WILEY

SmartBot

REVIEW ARTICLE OPEN ACCESS

A Comprehensive Review of Humanoid Robots

Qincheng Sheng[1] | Zhongxiang Zhou[2] | Jinhao Li[2] | Xiangyu Mi[2] | Pingyu Xiang[2] | Zhenghan Chen[2] | Haocheng Xu[2] | Shenhan Jia[2] | Xiyang Wu[2] | Yuxiang Cui[2] | Shuhao Ye[2] | Jiyu Yu[2] | Yuhan Du[2] | Shichao Zhai[2] | Kechun Xu[2] | Yifei Yang[2] | Zhichen Lou[2] | Zherui Song[2] | Zikang Yin[2] | Yu Sun[2] | Rong Xiong[2] | Jun Zou[1] | Huayong Yang[1]

[1]State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China | [2]State Key Laboratory of Industrial Control Technology, School of Control Science and Engineering, Zhejiang University, Hangzhou, China

Correspondence: Zhongxiang Zhou (zhouzhongxiang@zju.edu.cn) | Rong Xiong (rxiong@zju.edu.cn) | Jun Zou (junzou@zju.edu.cn)

Received: 2 December 2024 | **Revised:** 13 January 2025 | **Accepted:** 5 February 2025

Funding: The authors received no specific funding for this work.

Keywords: challenges and trends | humanoid robot | humanoid robot components | key technologies

ABSTRACT

Humanoid robots, increasingly recognized for their potential to drive economic and social development, have garnered significant attention in recent years. This paper aims to provide a comprehensive overview of the progress, challenges, and future directions in humanoid robotics, with a particular emphasis on essential system components and key technological innovations. Through a review of historical milestones, this paper explores critical aspects such as the design of the head and body, and examines state-of-the-art technologies in areas like locomotion control, perception, and intelligent manipulation. By presenting a thorough analysis of the field, this work aims to serve as a valuable resource for researchers and inspire future innovations that will drive the continued evolution of humanoid robots.

1 | Introduction

Humanoid robots are a critical focus in national technological advancement, representing a promising emerging industry that is, the key to future economic and industrial competition. These robots will have a profound impact on national economic and social development. By integrating frontier technologies—such as advanced manufacturing, novel materials, and artificial intelligence—humanoid robots aim to emulate and eventually surpass human capabilities. Through continuous evolution and adaptation, humanoid robots will integrate smoothly into existing social systems. As a result, humanoid robots are expected to become a driving force in the transformation and upgrading of the manufacturing industry, both in China and globally, enhancing societal productivity and improving people's quality of life.

The technology and industry surrounding humanoid robots are advancing rapidly worldwide. In 2022, Tesla's Optimus made a landmark debut, reshaping perceptions and expanding functional boundaries for humanoid robots. Its unveiling ignited global discussions and sparked imaginations about the potential for general-purpose applications in robotics. In 2023, rapid advancements in large language models accelerated the emergence of embodied intelligence. As a prominent example of embodied intelligence, humanoid robots have become an ideal testing ground to validate large models' abilities to interact effectively with the physical world, further driving advancements in humanoid robotics. Today, with cross-industry integration between tech giants and traditional enterprises, along with collaboration among academia, industry, government, and capital, humanoid robots are at the forefront of industrialization. The application-driven development of humanoid robot

Qincheng Sheng, Jinhao Li, Shenhan Jia, Yuxiang Cui, and Kechun Xu contributed to the work equally and should be regarded as co-first authors.

products, coupled with the growth of a supporting component industry, is flourishing, revealing immense market potential.

Globally, the commercialization of humanoid robots remains in an early phase. For humanoid robots to achieve widespread practical application, they must overcome several key technical challenges:

1. Stably Whole-body Control. Humanoid robots, as inherently unstable systems, rely on intermittent foot contact with the ground, creating minimal time and spatial domains for stability control. Stability is further challenged by factors such as dynamic model inaccuracies, uneven terrain, changing tasks or loads, and human-robot interactions. This makes it essential to develop technology capable of achieving human-like, full-body coordinated motion to maintain balance and stability.
2. Intelligence Upgradation. As versatile systems, humanoid robots should possess general-purpose intelligence that adapts to diverse environments and tasks. This represents a major challenge and key focus in artificial intelligence development, necessitating advanced abilities in semantic perception, environmental and object comprehension, and highly generalized human-like skill learning.
3. Human-Robot Interaction: Aiming to integrate into human society and fulfill roles in household settings, humanoid robots encounter substantial challenges in emotional expression and interaction. Effective interaction requires robots to interpret human emotions accurately while also expressing their own through facial movements, vocal tones, and gestures. Realizing such natural emotional engagement necessitates not only sophisticated perception and control technologies but also the seamless integration of multifunctional systems to enable genuine, human-like emotional responses.
4. System Integration: For reliable, sustained operation in complex and dynamic settings, enhancements in mobility, flexibility, reliability, load capacity, and battery life are required. These demands place high expectations on component performance, size, and weight, while necessitating the integration of hundreds of parts and seamless hardware-software fusion. Achieving these calls for compact, lightweight, and highly integrated design solutions.

This paper is organized as follows: Section 1 outlines the study's scope and motivations. Section 2 provides a review of the historical development of humanoid robots. Sections 3 and 4 focus on robot components—Section 3 examines head designs (both non-anthropomorphic and anthropomorphic), while Section 4 delves into body components, covering both hardware and software structures. Section 5 discusses key technologies such as locomotion control, perception, navigation, and intelligent manipulation, with an emphasis on both model-based and learning-based approaches. Section 6 explores application scenarios, ranging from industrial workers to disaster rescuers. Section 7 discusses challenges and trends, including design, human-robot interaction, security, cost, modularity, and embodied intelligence. The technological framework of humanoid robot is illustrated in Figure 1.

