception of dexterous hands. To seamlessly integrate anthropomorphic hands into human-centered applications and foster more harmonious human–robot interactions, we must focus more on user experience in HRI. Therefore, in this work, we review and summarize the role of dexterous hands in human–robot interaction (see Fig. 1).

Robots can enhance naturalness in the delivery of objects or everyday interactions with users by borrowing from the way objects are exchanged between people [5]. In retail environments, it has been found that robots enhance the customer experience by mimicking the "supportive" hand movements of human

\* Corresponding author.

E-mail address: [yuanhan@hit.edu.cn](mailto:yuanhan@hit.edu.cn) (H. Yuan).

<sup>1</sup> Given his role as Associate Editor of this journal, Wenfu Xu had no involvement in the peer-review of this article and had no access to information regarding its peer-review. This article was handled by Prof. Max Q.-H. Meng.

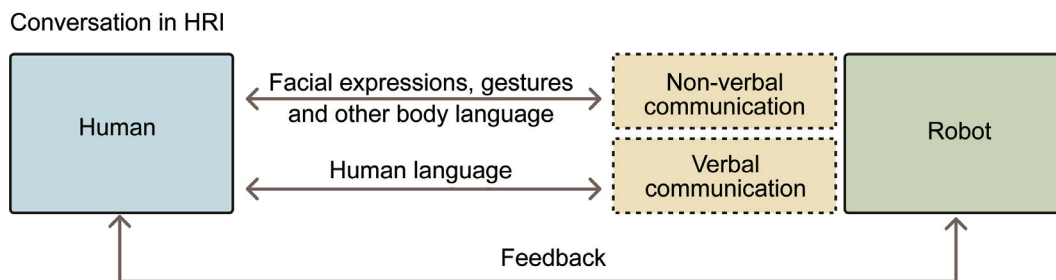


Fig. 1. Interaction information flow between human and humanoid robots.

shopkeepers, making their behavior seem more polite and competent [6]. The ability of robots to recognize and express emotions is key to enhancing the user interaction experience, but relatively little research has been conducted on human–robot interaction of dexterous hands. Research on gesture robot interaction design needs to enhance the naturalness of human–robot interaction experience by exploring the application law of gesture interaction in communication, basing it on user experience, and exploring the gesture interaction design of robots in combination with related disciplines such as cognitive psychology, gesture semantics, and ergonomics.

An analysis of the titles, keywords, and abstracts from the Web of Science database over the past five years shown in Fig. 2, utilizing the keyword “dexterous hands” via VOSviewer, reveals that the predominant research directions are centered on reinforcement learning, grasping, and manipulation. This review systematically compiles the research progress in the field of dexterous hands for humanoid robots. The existing literature emphasizes grasp stability analysis, manipulation trajectory planning, and reinforcement learning control strategies, achieving notable advancements in physical interactions; however, there remains a relative deficiency in research concerning gesture interactions within human–robot interaction (HRI).

This review discusses the innovative value of dexterous robotic hands as a nonverbal communication modality for enhancing user experience. In prototypical application scenarios—such as social robots and service robots—gesture-based interaction has transcended mere mechanical operation to become a vital vehicle for emotional expression, intention signaling, and cultural etiquette. We systematically trace the research trajectory in the field of anthropomorphic dexterous hands: existing advances have already surmounted key technical challenges, including biomimetic multi-joint design, high-precision tactile sensing networks, and adaptive grasping strategy. However, significant gaps remain in the semantic-bionics dimension of dynamic gesture interaction. This review proposes user-experience-oriented design guidelines for dexterous hands, effecting a paradigm shift from “functional implementation” to “cognitive empathy”. These theoretical breakthroughs offer crucial support for the development of next-generation humanoid robots endowed with social intelligence. Fig. 3 illustrates the logical structure of this literature review. We will describe how anthropomorphic dexterous hands have developed over time and ultimately emphasize the importance of current research on semantic gesture expression in dexterous hands.

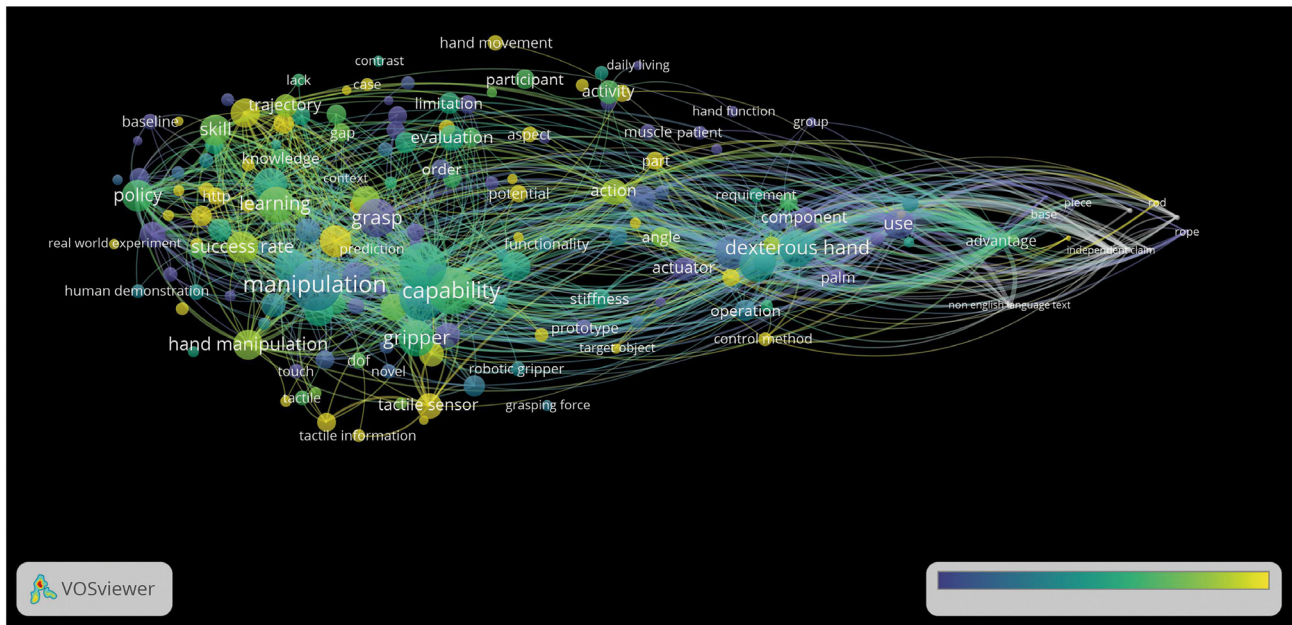
## 2. The evolution of humanoid dexterity

The evolution of humanoid dexterity has been driven by a synergistic progression across structural bionics, perceptual integration, cognitive generalization, and interactive expressiveness, as mapped in the Development Roadmap of Humanoid Five-Finger Dexterous Hands (see Fig. 4). This journey reflects a paradigm

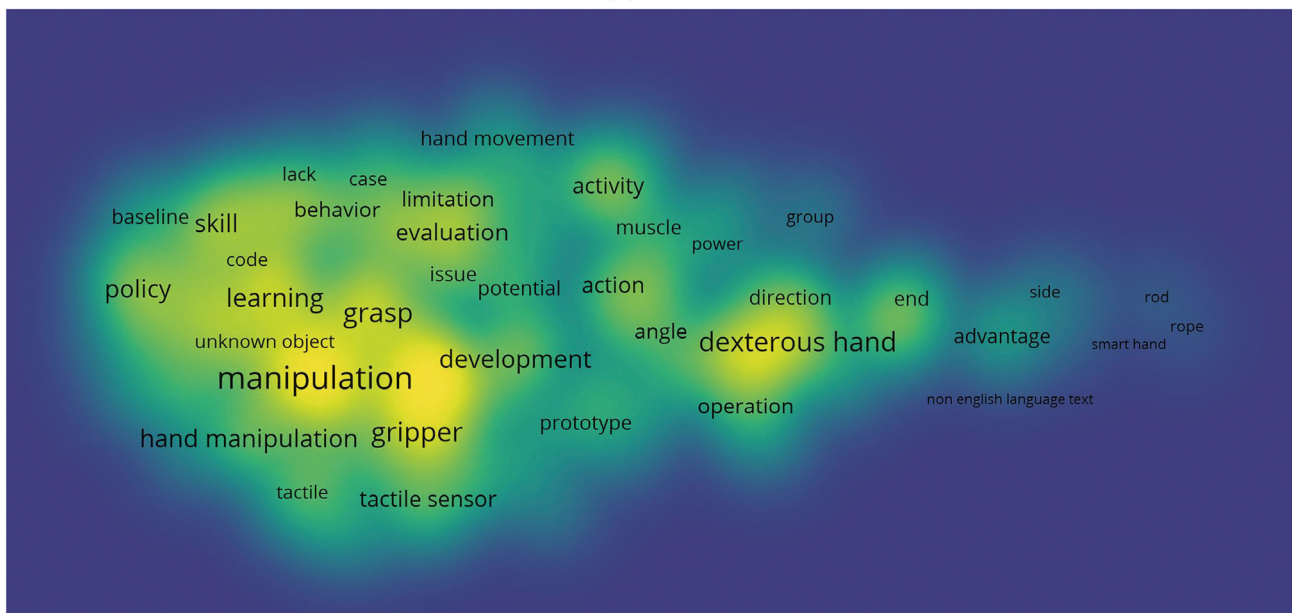
shift from mechanical replication of human hand anatomy to functional optimization, multimodal environmental adaptation, and socially-aware interaction capabilities. Early advancements in structural bionics enabled anthropomorphic designs with tendon-driven or linkage-based actuation, achieving human-like kinematics and tool-use proficiency. Subsequent breakthroughs in perceptual bionics fused tactile, visual, and haptic sensing to mimic biological sensory systems, while decision-making bionics leveraged imitation learning and reinforcement algorithms to bridge imitation and generalization. Ultimately, interaction bionics expanded dexterity beyond physical tasks into semantic gesture representation and emotional communication, marking the transition of dexterous hands from functional tools to socially intelligent partners. The following sections systematically dissect these interconnected advancements — from foundational structural innovations to the emerging frontiers of perceptual, cognitive, and interactive bionics — to unravel how layered biomimetic strategies collectively propel dexterity toward human-like adaptability and collaborative intelligence.