2 | Development of Humanoid Robots

The first bipedal robot was created in 1969 by Ichiro Kato from Japan's Waseda University. It featured only the lower body with bipedal legs, powered by hydraulics and cable traction for static walking. By 1972, the first full-body humanoid robot was developed. Since then, the evolution of humanoid robots has gone through several phases: from foundational theoretical methods and system formation to the rapid advancement of key technologies and system diversification. Currently, humanoid robots have entered an industrialization phase driven by applications and oriented toward products, with the potential to achieve low-cost, large-scale production and application in the future. During this development journey, numerous representative prototypes have emerged both domestically and internationally.

Honda began humanoid robot research in 1986, and after a decade, developed the P2 model, which featured 30 degrees of freedom (DOF) and a walking speed of 2 km/h. In 2000, Honda introduced ASIMO, with 26 DOF and a walking speed of 1.6 km/h, later achieving speeds of 2 km/h while walking and 6 km/h while running by 2005. Between 2005 and 2018, ASIMO's capabilities continued to evolve, reaching a running speed of 9 km/h and acquiring skills such as hopping on one or two legs, climbing stairs, and integrating modules for intelligent perception, manipulation, and interaction. ASIMO demonstrated impressive tasks like kicking a ball, unscrewing bottle caps, pouring water, serving tea, and following people. However, due to high production costs and limited scalability, the ASIMO project was officially discontinued in 2018.

Boston Dynamics' work on bipedal walking theory dates back to 1986, when its founder, Marc Raibert, introduced the concept of dynamic balance control. The company began formal research on humanoid robots in 2009, developing Petman, a robot designed to test U.S. military chemical suits. In response to the failure of robots to perform tasks during the Fukushima nuclear disaster, Boston Dynamics released the humanoid robot Atlas in 2013, designed for disaster rescue and special operations, with the ability to walk stably on rubble-filled terrain. Over the next decade, the company enhanced Atlas's dynamic movement capabilities, achieving breakthroughs in virtual model control, nonlinear model predictive control, and full-body control. This progress enabled Atlas to perform complex actions such as parkour, triple jumps, whole-body coordinated dances, running and jumping onto steps, and walking on a balance beam. In April 2024, Boston Dynamics unveiled a fully electric version of Atlas, which can rise from the ground and features joints capable of rotating 360°.

Tesla introduced its humanoid robot Optimus prototype in 2022, standing about 173 cm tall and weighing approximately 56 kg. The robot demonstrated basic tasks such as walking, watering plants, and moving metal rods. After two iterations, in 2023, the

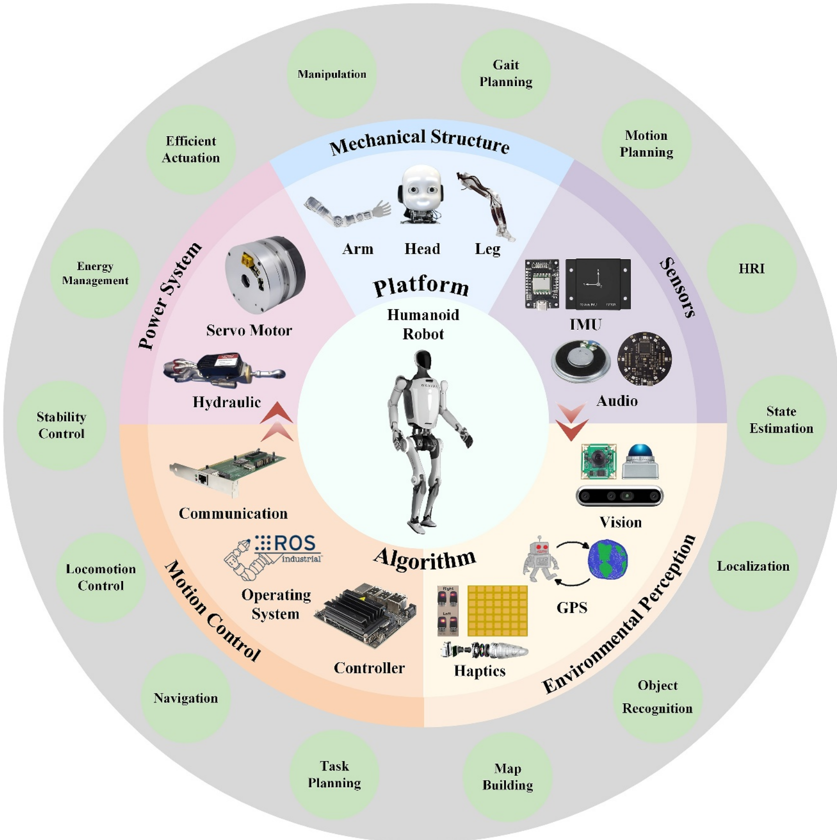

FIGURE 1 | Technological framework of humanoid robot.

Optimus Gen2 achieved stable walking in factory environments, the ability to stand still while grabbing objects, learning human movements through mapping, and using autonomous driving technology for environmental perception and intelligent navigation. Future plans include deploying Optimus on a large scale in industrial, service, and household applications.

The U.S. company figure.ai released the humanoid robot Figure 01, which stands 150 cm tall and weighs 60 kg. The robot demonstrated capabilities such as making coffee and moving materials into bins. On March 1, 2024, figure.ai announced a collaboration with OpenAI to integrate multimodal large models into Figure 01, enabling it to achieve desktop environment perception, facilitate smooth human-robot dialog, and perform tasks such as placing a black plastic bag into a box and arranging cups and plates on a draining rack based on a neural network strategy model.

China began research on humanoid robots in the 1990s, with several institutions, including the National University of Defense Technology, Harbin Institute of Technology, Tsinghua University, Beijing Institute of Technology, and Zhejiang University, making significant research achievements. In recent years, domestic universities, research institutes, and technology companies have increasingly engaged in humanoid robot research and industrialization.