### 2.1. Structural bionics: from mechanical reproduction to functional optimization

The key to realizing human-like dexterity lies in the development of anthropomorphic five-fingered manipulators [7]. The development of dexterous hands in terms of structure and function has gone through an evolutionary process from simple to complex, and the core logic lies in improving the manipulator’s operational flexibility and environmental adaptability through bionic design. The current dexterous hand can already do the movement very close to the human hand; for example, Tesla’s Optimus Gen3 dexterous hand has 22 degrees of freedom, which is very close to the degrees of freedom of the human hand. Currently, academics focus on the research of humanoid five-finger dexterous hands, such as Shadow Dexterous Hand [8], DLR/HIT Hand II [9], ILDA hand [10], etc., because it highly reproduces the anatomical features of human hand in its structure: five independent fingers with 27 skeletal joints in the topology, and 20+ DoF synergistic motion is realized by tendon drive or linkage mechanism [11]. This design not only meets the requirements of anthropomorphism but also realizes dexterity at the functional level. For example, the thumb’s palm-to-palm motion gives the manipulator precise pinching functions, while the kinetic model of multi-finger synergy supports complex operations such as tool use and two-handed robotic arm coordination [12]. It has been shown that anthropomorphic design of five-finger structures can effectively inherit human tool-use experience and reduce task migration costs by sharing human infrastructure (e.g., handle size, tool interface) [13]. Studies have also been conducted to replicate the structure and function of the human hand through biomechanical modeling [14] and synergistic control strategies (a synergy-based framework) [15]. The breakthrough of structural bionics lays the foundation for physical interaction of dexterous



(a)



(b)

**Fig. 2.** Research trends (a) and heat map (b) in the past five years centered on a Web of Science search for the keyword “dexterous hands”.

hands, but the full realization of its function still needs to rely on the synergistic evolution of multimodal sensing ability. Thus, perceptual bionics becomes the next key link for the humanoid dexterous hand to advance towards environmental adaptation.

*2.2. Perceptual bionics: synergistic evolution of multimodal sensors*

Structural bionics provides a physical foundation for perceptual bionics, which endows dexterous hands with human-like environment perception through multimodal sensor fusion. The perception techniques for robotic multi-fingered dexterous hands can be categorized into internal and external perception [16]. Similar to traditional robots, internal perception mainly detects

motion parameters such as joint position, velocity, and acceleration and feeds them back to the controller. In recent years, the development of sensing technologies and devices for humanoid skin and nervous systems has received extensive attention, and e-skin [17–19] has become an ideal solution for multi-fingered dexterous hand sensing systems due to its high-density distributed tactile sensing capability. External sensing technology, on the other hand, realizes environmental interaction through multimodal sensor fusion: visual sensing obtains object position information through depth cameras [20,21]; haptic sensing employs flexible capacitive/piezoresistive arrays [22] to capture contact force distribution in real time, providing human-like multi-dimensional sensing support for complex operational tasks.

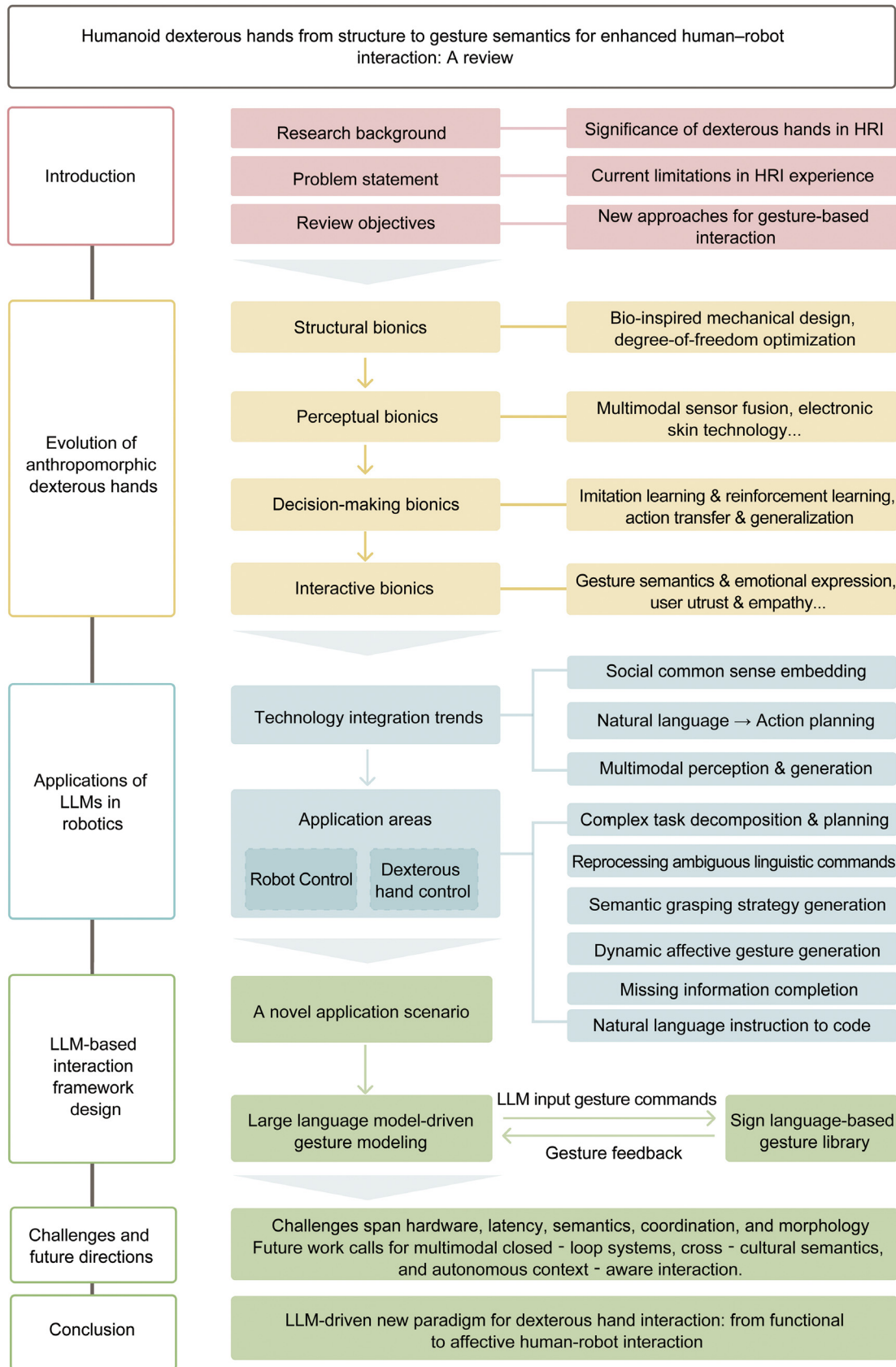


Fig. 3. The writing framework.