In terms of universities, Beijing Institute of Technology started its humanoid robot research in 2000 and released the BHR-1 in 2001, which could walk independently without external cables, reaching a maximum walking speed of 2 km/h. After several iterations, it gained abilities such as performing Tai Chi, playing table tennis with humans, and fall protection. From 2008 to 2011, Zhejiang University developed "Wukong I," which could play hundreds of rounds of table tennis with humans or another robot, gaining international attention. It was featured by outlets such as the Associated Press, Reuters, and National Geographic, and included in a summary report to the U.S. President by the National Science Foundation. "Wukong IV," a humanoid robot developed later, achieved a maximum speed of over 6 km/h, a high jump of 0.5 m, and the ability to climb 25° slopes and 10 cm steps. It could adapt to various terrains, including outdoor surfaces, grass, mud, stairs, and slopes, and also had environmental perception and autonomous navigation capabilities.

On the enterprise side, UBTECH Robotics began developing humanoid robot products in 2012, releasing the Walker series humanoid robots, which can walk stably across multiple terrains, climb stairs, and navigate autonomously. In 2023, Unitree launched the H1, a general-purpose humanoid robot standing approximately 180 cm tall and weighing about 47 kg. The H1 is capable of fast running, stair climbing, and resisting external disturbances. In 2024, Unitree released its second humanoid robot, the G1, standing around 127 cm tall and weighing approximately 35 kg, which can integrate a three-finger force-controlled dexterous hand and demonstrate flexible full-body movement abilities such as getting up and jumping. Additionally, companies like Xiaomi, Agibot, Fourier, Dreame, and Xpeng have also unveiled humanoid robot prototypes or products.

3 | Components of Humanoid Robot Head

As humanoid robots integrate into society and domestic settings, public acceptance is crucial for their adoption. The robot's head, especially the face, plays a key role in conveying emotions, providing social cues, and building trust in human-robot interactions. Effective design must balance functionality and psychological comfort, addressing or surpassing the uncanny valley [1]. A well-designed face shapes first impressions, facilitates communication, and adapts to cultural norms, ensuring successful integration into daily life. Therefore, this paper begins by introducing the components of humanoid robot head. Due to the uncanny valley effect, two distinct design approaches have emerged. One approach intentionally designs humanoid robot heads to look less human in order to avoid triggering the uncanny valley, which can evoke discomfort or fear in people. The other approach aims to make humanoid robot heads as human-like as possible, with the goal of surpassing the uncanny valley and increasing the robot's acceptance and likability [2].

As the level of anthropomorphism in humanoid robot heads increases, they progressively develop more complex human-like features, starting from non-human facial characteristics, then advancing to basic human traits such as eyebrows, eyes, nose, and mouth, and ultimately including highly realistic details like lifelike skin and hair. Concurrently, their facial expressions become more nuanced, enhancing their human-robot interaction capabilities. Based on the degree of anthropomorphism, humanoid robot heads can be categorized into two types: non-anthropomorphic and anthropomorphic. Non-anthropomorphic humanoid robot heads exhibit few or no human-like features, making them easily recognizable as robots. In contrast, anthropomorphic humanoid robot heads possess all human facial features, often making it difficult to distinguish them from actual humans based on appearance alone. The comparison of different types of humanoid robot heads is shown in Figure 2.

The distinct appearances of these two types of humanoid robot heads make them suitable for different application scenarios. Non-anthropomorphic humanoid robot heads are more commonly employed in industrial manufacturing, commercial services, family companionship, healthcare, and education. In contrast, anthropomorphic humanoid robot heads are more often used in roles such as guides, receptions, commercial promotions, shows, customer service, and similar fields. In this section, we will discuss these two types of robots in detail.

3.1 | Non-Anthropomorphic Humanoid Robot Head

Non-anthropomorphic humanoid robot heads can be categorized into two types based on the presence of human facial features. The first type is entirely non-human in appearance, such as humanoid robots that use electronic screens as face. The second type incorporates some human-like features, such as facial traits, but lacks lifelike skin and hair. This type of humanoid robot head often features a face designed in a cartoonish, mechanical, or science fiction-inspired style. The comparison of non-anthropomorphic humanoid robot heads is shown in Table 1.

The first type of non-anthropomorphic humanoid robots is primarily used in industrial settings, performing practical tasks such as autonomous delivery or factory logistics, with no emphasis on interactive capabilities. As a result, human-like characteristics, such as facial features, are intentionally omitted to avoid triggering the uncanny valley effect. Functionally, their heads serve two purposes: decoration and housing essential sensors, such as lidar, depth cameras, stereo sensors,

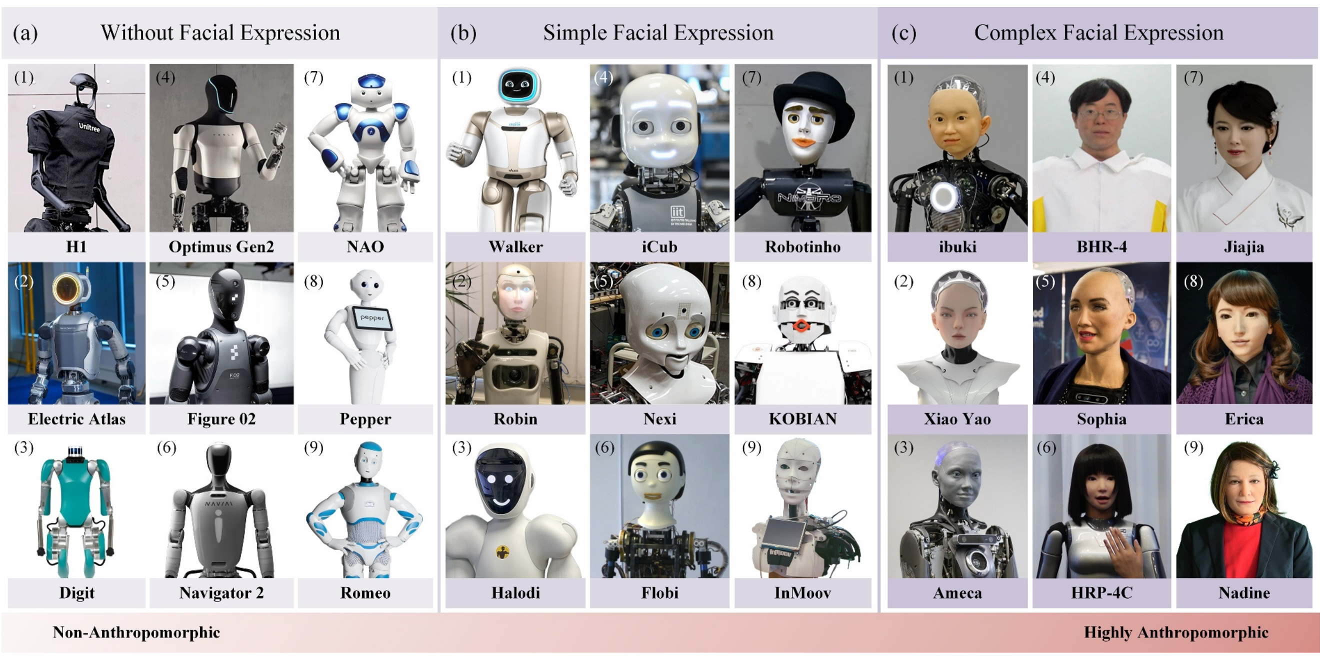

FIGURE 2 | Different types of humanoid robot heads.

TABLE 1 | Comparison of non-anthropomorphic humanoid robot heads in Figure 2.

Robot name	Year	Head DOFs	Face features	Reference
H1	2024	0	No facial features	[3]
Electric Atlas	2024	Neck: 1	No facial features	[4]
Digit	2023	0	LED eyes	[5]
Optimus Gen2	2023	Neck: 2	No facial features	[6]
Figure 02	2024	Neck: 2 or 3	No facial features	[7]
Navigator 2	2024	Neck: 2	No facial features	[8]
Walker	2019	Neck: 2	Digit screen face	[9]
Robin	2011	Neck: 3	Projected screen face	[10]
Halodi	2020	Neck: 1	LED screen face	[11]
iCub	2004	Eyes: 3; neck: 3;	Mechanical eyes with LED mouth and eyebrows	[12]
NAO v6	2018	Neck: 2	Fixed face	[13]
Pepper	2014	Neck: 2	Fixed face	[14]
ROMEO	2014	Neck: 2; eyes: 2×2	Fixed face	[15]
Nexi	2008	Eyebrows: 4; eyes: 3; eyelids: 4; mouth: 1; neck: 4	Mechanical face	[16]
Flobi	2009	Eyebrows: 2; eyes: 3; eyelids: 4; mouth: 6; neck: 3	Mechanical face	[17]
Robotinho	2009	Eyebrows: 4; eyes: 4; eyelids: 2; mouth: 5; neck: 2	Mechanical face	[18]
KOBIAN	2012	Eyebrows: 8; eyes: 3; eyelids: 5; mouth: 8; neck: 4	Mechanical face	[19]
InMoov	2012	Eyes: 4; eyelids: 2; mouth: 1; neck: 2	Mechanical face without eyebrow	[20]

microphones, and speakers, to support other operations. These robots do not even require a human-like head shape. For example, Unitree's H1 humanoid robot head serves as a mounting platform for a depth camera [3] (see Figure 2a1), Boston Dynamics' Electric Atlas features a large circular display with a ring of lights [4] (see Figure 2a2) and Agility Robotics' Digit robot uses a cylindrical head [5] (see Figure 2a3). However, for esthetic reasons and wider acceptance, an increasing number of humanoid robots are being designed with more human-like head shapes, such as Optimus Gen2 [6] (see Figure 2a4), Figure 02 [7] (see Figure 2a5), Navigator 2 [8] (see Figure 2a6), CyberOne [21] and other humanoid robots.

Although this type of humanoid robots lacks facial features, many companies have employed alternative methods to express the robot's status and enhance human-robot interaction. Agility Robotics' Digit robot, for example, uses two LED-animated eyes to convey information, such as indicating changes in direction and other actions during operation. Optimus Gen 2 humanoid robot utilizes light strips to signal working status, red for errors and green for normal status. Figure 02 humanoid robot incorporates an LCD screen that displays dynamic icons based on interactions, improving the user experience. CyberOne humanoid robot uses a 2D curved OLED screen for visual communication. In addition, many other humanoid robots also use similar settings, such as Neo [22], Walker-S [23], etc.

The second type of non-anthropomorphic humanoid robots is typically used in fields such as reception, service, education, and healthcare, where interaction with users is essential. These robots have some human-like facial features and can make simple expressions, but they lack lifelike skin and hair. A straightforward method for achieving facial expressions is to use digital faces displayed via screens, projectors, or LEDs. Robots like Walker [9] (see Figure 2b1), Arash [24], and Iromec [25] use display screens to express emotions. Some robots, such as Robin [10] (see Figure 2b2), Mirokai [26], and Furhat [27], project animated faces onto masks to create dynamic facial expressions. LEDs are another approach, as seen in Halodi [11], which uses an entire LED screen for its face (see Figure 2b3), and iCub [12] which combines a cartoon facial shape with LED eyebrows and mouth to adjust expressions and convey emotions (see Figure 2b4).

However, the expressions on digital faces tend to lack realism, which can limit effective interaction. To facilitate better communication, researchers have developed mechanical structures to enable dynamic facial expressions. These robots typically feature a cartoon-like appearance to enhance approachability and foster a desire for communication. For example, robots like NAO [13] (see Figure 2a7), PEPPER [14] (see Figure 2a8), and ROMEO [15] (see Figure 2a9) use cartoonish faces with fixed smiling expressions, while Nexi [16] (see Figure 2b5) and Flobi [17] (see Figure 2b6) use cartoonish appearances with more dynamic faces. In addition to appearance, non-verbal communication, particularly facial expressions, is crucial.

Alternatively, the face can also be designed to resemble human features more closely. Since these non-anthropomorphic robot heads lack lifelike skin, facial features are constructed as separate mechanical components attached directly to the face mask, such as Robot Robotinho [18] (see Figure 2b7), KOBIAN [19] (see Figure 2b8) and InMoov [20] (see Figure 2b9).