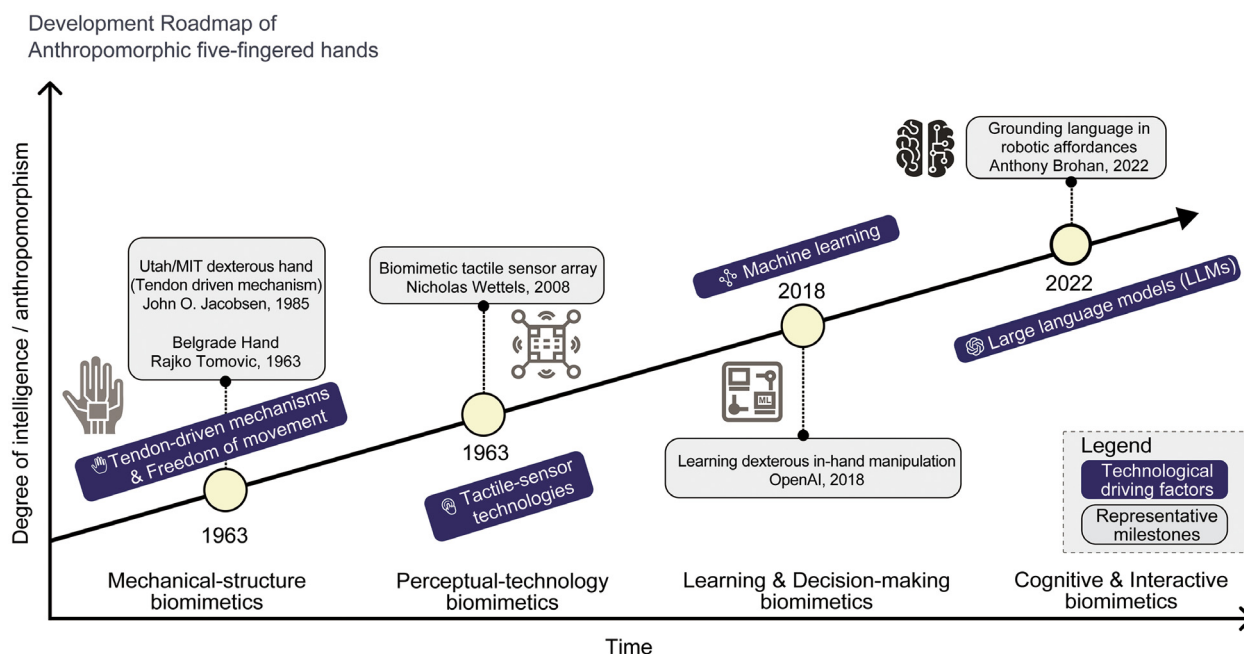


Fig. 4. Development Roadmap of Humanoid Five-Finger Dexterous Hands.

### 2.3. Decision-making bionics: from imitative learning to capability generalization

Current dexterous hands have driven the development of humanization in dexterous hand manipulation by studying human grasping through imitative learning [23]. In humanoid research on grasping strategies, researchers often collect human hand movement data through motion capture systems [24,25], which are converted into robot control commands to guide strategy learning. In recent years, vision-based imitation learning methods have achieved cross-morphological action migration from human hand movements to high degree of freedom robot hands by extracting 3D hand-object interaction gestures from large-scale human videos [26,27], combined with optimization-driven motion reorientation techniques [28,29]. Compared to the limited demonstration data collected by traditional VR devices, video-based imitation learning significantly reduces the data collection cost while supporting behavioral diversity for complex tasks [30, 31]. At the level of policy generation, researchers fused state-action demonstrations with reinforcement learning objectives to significantly improve the sample efficiency and generalization ability of complex manipulation tasks through algorithms such as Generative Adversarial Imitation Learning (GAIL) [32] and Demonstration Augmented Policy Gradient (DAPG) [33,34]. All of the above studies provide a bionic foundation for the generalization ability of intelligent decision-making of humanoid dexterous hands, and with the technical accumulation of structural, perceptual, and decision-making bionics, interaction bionics focuses on the in-depth bionicization of gesture semantics and emotion expression. The breakthrough in this dimension will promote the paradigm shift of dexterous hands from functional tools to socialization partners.

### 2.4. Interaction bionics: synergistic evolution of expressive human-robot interaction and semantic representation of gestures

Although dexterous hands have realized a high degree of bionics in multiple dimensions, such as mechanical structure, sensing mechanism, and intelligent decision-making, the current research gap in the bionic development of dexterous hands lies in

the bionic design of robot gesture interaction. In the interaction process of humanoid robots, different gestures convey nuanced information to strengthen collaboration by expressing functional states (e.g., actions, intentions, or emotions) [35–38], which in turn affect human psychological feelings [35,39,40], strengthen emotional communication [41], and enhance the effectiveness of semantic information that is most effectively retained [42].

At the level of the impact of the naturalness of gesture expression on the interaction experience, Saunderson and Nejat [43] revealed a strong correlation between non-verbal behavior and user trust, and Zabala et al. [44] demonstrated that the expression of body language reflects the personality of the robot, findings that emphasize the impact of the naturalness of gesture expression on the human perceptual level. In response to the cognitive empathy turn in gesture interaction, the Generative Adversarial Network gesture model proposed by Rodriguez et al. [45] and the pattern inference capabilities of large language models (LLMs) explored by Mirchandani et al. [46] exemplify the technological evolutionary path from mechanical trajectory replication to semantic understanding. Roy et al. [47] further advance the integration of gesture interaction with LLMs and affective computing by proposing a GPT-4-driven framework that maps abstract states (e.g., confused or waiting) to parameterized expressive motions (e.g., body tilt, speed), directly linking affective computing to gesture generation. Marmpena et al. proposed a data-driven framework for generating emotional body language in humanoid robots, achieving lifelike, context-relevant expressions that were perceived as equally anthropomorphic and emotionally expressive as hand-designed motions [48]. In addition, Ekman and Friesen's [49] classification theory of nonverbal behavior provides a cognitive psychological basis for the construction of the semantic system of gestures in this paper, while Gallagher's [50] theory of empathic direct perception lays a theoretical foundation for the elaboration of the emotion transfer mechanism of gestures.

Dexterous, gesture-capable robot hands are poised to transform a variety of human-centered domains by bringing natural, intuitive non-verbal interaction to everyday tasks. In education, they can reinforce teaching by synchronizing demonstrations with natural hand motions; in service (e.g. retail and

hospitality), subtle gestures enhance human–robot rapport and customer engagement; and in collaborative manufacturing, intuitive hand-off and alignment gestures streamline human–robot teamwork. Embedding gesture semantics into robotic control will thus bridge manipulation and non-verbal communication across these application scenarios.

### 3. LLMs allow dexterous hands to make the leap from mechanical mapping to semantic representation

The integration of large language models with humanoid dexterous hands is driving a shift from rigid command execution to dynamic semantic understanding, endowing robots with contextual reasoning, adaptive interaction, and social intelligence. Existing systems, however, remain confined to one-way instruction execution without the capacity for proactive adaptation or intent prediction in complex environments. Although Google's RT-X demonstrates task decomposition and GenEM generates socially compliant gestures, both fall short in nuanced context analysis and spontaneous behavioral feedback. The key challenge is to move beyond mere instruction translation toward autonomous context interpretation aligned with human cognition, enabling natural interaction and fulfilling deeper human expectations for collaborative partners. The following sections will detail the technical pathways and challenges for realizing this vision and explore how synergizing semantic intelligence with human-centered interaction paradigms can expand the intelligent reach of humanoid dexterous hands.

#### 3.1. Trends in combining robotics and LLMs

The convergence of robotics and large language models is a major trend. LLMs already exhibit exceptionally powerful conversational intelligence – OpenAI's ChatGPT being a prime example – so their integration into humanoid platforms naturally promises to elevate dialogue-based interactions. Yet verbal exchange represents only one facet of communication; non-verbal cues – gestures, postures, and expressive motions – account for a substantial portion of social presence. Accordingly, recent research has begun embedding LLMs directly into motion-generation pipelines, using semantic understanding to drive multimodal behavior and optimize the synergy between linguistic commands and physical actions. Google's RT-X platform combines LLMs with robotic manipulation, enabling a robotic arm to understand natural language commands (e.g., “Organize your room”) and autonomously break down the steps of a task (identifying clutter, sorting, and organizing). This technology has already demonstrated its potential in scenarios such as folding clothes and fine assembly. Large models allow dexterous gestures to be expressed more richly by processing complex scene information. Recently the rapid development of language models has revolutionized multimodal data processing [51–56], whose core architecture is based on Transformer [57], which achieves end-to-end generation through serialization of text tokens.

Large language models and vision language models (VLMs) can be used in almost every part of robot development [58,59]. VLM agents are involved at the perception-to-understanding level, where visual information is analyzed to make more contextual perceptual judgments and contextual information is fed back and interpreted Chain of Thought (CoT). LLM agents are primarily translated into executable actions at the decision-making level. In perception, visual-linguistic models and visual-linguistic-action models have been shown to significantly enhance the generalization capabilities of robots [60–64]. LLM production has been used to make planning and execution more flexible and context-sensitive [65–71]. In control, linguistic conditional strategies and

transformer-based robots have been extensively studied in recent research on control [72–78]. LLMs can significantly improve robot–environment and robot–human interaction [79,80]. Table 1 provides an overview of the diverse applications of LLMs in robotics, highlighting their contributions to improved semantic understanding and task execution.