3.2 | Anthropomorphic Humanoid Robot Head

The anthropomorphic humanoid robot head, typically fitted with lifelike artificial skin and hair, can closely resemble a human, often making it difficult to distinguish from an actual person. This type of robot, also known as Android Robot, is designed to convey emotions more naturally and enhance human-robot interaction through its realistic appearance. The comparison of anthropomorphic humanoid robot heads is shown in Table 2.

The development of anthropomorphic humanoid robot heads progresses from a focus on physical appearance to the integration of advanced behaviors and emotions, moving through three main stages, as shown in Figure 3. In the first stage, the focus is on achieving a human-like appearance with realistic skin and facial features—a milestone most anthropomorphic robot heads have already achieved. In the second stage, the robot head acquires anthropomorphic movements, including eye contact, lip synchronization, and natural expressions, to enable more fluid and coordinated interactions. Most current designs are in this stage, demonstrating basic but improving capabilities for dynamic interaction. In the third stage, the robot head develops an anthropomorphic psychology. By Using an emotion model, it can adjust its "mental state" based on its environment, facilitating more natural and responsive emotional interactions with humans. This stage remains in the experimental phase, facing challenges in achieving real-time psychological and emotional adaptability. This section will examine the evolution of anthropomorphic humanoid robot heads across these three stages.

3.2.1 | First Stage: Anthropomorphic Appearance

In the first stage, an anthropomorphic humanoid robot head requires a highly human-like appearance, particularly a realistic face capable of expressing delicate facial movements and adjusting in real-time to external changes. Achieving this involves not only lifelike skin but also a mechanical structure to drive skin movement and a sensor system to receive external stimuli.

3.2.1.1 | Realistic Skin. There are currently several options for creating realistic skin. The most common materials include rubber or silicone. Some robots, like Affeto, use polyurethane elastomer gel [43], while SAYA uses urethane resin [44] for facial skin. However, these materials can be challenging to control; overly soft materials may cause the face to collapse, while harder materials may limit facial flexibility. To address

TABLE 2 | Comparison of anthropomorphic humanoid robot heads in Figure 2.

Robot name	Year	Head DOFs	Face features	Emotion model
Ibuki [28]	2018	Eyebrows: 3; eyes: 3; eyelids: 2; mouth: 7; neck: 3	Childlike Face 15 DoFs 18 AUs	/
Xiao Yao [29]	2024	Eyebrows: 4; eyes: 2; eyelids: 4; mouth: 13; neck: 3	Female Face 23 DoFs 19 AUs	/
Ameca [30]	2021	Eyebrows: 4; nose: 1; eyes: 4; eyelids: 4; mouth: 14; neck: 5	Female Gray face Face 27 DoFs	Empathic mechanized anthropomorphic humanoid system
BHR-4 [31]	2010	Eyebrows: 2; cheek 2; eyelids: 6; mouth: 3; neck: 2	Male Face 13 DoFs 14 AUs	/
Sophia [32]	2016	Face 27; neck: 5	Female face	/
HRP-4C [33]	2009	Eyebrows: 1; eyes: 2; cheek 1; eyelids: 1; mouth: 3; neck: 3;	Female Face 8 DoFs	/
JiaJia [34]	2016	/	Female	/
Erica [35]	2015	Eyebrows: 2; eyes: 3; eyelids: 2; mouth: 6; neck: 3	Female Face 13 DoFs	Emotional voice conversion model
Nadine [36]	2013	Face: 7; neck: 3	Female Face 7 DoFs	Personalized-characterized mood dynamics model