The ability of the dexterous hand to perform complex and varied tasks is not achieved by a highly bionic human structure alone. Just as in the case of human beings, neither dexterous grasping nor dexterous gestures can be achieved without a “brain”. Currently, dexterous hands are making the leap from the perceptual to the cognitive level. Traditional approaches focus on grasping stability (e.g., trajectory planning and reinforcement learning) but neglect the importance of human behavioral patterns and semantic interactions [81,82]. In recent years, the fusion of LLMs and multimodal data has provided new ideas for natural interaction and task generalization for dexterous hands. RealDex improves the grasping success rate by filtering the grasping gestures that are most compatible with human preferences using multimodal LLMs and outperforms other grasping approaches as tested by both qualitative and quantitative methods [81]. Semantic understanding of linguistic commands significantly improves the task adaptability of grasping, and SayFuncGrasp proposes a layered framework [83]: Grasping function inference: a few-shot cue guides the LLMs to parse the commands and output the type of grasping (e.g., “grip” or “pinch”) and the functional part (e.g., “hold” or “pinch”). Li et al. proposed a cognitive grasping system that mimics the intuitive reasoning process of human beings by using LLM to infer missing features (e.g., fragility, texture) from limited information such as object name and material and by accurately defining feature categories (e.g., categorizing “shape” into “object”). By precisely defining the feature categories (e.g., categorizing “shape” into cube, cylinder, etc.), we can improve the complementation accuracy [82]. It can be seen that incorporating LLMs into dexterous hands can significantly improve their semantic reasoning and generalization ability, thus enhancing the grasping adaptability of dexterous hands.

#### 3.2. Trends in combining dexterous hands and LLMs

In terms of semantically driven paradigms for dexterous hand gesture generation, Brown et al.'s [84] proposal of a few-shot learning capability provides a theoretical basis for semantic generalization based on large models, while Mahadevan et al.'s [85] framework for generating robotic control code via LLMs further validates the potential of language models for the generation of complex behavioral sequences. Wang et al.'s [86] illustrated that LLMs can generate low-level control commands to enable quadrupedal robots to walk.

Approaches based on large language models (LLMs) [87–89] are able to transform natural language commands into parameterized control code through chain-of-thought, supporting dynamic generation of multimodal behaviors. For example, the GenEM framework [90] can decompose the “nod to greet” command into combinations of head panning, light display, and other actions, and incorporate social norms (e.g., maintaining eye contact) [91], so that gestures are both functional and socially adaptive.

Palo et al. [92] used Vision Transformers to extract 3D key-point tokens from images and utilized expert demonstrations as contexts for few-shot imitation learning with LLMs. (Utilize expert demonstrations as contexts for few-shot imitation learning with LLMs.) Wang et al. [93] used LLMs directly as the underlying feedback controller by initializing cues with a small amount of physical environment data and combining descriptive text with historical observation. The cue design of action sequences controlled a quadrupedal robot to perform a simple walking task in

**Table 1**  
Application areas and functions of LLMs in the field of robotics.

Application areas	Specific functions	Technical realization
Language instruction parsing and reasoning	Reasoning about task requirements and processing unseen task instructions through in-context learning (ICL) and chain of thought (CoT)	LLMs parse natural language instructions to extract implicit requirements → CoT reasoning decomposes into atomic action sequences
multimodal task planning	Combining verbal instructions with visual scenes to generate task-oriented grasping strategies	Align point-cloud features with language features → LLMs generate grasp-pose encoding → Reinforcement learning optimizes grasp success rate
Collaborative control optimization	Reduced complexity of high-freedom hand control	LLMs infer the appropriate grasp type → generate the corresponding motion in a low-dimensional synergy space → dynamically adjust joint trajectories (e.g., pinching or gripping)

simulation. The EMOTION framework, on the other hand, combines VLMs and LLMs to realize a robot perception-to-decision path [94]. The study couples the contextual understanding of VLMs with the motion generation of LLMs to construct a closed-loop system of “perception-reasoning-expression”. As a result, the large language model can generate natural gesture interaction sequences through contextual analysis. This method provides a new paradigm for multimodal human–robot interaction and aids the robot’s expressive ability.

Currently, LLMs are used in four major areas of robot dexterous manipulation: task planning and decomposition, multimodal perception and action generation, cross-dexterous hand generalization control, and data-driven grasping strategy optimization (see Table 2). In terms of task planning, LLMs decompose high-level commands (e.g., ‘pour a glass of water’) into executable atomic action sequences [95,96] (e.g., ‘grasp the cup’ and ‘tilt the arm’) via semantic parsing, and resolve the dual-arm manipulation challenge through spatiotemporal coordination strategies [97]. For multi-modal fusion, research combines VLMs and diffusion model (e.g., DexHandDiff) to fuse verbal commands with perceptual information, such as point clouds and images, to generate physically feasible grasping gestures [98,99]. For generalization capability, Multi-GraspLLM achieves unified control across manipulators through semantic parameter mapping to avoid modeling each type of dexterous hands separately [100]. In addition, the LLM-driven data annotation system [101] with simulation dataset [81] significantly improves the diversity and adaptability of grasping strategies.

The value of combining humanoid dexterous hands and LLMs lies in the ability to enhance semantic interpretation. LLMs provide ideas for shaping social robots with common sense [105] and allow robots to interpret human social cues and respond with natural and engaging gestures [94]. LLMs semantically understand and describe complex scenarios and can help dexterous hands better understand their operating environment [100]. Zhan et al. (2024) used LLMs to decompose a complex task into a series of basic operations [96]. Generating linguistic guidance through LLMs can also help the model to understand the task intent and generate corresponding grasping actions [83,101].

### 3.3. Current applications of LLMs in human–robot interaction

Current applications of LLMs in human–robot interaction (HRI) have significantly expanded robots’ multimodal generation, social interaction, and task execution capabilities. In the field of multimodal generation, LLMs enable mapping from user queries to diverse responses by integrating textual, visual, and auditory information. For example, VideoCLIP supports zero-shot transfer learning by associating video with text [106], while MotionGPT generates motion sequences of avatars based on textual cues [107], which further enhances the intuitiveness of interaction. In the field of social robotics, LLMs empower robots with

commonsense reasoning capabilities, enabling them to serve educational, medical, and other scenarios. For example, chatbots can generate personalized stories for children [108], while chat systems combining Bi-LSTM and attentional mechanisms can assist autistic children in interaction [109]. In addition, GPT-3-powered philosophical dialog robots demonstrate the potential of LLMs in complex semantic understanding [110]. In terms of command execution and task completion, LLMs enhance robot autonomy by transforming fuzzy commands into concrete action planning. For example, ProgPrompt directly generates robot control strategy code [111]. However, the wide application of LLMs in HRI still faces challenges such as data ethics and security [112], contextual understanding [113], and generalization capability [114], which need to be continuously optimized through interdisciplinary collaboration.

Despite the fact that numerous studies have been conducted to improve the grasping performance and task decomposition of dexterous hands using LLMs, there are limitations in the application of LLMs, which are more suitable for executing abstract, broad, and generalized tasks rather than tasks that require high precision. Applying LLMs to task execution in complex scenarios often fails to achieve the desired results [114,115]. However, LLMs show their significant advantages in dealing with linguistic and comprehension-based information processing tasks, such as contextual understanding and judgment, and linguistic feedback [57].

### 3.4. LLMs enable novel approaches to gesture expression

The semantic expression of sign language is consistent with basic human linguistic cognition and with the linguistic laws of human sign language. The lexical indexing model suggests that gestures are associated with the integration of spoken surface forms that enhance lexical activation and facilitate semantic access [116]. The Broca’s area of the brain (the main brain area responsible for language processing) is also activated when people gesture, suggesting that gesture and language may have a common neural basis in the brain [117]. Evidence from evolutionary neuroscience suggests that there is an evolutionary continuity between gesture and language and that Broca’s area, as the main brain area responsible for language processing in humans, may control both gesture and word pronunciation simultaneously [118–122]. Moreover, it has also been shown that LLMs can effectively handle new languages by leveraging shared commonalities with previously learned languages [123]. Therefore, the generalization of sign language and gestures through LLMs is a feasible way to express robotic gestures.