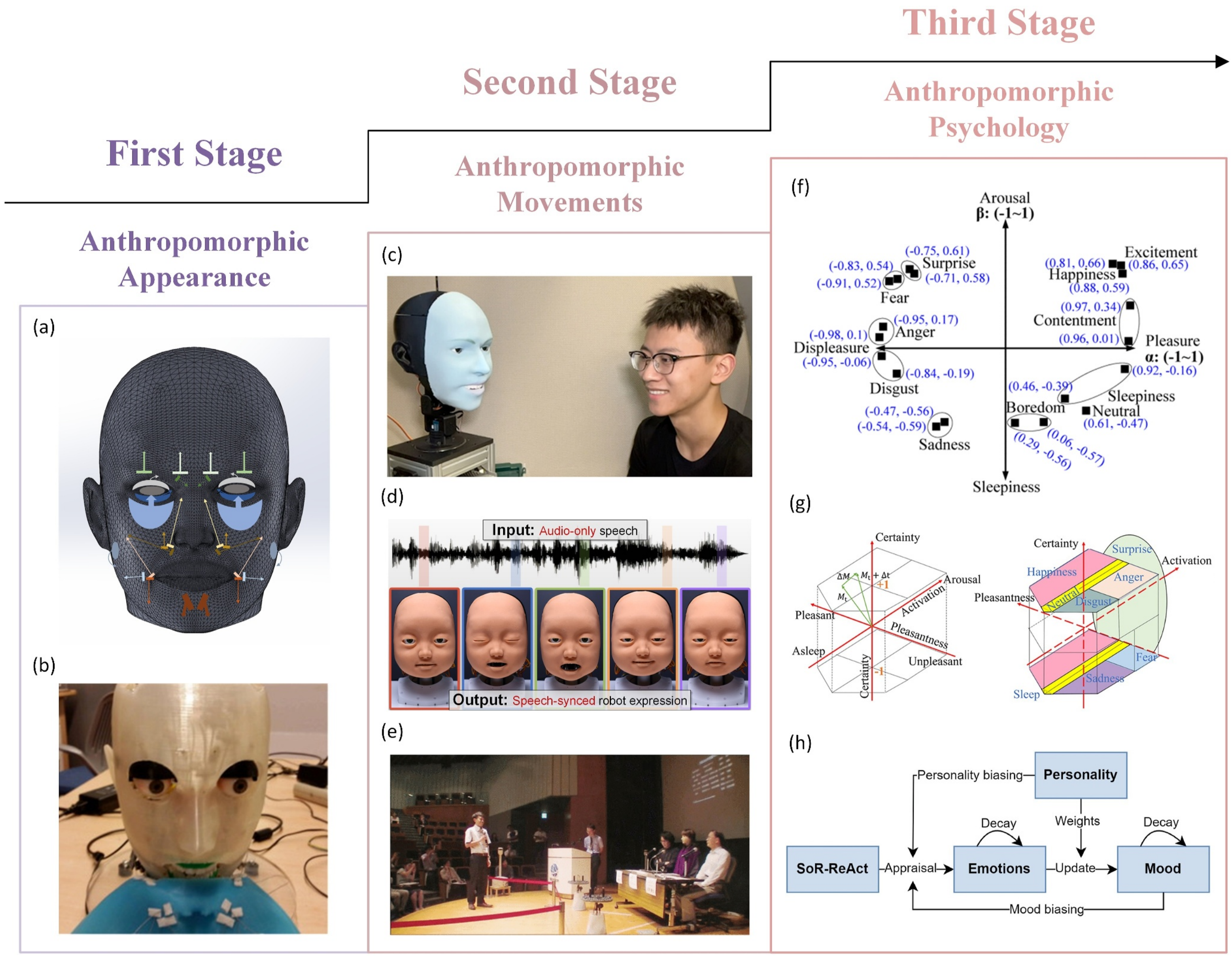

FIGURE 3 | Developmental stages of the anthropomorphic humanoid robot head: (a) 15 AUs of the EVA robot and their movement directions. Reproduced with permission [37]. Copyright 2020, The Authors. (b) Distribution of steel wires on the EVA robot's silicone mask. Reproduced with permission [37]. Copyright 2020, The Authors. (c) EMO robot coexpression. Reproduced with permission [38]. Copyright 2024, The American Association for the Advancement of Science. (d) Dynamic animatronic facial expressions generated by speech. Reproduced with permission [39]. Copyright 2024, IEEE. (e) ERICA public demonstration. Reproduced with permission [35]. Copyright 2016, IEEE. (f) Facial expression distribution based on the PA emotion model. Reproduced with permission [40]. Copyright 2013, IEEE. (g) WE-4 robot emotion model. Reproduced with permission [41]. Copyright 2023, Science China Press. (h) Nadine robot psychodynamic architecture. Reproduced with permission [42]. Copyright 2024, The Authors.

this, researchers have developed Frubber [45], a skin material specifically designed for robot faces, made from a mixture of sponge-like artificial rubber and foam. With moderate hardness, Frubber allows for natural expressions such as smiles and frowns, making it highly suitable for robot faces.

3.2.1.2 | Actuation Methods. The mechanical structure that drives skin movement is typically based on the Facial Action Coding System (FACS), developed by Ekman et al. [46]. FACS divides facial muscle movements into 44 basic action units (AUs), which can be combined to create a wide range of expressions. Using this model, an anthropomorphic robot head employs control points linked to these AUs, enabling the robot to accurately simulate expressions by moving these control points.

There are three primary methods to drive these control points: electric, pneumatic, and soft material deformation. Electric drives rely on motors and servos, using gears, connecting rods, pulleys, and cables to move control points in specific directions. For example, Ameca [30] uses motors to drive connecting rods for 19 control points (covering the eyebrows, eyelids, nose, and mouth), while EVA [37] uses servos and lines to control 15 points on the face (see Figure 3a,b). Pneumatic drives utilize pneumatic actuators behind the skin to directly pull control points and create facial expressions. For instance, SAYA uses McKibben pneumatic actuators, which simulate human muscles, to control the movement of 19 facial points, while Nikola uses pneumatic actuators with pressure control valves [47]. Although both electric and pneumatic methods offer precise control, pneumatic actuators are often noisy, and electric actuators take up substantial head space due to their weight and size, limiting the robot head's freedom of movement. To address these limitations, researchers have explored lightweight and compact soft actuators for controlling facial expressions. For example, Tadesse et al. [48] used shape memory alloys (SMAs) to drive eye and mouth movements; Pioggia et al. [49] used dielectric elastomer actuators (DEAs) to achieve seven basic expressions, and Y. Li et al. [50] utilized multilayered PVC gel actuators (MPGAs) to control eye and jaw movements. However, these materials still require improvements to achieve rapid expression changes and precise control.

Additionally, to better perceive the external environment, the humanoid robot head incorporates various sensors that simulate human senses, including visual and auditory sensors, microphones, tactile sensors, and even olfactory sensors. With lifelike skin, an advanced movement mechanism, and a suite of sensory inputs, the robot head can further perform anthropomorphic movements more realistically.

3.2.2 | Second Stage: Anthropomorphic Movements

In the second stage, the anthropomorphic humanoid robot head can perform natural human-like movements, including eye contact, lip synchronization, and expressive facial gestures.

3.2.2.1 | Eye Contact. Eye contact, in particular, is essential. Research by Kiilavuori et al. [51] found that eye contact with a robot can evoke similar emotional and attentional responses as eye contact with a human, making the robot's gaze critical for building trust and a sense of closeness during human-robot interactions. To achieve natural eye contact, the robot's eye movements must be both coordinated and responsive. On one hand, natural eye movement requires coordination with neck movement, ensuring that the speed of eye and neck rotation aligns with human psychophysical principles. Pateromichelakis et al. [52], for example, simulated the human vestibular-ocular reflex in the Romeo robot, allowing its eyes to stay fixed on a target even when the head moves. On the other hand, the robot's eye focus should adapt dynamically to its environment. In dyadic interaction, for instance, the robot's sightline is typically focused on the user. Raković et al. [53] developed a gaze dialog model for the iCub robot, enabling it to alternate between leader and follower roles, infer human behavior, and adjust its point of gaze accordingly. In multi-person interactions, the gaze direction should shift to follow the speaker, fostering effective communication. Zaraki et al. [54] designed a gaze control system for the FACE robot that uses an attention-based mechanism to direct gaze according to verbal and non-verbal cues, proving effective in multi-person conversations.

3.2.2.2 | Lip Synchronization. The second aspect is lip synchronization. The McGurk effect demonstrates that lip movements influence auditory perception [55], making the alignment of lip shape with audio essential for authentic human-robot interactions. Achieving this requires not only designing a mechanical mouth structure that can replicate various human mouth shapes but also establishing a mapping relationship between audio and the mouth movements of the humanoid robot head. Strathearn and Ma [56] developed an innovative robot mouth structure capable of opening, contracting, and raising or lowering the lip corners, closely mimicking human lip movements. For mapping audio to mouth movements, researchers can draw from lip synchronization techniques used in computer graphics animation. For example, Heisler et al. [57] compared deep learning models, VisemeNet (based on visemes) and FaceFormer (based on mesh structures), for Android robots and found the mesh-based method to be most effective for achieving realistic lip synchronization.

3.2.2.3 | Expression Generation. The last aspect focuses on enabling robots to make anthropomorphic expressions based on external information they receive. By processing and analyzing signals from different sources, robots determine appropriate expressions to convey emotions. Expression generation methods can be categorized into four types based on signal sources: visual, audio, tactile, and multimodal data.

Visual signals are a primary source for generating expressions, with current research emphasizing the simulation of human expressions. Hu et al. [38] developed two deep learning models for the Emo robot: one model predicts human facial expressions, and the other generates motion commands, allowing the robot to show co-expressions with human in real time (see Figure 3c).

Audio signals, particularly emotional cues in speech, also play a key role in expression generation. While audio driven expression is well-developed in computer graphics animation, complex humanoid robot head structures pose challenges for direct adaptation. To address this, B. Li et al. [39] proposed a design and motion synthesis approach based on linear blending skinning (LBS), enabling real-time, speech-driven facial expression generation with greater accuracy, efficiency, and naturalness (see Figure 3d).

Tactile signals offer another channel for robots to generate expressions. By using tactile sensors, robots can detect the strength and the location of touch, allowing for responsive facial expressions. Martinez-Hernandez et al. [58] developed a Bayesian framework to interpret different types of touch, which is then translated into emotional expressions through an emotional control architecture, allowing the iCub robot to dynamically adjust expressions based on tactile input.

Multimodal data combines signals from vision, audio, and text, providing robots with comprehensive information for generating expressions. Glas et al. [35] integrated multimodal data (vision, audio, and position) into the ERICA robot, allowing it to accurately locate the speaker, adjust eye gaze and expressions, and provide natural verbal and non-verbal feedback, enhancing its anthropomorphic performance and interaction quality in human-robot communication (see Figure 3e). Although robots can generate expressions from external signals, they lack emotional cognition and complex emotional management. Establishing a robot emotion model is essential for robots to not only produce immediate expressions in continuous interactions but also to adjust their emotional states based on environmental feedback, achieving genuine "emotional intelligence."

3.2.3 | Third Stage: Anthropomorphic Psychology

In the third stage, the anthropomorphic humanoid robot head incorporates anthropomorphic psychology. To achieve this, an emotion model and emotion state management system are developed, allowing the robot to simulate human emotional responses when processing external information.

3.2.3.1 | Types of Emotion Models. The emotion model classifies emotions, enabling the robot to understand and express distinct emotional characteristics, such as happiness, anger, and sadness. The emotion models include discrete, dimensional and other specialized types. In discrete emotion models, each emotion is independent, as seen in Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [59]. Dimensional emotion models treat emotions as continuous, showing their similarity and differences in dimensional space. Notable dimensional emotion models include Russell's two-dimensional valence-arousal (VA) model, where "valence" reflects emotion polarity, like positive or negative, and "arousal" represents intensity [60]. Building on this, Mehrabian's three-dimensional pleasure-arousal-dominance (PAD) emotion model introduces "dominance," representing how much a person feels in control or, alternatively, submissive and controlled [61]. Additionally, Plutchik's emotion wheel model presents eight basic emotions along with their gradients and combinations. Building on this, his cone model represents emotion intensity on the cone's vertical axis, with eight sectors for each basic emotion [62]. Other emotion models include the Ortony, Clore and Collins's (OCC) model [63] and hidden Markov emotion model [64], etc.

3.2.3.2 | Emotion Systems of Humanoid Robot Heads. Based on the emotion model, an emotion state management system refines emotional expression by combining real-time interaction data and the robot's set personality (often using McCrae's five personality traits [65]). For example, Han et al. [40] proposed an emotion expression generation method for robots based on a two-dimensional PA emotion model (pleasure-arousal), combining user emotional intensity with the robot's personality traits to produce natural expressions using the fuzzy Kohonen clustering network (FKCN) for smooth emotional transitions (see Figure 3f). Similarly, Miwa et al. [41] developed a psychological model for the humanoid robot WE-4, creating a three-dimensional emotional space (pleasure, activation, and certainty). This system allows the robot to generate instant responses, sustain long-term emotional tones, and dynamically adjust intensity based on external stimuli, providing complex and personalized emotional responses (see Figure 3g). J. Zhang et al. [42] proposed a personalized-characterized mood dynamics (PCMD) model using a three-dimensional PAD emotion model, applied to the Nadine robot [36]. In this model, personality acts as an emotional baseline, while emotions and moods shape immediate responses and long-term states. The system uses convex optimization to calculate emotional weights, integrates emotions into mood shifts, and enables the robot to express emotions consistently with its personality over multiple interactions, enhancing personalized emotional expression (see Figure 3h).

4 | Components of Humanoid Robot Body

Humanoid robots are designed to replicate human physical and cognitive abilities, and as such, their components are a sophisticated blend of hardware and software systems. In this section, we explore the fundamental components of humanoid robot body, which include their hardware architecture—comprising mechanical structure, sensor systems, and power systems—and their software architecture, which governs the robot's operation and communication.