At present, the key to realizing the intelligence of dexterous hands is to solve the problem of “hand-eye-brain synergy”. Through LLMs, the dexterous hands can be trained to understand sign language and then search for the correlation between language and sign language. The “sign language lexicon” (the internal semantic mapping between gestures and language within

**Table 2**  
Technical routes for the control of dexterous hands in large language models.

Author (year)	Application scenario	Main contributions	Technology paths for combining large language models
Li et al. (2023) [82]	Grabbing Planning with Incomplete Awareness	Completing object features and generating grasping strategies using LLMs	Completion of missing attributes by LLMs after object feature extraction
Gbagbe et al. (2024) [102]	Dexterous operation of dual-armed robots	Integration of vision, verbal, and movement for bimanual synergy	A VLM and motion module parse instructions into action sequences, then convert them into predefined API calls.
Feng et al. (2024) [98]	Task-oriented dexterous grasping	Combining MLLM to generate task-oriented grasping actions	Combining MLLM and VLMs for task-oriented grasping
Chu et al. (2024) [97]	Coordinated control of dual-arm robots	Solving the two-handed robot spatio-temporal coordination problem using LLMs	LLMs decompose tasks into uncoordinated and coordinated phases, with step-by-step control instructions
Liu et al. (2024) [81]	Dexterous grasping by humanoid robots	Enhancing crawl detail understanding with LLMs	Rendering of dexterous hand-object interaction images, input to Gemini model combined with prompt understanding of grasping details
Zhan et al. (2024) [96]	Two-handed object handling	Propose a functional layering framework to support complex task decomposition and action synthesis	Development of LLM-based task decomposer and diffusion model-driven motion generator to verify the feasibility of task and motion planning
Wu et al. (2024) [103]	dexterous hand control	Using LLMs to Generate Gesture Descriptions from Languages	Combine a well-trained hand model with a large language model to generate gestures
Wei et al. (2024) [101]	Dexterous grasping of linguistic guidance	For dexterous crawling generation based on natural language instructions	Construction of datasets by hand-object interaction retargeting strategy and LLM-assisted language-guided annotation system
Li et al. (2025) [100]	Control generalization across dexterous hands	Semantic Gripping Parameter Generation for Different Manipulators via LLMs	Generating semantically compliant gripping poses for different types of manipulators by combining verbal commands and 3D point cloud data
Li et al. (2025) [83]	Language-guided dexterity feature capture	Grab function generated using LLMs	Utilizing LLMs to parse language commands and reason about grip function parameters (position and type)
Liang et al. (2025) [104]	interaction-aware diffusion planning	Modeling dynamic interactions through two-stage diffusion	Translation of high-level task descriptions into executable robot control programs, guided by LLMs
Zhong et al. (2025) [99]	Robot Dexterity Grab	Fusing point cloud features with semantic descriptions to improve complex object understanding	Enhancing the model's semantic understanding of complex objects through LLMs

the robot's reasoning system) will be explored and expanded by reasoning about the possible "language model" of human gestures through LLMs and the existing sign language dictionary. Moreover, this lexicon can provide more rational and intuitive gesture movement patterns for gesture output and enhance the naturalness of humanoid robot interaction. At present, the multimodal sign language model SignLLM has been able to realize the dynamic generation of sign language to gestures [124]. And it can realize the generalization of gesture semantics, which extends the possibility of gesture interaction.

Most of the previous gesture generation studies rely on predefined action libraries, mainly through the joint angle control to realize the basic grasping action, and lack of dynamic contextual response ability, while this paper, based on the in-context learning mechanism of the EMOTION framework [94], proposes a contextual gesture generation scheme based on a large language model.

#### 4. A novel framework for humanoid dexterous hand interaction based on gesture-semantic mapping

Although LLMs facilitate semantic generalization of gestures, existing systems lack a unified closed-loop framework for evaluation and iterative, user-driven refinement. To enable semantic interaction and cognitive empathy in humanoid dexterous

hands (Fig. 5), we propose a closed-loop framework that integrates multimodal perception, LLM-driven cognition, and human-centered assessment for context-adaptive gesture generation and refinement.

**Perceptual Layer:** multimodal input processing, contextual semantics are extracted and understood through visual language models (VLMs), including but not limited to images/videos (visual), textual descriptions (linguistic), and physical motion data in the environment (action). This information is preprocessed and fed into the appropriate encoder. The visual encoder will use Transformers techniques to extract image features;

**Cognitive Layer:** a language model based on the Transformer architecture, responsible for understanding and parsing natural language commands into a machine-understandable form. The mapping rules from social contexts to gesture symbols are parsed through large model instructions;

**Generation layer:** using the sequence generation capability of LLMs to transform semantic symbols into joint space trajectories or accessing a large model of sign language gestures with strong generalization capability for gesture action extraction and at the same time, outputting the action coordinate information of such gestures to the dexterous hands' driver module in order to perform the corresponding operations;

**Evaluation Layer:** Based on the user's physiological signals (e.g., eye movement, EEG, etc.) and subjective feedback, the effectiveness and naturalness of semantic communication of gestures

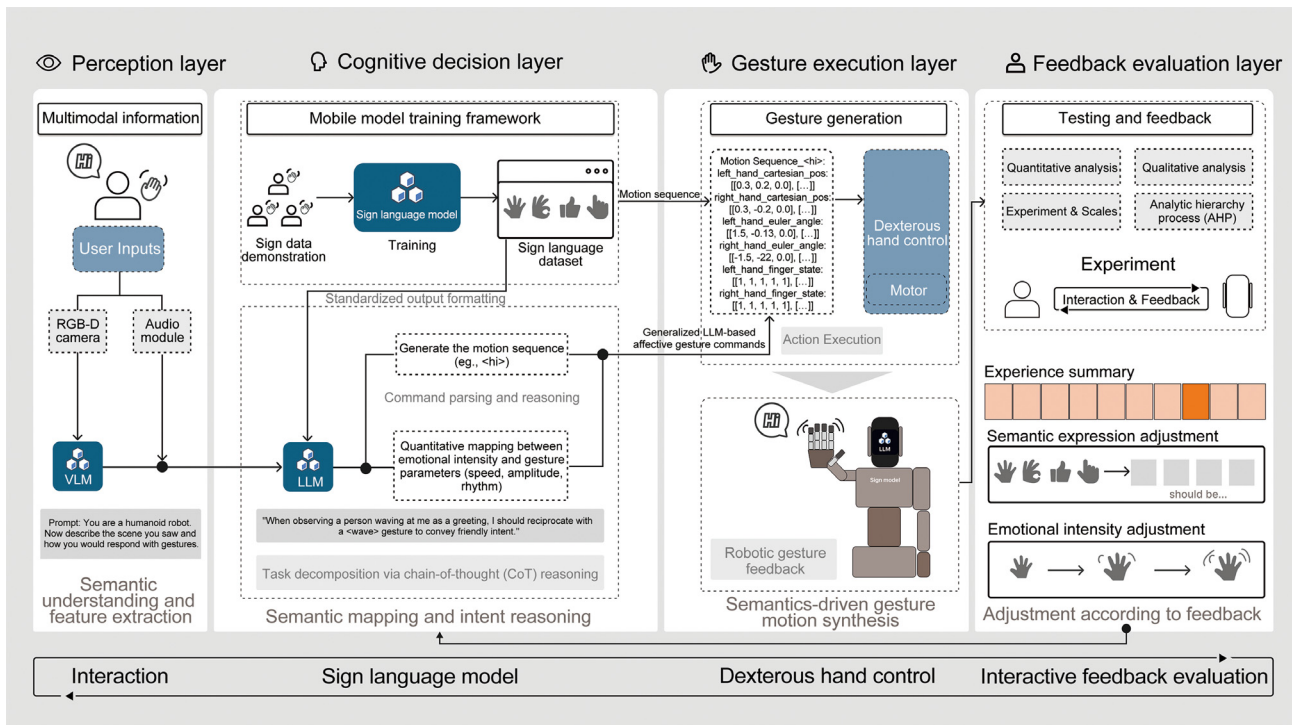


Fig. 5. Framework of the closed-loop system "Perception-Cognition-Generation-Evaluation".

are quantified to form a closed loop of optimization. At the quantitative level, cognitive load is analyzed through eye tracking (gaze point distribution, dwell time) and EEG, combined with behavioral indicators such as task completion time, misuse rate, etc. At the qualitative level, gesture naturalness and semantic clarity are evaluated using Likert scale, supplemented with semi-structured interviews to dig out the feedback on emotional and cultural appropriateness.

## 5. Challenges and future directions

### 5.1. Key challenges

Although recent methods have partially addressed issues concerning dexterous hands in human-robot interaction, achieving human-level dexterity and expressiveness remains a long-term endeavor. Mechanical complexity in humanoid dexterous hands, coupled with the instability of large language models, has hindered progress in using five-fingered dexterous hands for emotional expression in human-robot interaction. In general, the principal challenges in this field are as follows:

- **Hardware limitations:** the lack of fine-motion capabilities of the robot's fingers limits the accurate reproduction of certain human gestures. This affects the expressive accuracy of some of the gestures.
- **Latency Challenges of Large Language Models (LLMs) in Real-Time Control:** Generating the initial action sequence via LLM APIs and performing single-round feedback optimization introduces significant delays, making it difficult to achieve the low-latency control required for real-time interaction.
- **Depth of comprehension:** While LLMs excel at parsing natural language, they still struggle to infer implicit relationships, logical hierarchies, and context nuances in task instructions—and to convert those insights into concrete, multi-step gesture plans.

- **Trajectory coordination:** Integrating high-level intent from LLM outputs with low-level joint-trajectory planners requires seamless harmonization of abstract semantic goals and precise motor commands.
- **Robot Morphology Variability:** Adapting the same framework to robots with different shapes, degrees of freedom, and kinematic constraints demands robust generalization strategies.

### 5.2. Future directions

To propel humanoid dexterity toward cognitive intelligence and richer expressiveness, we identify the following synergistic research avenues:

- **Combining multimodal large language models (e.g., GPT-4o) to realize gesture-language-emotion dynamic mapping, multimodal interaction with humans through gestures, and broadening the robot's non-verbal emotional expressions**
- **Achieving a closed loop from perception to cognition is a vital future direction.** For example, once the robot perceives the user's semantic intent, the large language model processes and interprets this information, enabling the dexterous hands to produce corresponding gestural feedback. In this way, the system can autonomously perceive, understand, and generate a gestural language repertoire that aligns with human cognitive conventions.
- **Establishing an emotion mapping model, mapping basic emotion through gesture speed, amplitude, and rhythm, developing an "emotion-gesture-speech" synergy system, realizing multimodal emotion consistency, and improving emotion recognition accuracy and empathy score.**
- **Expanded Behavioral Repertoire:** Incorporating a wider action space enables generation of complex, sequential, and context-rich behaviors beyond static poses.
- **Multi-Modal Expressiveness:** Fusing gestures with audio cues promises richer communication channels and heightened user engagement.

- Autonomous State Selection: Future integration of LLMs for context-aware state input aims to minimize human intervention, enhancing the framework's scalability for novel tasks.
- Establishment of user experience metrics: building user trust metrics to assess robot emotionality.
- Cross-cultural gesture semantic base construction: building a semantic base of gesture expressions covering cultural differences in different countries.

Moreover, the majority of dexterous hand systems remain predominantly passive and command-driven, relying on explicit semantic instructions for tasks such as object grasping. This paradigm limits their capacity for genuine human-robot engagement, as these systems do not autonomously interpret environmental context or social norms to generate appropriate gestural behaviors. Future research should prioritize the development of proactive interaction capabilities, enabling dexterous hands to dynamically assess situational cues and execute socially contextualized gestures without explicit user commands, thereby advancing toward more natural and intuitive human-robot collaboration.

By systematically tackling these challenges – through tighter hardware-software co-design, deeper semantic reasoning, and enriched multimodal frameworks – we can move closer to robotic hands that not only mimic human dexterity but also participate meaningfully in social and emotional exchange.

## 6. Conclusion

In summary, dexterous robotic hands are evolving beyond mechanical replication toward cognitive empathy, where gesture naturalness directly influences user trust and satisfaction. Large language models play a pivotal role in this transition, moving gesture control from rigid, rule-based routines to flexible, context-sensitive semantic generation. By systematically reviewing gesture generation technologies and grounding them in theories of human cognition and language, we have articulated a novel “perception-cognition-generation-assessment” loop. This closed-loop approach integrates multimodal sensing, transformer-based semantic understanding, dynamic trajectory synthesis, and quantitative and qualitative user feedback. Adopting this framework promises to advance humanoid robots from mere manipulators to socially intelligent partners capable of authentic, empathetic non-verbal interaction.

## CRedit authorship contribution statement

**Xin Li:** Writing – original draft, Visualization, Conceptualization. **Wenfu Xu:** Writing – review & editing, Methodology. **Zaiqiao Ye:** Writing – review & editing. **Han Yuan:** Writing – review & editing, Supervision, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (62173114), Guangdong Basic and Applied Basic Research Foundation (2024A1515011228), Guangdong Provincial Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics (2023B1212010005), the Shenzhen Science and Technology Program (KJZD20240903100501002 and GXWD20231129174132001), and the Program of Shenzhen Peacock Innovation Team (KQTD20210811090146075).

## References

- [1] Z. Xia, Z. Deng, B. Fang, Y. Yang, F. Sun, A review on sensory perception for dexterous robotic manipulation, *Int. J. Adv. Robot. Syst.* 19 (2) (2022) <http://dx.doi.org/10.1177/17298806221095974>.
- [2] Kelsey Tonner, What to do with your hands while speaking? Hand gestures!, 2017.
- [3] Cohen Doron, Geoffrey Beattie, Heather Shovelton, Tracking the distribution of individual semantic features in gesture across spoken discourse: new perspectives in multi-modal interaction, *Semiotica* 2011 (185) (2011).
- [4] J. Li, M. Chignell, Sachi Mizobuchi, Michiaki Yasumura, Emotions and messages in simple robot gestures, in: *Lecture notes in computer science*, 2009, pp. 331–340.
- [5] H. Duan, P. Wang, Y. Li, D. Li, W. Wei, Learning human-to-robot dexterous handovers for anthropomorphic hand, *IEEE Trans. Cogn. Dev. Syst.* 15 (3) (2023) 1224–1238.
- [6] Xiang Pan, Malcolm Doering, 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, HRI'24*, Association for Computing Machinery, New York, NY, USA, 2024, pp. 552–560.
- [7] Y. Huang, D. Fan, H. Duan, et al., Human-like dexterous manipulation for anthropomorphic five-fingered hands: A review, *Biomim. Intell. Robot.* (2025) 100212.
- [8] Shadow dexterous hand e1 series, 2013, <http://www.shadowrobot.com/wpcontent/uploads/shadowdexteroushandtechnicalspecificationE120130101.pdf>.
- [9] H. Liu, K. Wu, P. Meusel, et al., Multisensory five-finger dexterous hand: The DLR/HIT hand II, in: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2008, pp. 3692–3697.
- [10] Park, integrated linkage-driven dexterous anthropomorphic robotic hand, *Nat. Commun.* 12 (1) (2021) 1–13; *Gen Robots and Systems, IEEE*, 2008, pp. 3692–3697.
- [11] Y. Liu, Z. Li, H. Liu, Z. Kan, B. Xu, Bioinspired embodiment for intelligent sensing and dexterity in fine manipulation: a survey, *IEEE Trans. Ind. Inf.* 16 (7) (2020) 4308–4321.
- [12] A.R. Sobinov, S.J. Bensmaia, The neural mechanisms of manual dexterity, *Nature Rev. Neurosci.* 22 (12) (2021) 741–757.
- [13] L. Biagiotti, F. Lotti, C. Melchiorri, G. Vassura, How Far Is the Human Hand? A Review on Anthropomorphic Robotic End-Effectors, *Tech. Rep.*, University of Bologna, 2004.
- [14] A.M. Dollar, R.D. Howe, The highly adaptive SDM hand: design and performance evaluation, *Int. J. Robot. Res.* 29 (5) (2010) 585–597.
- [15] F. Ficuciello, D. Zaccara, B. Siciliano, Learning grasps in a synergy-based framework, in: *International Symposium on Experimental Robotics*, Springer, 2016, pp. 125–135.
- [16] H. Yousef, M. Boukallel, K. Althoefer, Tactile sensing for dexterous in-hand manipulation in robotics: a review, *Sensors Actuators A - Phys.* 167 (2) (2011) 171–187.
- [17] Junchang Yang, J. Mun, S. Kwon, et al., Electronic skin: Recent progress and future prospects for skin-attachable devices for health monitoring, robotics, and prosthetics, *Adv. Mater.* 31 (48) (2019) 1904765.
- [18] Li Shuo, Bai Hedan, R.F. Shepherd, et al., Bio-inspired design and additive manufacturing of soft materials, machines, robots, and haptic interfaces, *Angew. Chem. - Int. Ed.* 58 (33) (2019) 11182–11204.
- [19] T. Yamaguchi, T. Kashiwagi, T. Aire, et al., Human-like electronic skin-integrated soft robotic hand, *Adv. Intell. Syst.* 1 (2) (2019) 1900018.
- [20] Y.W. Chao, W. Yang, Y. Xiang, et al., Dexycb: A benchmark for capturing hand grasping of objects, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9044–9053.
- [21] S. Hampali, M. Rad, M. Oberweger, V. Lepetit, Honnotate: a method for 3d annotation of hand and object poses, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3196–3206.
- [22] S. Brahmabhatt, C. Tang, C.D. Twigg, C.C. Kemp, J. Hays, ContactPose: A dataset of grasps with object contact and hand pose, in: *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August (2020) 23–28*, *Proceedings, Part XIII 16*, Springer, 2020, pp. 361–378.
- [23] C. Yu, P. Wang, Dexterous manipulation for multi-fingered robotic hands with reinforcement learning: A review, *Front. Neurobotics* 16 (2022) 861825.
- [24] M. Vecerik, T. Hester, J. Scholz, et al., Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards, 2017, arXiv preprint [arXiv:1707.08817](https://arxiv.org/abs/1707.08817).
- [25] P. Sharma, D. Pathak, A. Gupta, Third-person visual imitation learning via decoupled hierarchical controller, in: *Advances in Neural Information Processing Systems*, 2019, p. 32.
- [26] F. Liu, Z. Ling, T. Mu, H. Su, State alignment-based imitation learning, in: *ICLR*, 2020.

- [27] Y.W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y.S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, et al., Dexycb: A benchmark for capturing hand grasping of objects, in: *CVPR*, 2021.
- [28] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.W. Chao, Q. Wan, S. Birchfield, N. Ratliff, D. Fox, Dexipilot: vision-based teleoperation of dexterous robotic hand-arm system, in: *ICRA*, 2020.
- [29] D. Antotsiou, G. Garcia-Hernando, T.K. Kim, Task-oriented hand motion retargeting for dexterous manipulation imitation, in: *ECCV Workshops*, 2018.
- [30] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, S. Levine, Learning complex dexterous manipulation with deep reinforcement learning and demonstrations, 2018.
- [31] I. Radosavovic, X. Wang, L. Pinto, J. Malik, State-only imitation learning for dexterous manipulation, in: *IROS*, 2021.
- [32] J. Ho, S. Ermon, Generative adversarial imitation learning, in: *NeurIPS*, 2016.
- [33] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, S. Levine, Learning complex dexterous manipulation with deep reinforcement learning and demonstrations, 2017, arXiv.
- [34] I. Radosavovic, X. Wang, L. Pinto, J. Malik, State-only imitation learning for dexterous manipulation, in: *IROS*, 2021.
- [35] H. Knight, R. Simmons, Laban head-motions convey robot state: A call for robot body language, in: *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 2881–2888.
- [36] M. Dubois, J.-A. Claret, L. Basañez, G. Venture, Influence of emotional motions in human–robot interactions, in: *Proc. 29th IEEE Int. Conf. Robot Hum. Interactive Commun.*, 2020, pp. 1243–1250.
- [37] A.D. Dragan, S. Bauman, J. Forlizzi, S.S. Srinivasa, Effects of robot motion on human–robot collaboration, in: *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.–Robot Interact.*, 2015, pp. 51–58.
- [38] A. Zhou, D. Hadfield-Menell, A. Nagabandi, A.D. Dragan, Expressive robot motion timing, in: *Proc. ACM/IEEE Int. Conf. Hum.–Robot Interact.*, 2018, pp. 22–31.
- [39] A. Zinina, L. Zaidelman, N. Arinkin, A. Kotov, Non-verbal behavior of the robot companion: A contribution to the likeability, *Procedia Comput. Sci.* 169 (2020) 800–806.
- [40] M. Salem, K. Rohlfing, S. Kopp, F. Joubin, A friendly gesture: investigating the effect of multimodal robot behavior in human–robot interaction, in: 2011 RO-MAN, Atlanta, GA, USA, 2011, pp. 247–252, <http://dx.doi.org/10.1109/ROMAN.2011.6005285>.
- [41] J. Xu, J. Broekens, K. Hindriks, M.A. Neerinx, Effects of a robotic storyteller's moody gestures on storytelling perception, in: 2015 International Conference on Affective Computing and Intelligent Interaction, (ACII), Xi'an, China, 2015, pp. 449–455, <http://dx.doi.org/10.1109/ACII.2015.7344609>.
- [42] A. Esposito, J. Vassallo, A.M. Esposito, N. Bourbakis, On the amount of semantic information conveyed by gestures, in: 2015 IEEE 27th International Conference on Tools with Artificial Intelligence, ICTAI, Vietri Sul Mare, Italy, 2015, pp. 660–667, <http://dx.doi.org/10.1109/ICTAI.2015.100>.
- [43] S. Saunderson, G. Nejat, How robots influence humans: a survey of nonverbal communication in social human–robot interaction, *Int. J. Soc. Robot.* 11 (4) (2019) 575–608.
- [44] U. Zabala, I. Rodriguez, J.M. Martínez-Otzeta, E. Lazkano, Expressing robot personality through talking body language, *Appl. Sci. (Basel)* 11 (10) (2021) 4639.
- [45] I. Rodriguez, J.M. Martínez-Otzeta, I. Irigoien, E. Lazkano, Spontaneous talking gestures using generative adversarial networks, *Robot. Auton. Syst.* 114 (C) (2019) 57–65, <http://dx.doi.org/10.1016/j.robot.2018.11.024>, [Online]. Available.
- [46] S. Mirchandani, F. Xia, P. Florence, Brian. Ichter, D. Driess, M.G. Arenas, K. Rao, D. Sadigh, A. Zeng, Large language models as general pattern machines, in: 7th Annual Conference on Robot Learning, 2023.
- [47] L. Roy, E.A. Croft, A. Ramirez, D. Kulić, GPT-driven gestures: Leveraging large language models to generate expressive robot motion for enhanced human–robot interaction, *IEEE Robot. Autom. Lett.* 10 (5) (2025) 4172–4179, <http://dx.doi.org/10.1109/LRA.2025.3547631>.
- [48] M. Marmpena, F. Garcia, A. Lim, et al., Data-driven emotional body language generation for social robotics, 2022, arXiv preprint arXiv:2205.00763.
- [49] P. Ekman, W.V. Friesen, The repertoire of nonverbal behavior: categories, origins, usage, and coding, *Semiotica* 1 (1) (1969) 49–98.
- [50] S. Gallagher, Empathy and theories of direct perception, in: *The Routledge Handbook of Philosophy of Empathy*, Routledge, 2017, pp. 158–168.
- [51] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.* 33 (2020) 1877–1901.
- [52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [53] R. Po, W. Yifan, V. Golyanik, et al., State of the art on diffusion models for visual computing, in: *Computer Graphics Forum*, vol. 43, (2) 2024, e15063.
- [54] C. Raffel, N. Shazeer, A. Roberts, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (140) (2020) 1–67.
- [55] Y. Wu, Z. Zhang, J. Chen, et al., Vila-u: a unified foundation model integrating visual understanding and generation, 2024, arXiv preprint arXiv:2409.04429.
- [56] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen, A survey of large language models, 2024.
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, 2017.
- [58] F. Zeng, W. Gan, Y. Wang, N. Liu, P.S. Yu, Large language models for robotics: a survey, 2023, arXiv preprint arXiv:2311.07226.
- [59] C. Zhang, J. Chen, J. Li, Y. Peng, Z. Mao, Large language models for human–robot interaction: a review, *Biomim. Intell. Robot.* (2023) 100131.
- [60] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Chormanski, T. Ding, D. Driess, A. Dubey, C. Finn, et al., Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023, arXiv preprint arXiv:2307.15818.
- [61] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International Conference on Machine Learning*, pmlr, 2021, pp. 8748–8763.
- [62] D. Shah, B. Osinski, S. Levine, et al., Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action, in: *Conference on Robot Learning*, pmlr, 2023, pp. 492–504.
- [63] D. Driess, F. Xia, M.S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al., Palm-e: An embodied multimodal language model, 2023, arXiv preprint arXiv:2303.03378.
- [64] Y. Zhu, R. Mottaghi, E. Kolve, J.J. Lim, A. Gupta, L. Fei-Fei, A. Farhadi, Target-driven visual navigation in indoor scenes using deep reinforcement learning, in: 2017 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2017, pp. 3357–3364.
- [65] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al., Do as i can, not as i say: Grounding language in robotic affordances, 2022, arXiv preprint arXiv:2204.01691.
- [66] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, *Adv. Neural Inf. Process. Syst.* 35 (2022) 27730–27744.
- [67] M. Crosby, M. Rovatsos, R. Petrick, Automated agent decomposition for classical planning, in: *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 23, 2013, pp. 46–54.
- [68] B. Xu, Z. Peng, B. Lei, S. Mukherjee, Y. Liu, D. Xu, Rewoo: decoupling reasoning from observations for efficient augmented language models, 2023, arXiv preprint arXiv:2305.18323.
- [69] T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, *Adv. Neural Inf. Process. Syst.* 35 (2022) 22199–22213.
- [70] S.S. Raman, V. Cohen, D. Paulius, I. Idrees, E. Rosen, R. Mooney, S. Tellex, Cape: Corrective actions from precondition errors using large language models, 2022, arXiv preprint arXiv:2211.09935.
- [71] Z. Liu, A. Bahety, S. Song, Reflect: summarizing robot experiences for failure explanation and correction, 2023, arXiv preprint arXiv:2306.15724.
- [72] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, et al., Open x-embodiment: robotic learning datasets and rt-x models, 2023, arXiv preprint arXiv:2310.08864.
- [73] S. Reed, K. Zolna, E. Parisotto, S.G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J.T. Springenberg, et al., A generalist agent, 2022, arXiv preprint arXiv:2205.06175.
- [74] N.M. Shafiqullah, Z. Cui, A.A. Altanzaya, L. Pinto, Behavior transformers: cloning k modes with one stone, *Adv. Neural Inf. Process. Syst.* 35 (2022) 22955–22968.
- [75] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu, et al., Ai2-thor: an interactive 3d environment for visual ai, 2017, arXiv preprint arXiv:1712.05474.
- [76] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al., Rt-1: Robotics transformer for real-world control at scale, 2022, arXiv preprint arXiv:2212.06817.
- [77] C. Matuszek, E. Herbst, L. Zettlemoyer, D. Fox, Learning to parse natural language commands to a robot control system, in: *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, Springer, 2013, pp. 403–415.
- [78] D. Chen, R. Mooney, Learning to interpret natural language navigation instructions from observations, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, 2011, pp. 859–865.

- [79] J. Arkin, D. Park, S. Roy, M.R. Walter, N. Roy, T.M. Howard, R. Paul, Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions, *Int. J. Robot. Res.* 39 (10–11) (2020) 1279–1304.
- [80] A. Buckler, L. Figueredo, S. Haddadin, A. Kapoor, S. Ma, S. Vemprala, R. Bonatti, Latte: language trajectory transformer, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 7287–7294.
- [81] RealDex: Towards Human-like Grasping for Robotic Dexterous Hand.
- [82] Knowledge Augmentation and Task Planning in Large Language Models for Dexterous Grasping.
- [83] Language-Guided Dexterous Functional Grasping by LLM Generated Grasp Functionality and Synergy for Humanoid Manipulation.
- [84] T.B. Brown, Language models are few-shot learners, 2020, arXiv preprint arXiv:2005.14165.
- [85] K. Mahadevan, J. Chien, N. Brown, Z. Xu, C. Parada, F. Xia, A. Zeng, L. Takayama, D. Sadigh, Generative expressive robot behaviors using large language models, in: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, 2024, pp. 482–491.
- [86] Y.-J. Wang, B. Zhang, J. Chen, K. Sreenath, Prompt a robot to walk with large language models, in: Conference on Decision and Control, CDC, 2024.
- [87] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, Andy Zeng, Code as policies: language model programs for embodied control, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 9493–9500.
- [88] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, Animesh Garg, Prog-prompt: generating situated robot task plans using large language models, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 11523–11530.
- [89] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, Anima Anandkumar, Voyager: an open-ended embodied agent with large language models, 2023, arXiv preprint arXiv:2305.16291.
- [90] Jason Wei, Quezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, in: Large Language Models, in: Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824–24837.
- [91] Minae Kwon, Hengyuan Hu, Vivek Myers, Siddharth Karamcheti, Anca Dragan, Dorsa Sadigh, Towards grounded social reasoning, 2023, arXiv preprint arXiv:2306.08651.
- [92] N. Di Palo, E. Johns, Keypoint action tokens enable in-context imitation learning in robotics, *Robot.: Sci. Syst. (RSS)* 2024 (2024).
- [93] Y.-J. Wang, B. Zhang, J. Chen, K. Sreenath, Prompt a robot to walk with large language models, in: Conference on Decision and Control, CDC, 2024.
- [94] P. Huang, Y. Hu, N. Nechyporenko, et al., EMOTION: Expressive motion sequence generation for humanoid robots with in-context learning, 2024, arXiv preprint arXiv:2410.23234.
- [95] C. Wang, S. Hasler, D. Tanneberg, et al., Lami: Large language models for multi-modal human-robot interaction, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–10.
- [96] X. Zhan, L. Yang, Y. Zhao, et al., Oakink2: A dataset of bimanual hands-object manipulation in complex task completion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 445–456.
- [97] K. Chu, X. Zhao, C. Weber, et al., Large language models for orchestrating bimanual robots, in: 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids), IEEE, 2024, pp. 328–334.
- [98] Q. Feng, D.S.M. Lema, M. Malmir, et al., Dexgrasp: Dexterous generative adversarial grasping synthesis for task-oriented manipulation, in: 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids), IEEE, 2024, pp. 918–925.
- [99] Y. Zhong, Q. Jiang, J. Yu, et al., DexGrasp anything: Towards universal robotic dexterous grasping with physics awareness, 2025, arXiv preprint arXiv:2503.08257.
- [100] H. Li, W. Mao, W. Deng, et al., Multi-graspllm: A multimodal LLM for multi-hand semantic guided grasp generation, 2024, arXiv preprint arXiv:2412.08468.
- [101] Y.L. Wei, J.J. Jiang, C. Xing, et al., Grasp as you say: Language-guided dexterous grasp generation, 2024, arXiv preprint arXiv:2405.19291.
- [102] K.F. Gbagbe, M.A. Cabrera, A. Alabbas, et al., Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations, in: 2024 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, 2024, pp. 2864–2869.
- [103] T. Wu, S. Li, C. Lyu, et al., MoDex: Planning high-dimensional dexterous control via learning neural hand models, 2024, arXiv preprint arXiv:2409.10983.
- [104] Z. Liang, Y. Mu, Y. Wang, et al., DexHandDiff: Interaction-aware diffusion planning for adaptive dexterous manipulation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 1745–1755.
- [105] C. Zhang, J. Chen, J. Li, et al., Large language models for human-robot interaction: A review, *Biomim. Intell. Robot.* 3 (4) (2023) 100131.
- [106] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, C. Feichtenhofer, Videoclip: Contrastive pre-training for zero-shot video-text understanding, 2021, arXiv preprint arXiv:2109.14084.
- [107] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, T. Chen, MotionGPT: Human motion as a foreign language, 2023, arXiv preprint arXiv:2306.14795.
- [108] E. Nichols, L. Gao, R. Gomez, Collaborative storytelling with large-scale neural language models, in: Proceedings of the 13th ACM SIGGRAPH Conference on Motion, Interaction and Games, 2020, pp. 1–10.
- [109] X. Li, H. Zhong, B. Zhang, J. Zhang, A general Chinese chatbot based on deep learning and its-application for children with ASD, *Int. J. Mach. Learn. Comput.* 10 (4) (2020) 519–526.
- [110] E. Schwitzgebel, D. Schwitzgebel, A. Strasser, Creating a large language model of a philosopher, 2023, arXiv preprint arXiv:2302.01339.
- [111] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, A. Garg, Progprompt: generating situated robot task plans using large language models, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 11523–11530.
- [112] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models, 2021, 2021, arXiv preprint arXiv:2112.04359.
- [113] P. Tsarouchi, S. Makris, G. Chrysolouris, Human-robot interaction review and challenges on task planning and programming, *Int. J. Comput. Integ. Manuf.* 29 (8) (2016) 916–931.
- [114] S. Shahriar, K. Hayawi, Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations, 2023, arXiv preprint arXiv:2302.13817.
- [115] C.H. Song, J. Wu, C. Washington, B.M. Sadler, W.-L. Chao, Y. Su, Llm-planner: fresh-shot grounded planning for embodied agents with large language models, 2022, arXiv. arXiv preprint arXiv:2212.04088.
- [116] R.M. Krauss, Y. Chen, R.F. Gottesman, Lexical gestures and lexical access: a process model, in: D. McNeill (Ed.), *Language and Gesture* (261–283), Cambridge University Press, Britain, 2000.
- [117] Y.U. Wenhua, L.U. Zhongyi, A new perspective on the cognitive function of gestures: The spatializing gesture hypothesis, *Adv. Psychol. Sci.* 28 (3) (2020) 426–433.
- [118] G. Buccino, F. Binkofski, G.R. Fink, et al., Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study, *Eur. J. Neurosci.* 13 (2) (2001) 400–404.
- [119] M.A. Arbib, From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics, *Behav. Brain Sci.* 28 (2) (2005) 105–124.
- [120] G. Rizzolatti, R. Camarda, M. Fogassi, M. Gentilucci, G. Luppino, M. Matelli, Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements, *Exp. Brain Res.* 71 (3) (1988) 491–507.
- [121] V. Raos, M.-A. Umilta, A. Murata, L. Fogassi, V. Gallese, Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey, *J. Neurophysiol.* 95 (2) (2006) 709–729.
- [122] M.A. Umilta, T. Brochier, R.L. Spinks, R.N. Lemon, Simultaneous recording of macaque premotor and primary motor cortex neuronal populations reveals different functional contributions to visuomotor grasp, *J. Neurophysiol.* (2007).
- [123] J. Gong, L.G. Foo, Y. He, et al., Llms are good sign language translators, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18362–18372.
- [124] S. Fang, L. Wang, C. Zheng, et al., SignLLM: Sign languages production large language models, 2024, arXiv preprint arXiv:2405.10718.