4.1 | Hardware Architecture

4.1.1 | Body Mechanical Structure and DOF Analysis

The concept of humanoid robots was introduced to develop a versatile robot capable of imitating human functionality, requiring resemblance to human appearance as well as human-like dynamic and operational abilities. In the design of humanoid robots, the mechanical structure plays a vital role in determining the robot's versatility and efficiency in replicating human-like movements. The degrees of freedom (DOF) is a crucial metric in defining the robot's range of motion and its ability to perform complex tasks. To achieve the versatility, researchers have designed the mechanical structure of humanoid robots inspired by the human skeletal system, typically allowing 20–40 degrees of freedom for replicating human movements [66]. And as shown in Figure 4, hardware architecture of a humanoid robot includes mechanical structure, sensor and power system. Due to the limitations of current actuator technology, material characteristics, and control methods, this complex electromechanical system requires optimization in terms of mass distribution and mechanical resonance suppression rather than a direct replication of human anatomy [67]. This optimization process entails, firstly, reducing the number of actuators as much as possible to minimize weight and concentrating mass around the torso—typically near the waist—to reduce the inertia of the limbs. Secondly, it involves restricting flexibility at joints and connection points to maintain control precision, ensuring stable and reliable motion.

4.1.1.1 | Joint Design. Joints are crucial for humanoid robot mobility, serving as the primary elements driven by actuators. Rather than installing actuators, such as motors, directly at the joints, a common approach is to use cascaded reduction gears or harmonic drives to increase torque, enhancing joint movement efficiency. In many cases, the actuators are mounted remotely to drive the joints at a distance, utilizing cables, belts, lead screws, or various linkage mechanisms such as cranks and four-bar linkages. Although these mechanisms allow for spatial separation between the actuator and the joint, they also introduce non-linearities that make precise control more challenging. Joints can be categorized based on degrees of freedom (DOF), including single-degree and multi-degree types; the latter can be further divided into serial and parallel configurations to meet specific mechanical needs [68].

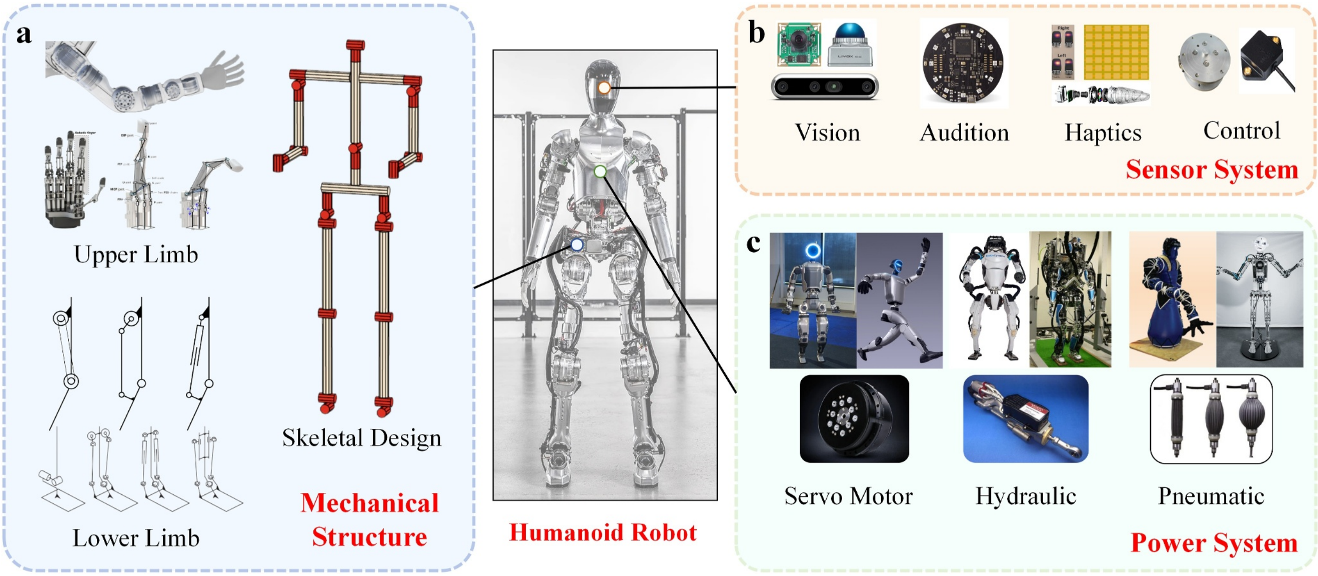

FIGURE 4 | Hardware architecture of a humanoid robot. (a) Mechanical structure design. (b) Sensor system. (c) Power system.

4.1.1.2 | Upper Limb Mechanical Structure Design. The upper limb comprises both the arm and the gripper. Advanced humanoid robots like Unitree G1 [69] and OpenAI Figure 02 [7] are designed with arms that have seven degrees of freedom (7-DOFs) to avoid singularities in movement, ensuring a wider range of motion and flexibility. For grippers, most humanoid robots adopt multi-fingered dexterous hands. When it comes to dexterous hands mounted on the upper limb, additional degrees of freedom are introduced, typically ranging from 6 to 20 degrees of freedom per hand. Due to constraints such as hand size and load-bearing requirements, each finger generally has one to two degrees of freedom, with the actuators integrated within the palm. The fingers are driven by a four-bar linkage mechanism, allowing the MCP (Metacarpophalangeal Joint), PIP (Proximal Interphalangeal Joint), and DIP (Distal Interphalangeal Joint) joints to bend, simulating human-like finger movements essential for delicate tasks [70, 71].

4.1.1.3 | Lower Limb Mechanical Structure Design. The lower limb structure typically contain six degrees of freedom (6-DOFs) including both the leg and foot, with the leg incorporating the hip and knee joints and the foot primarily involving the ankle joint. Some designs also include active or passive toes to facilitate jumping and to allow for a larger stride. The hip joint in a humanoid robot is engineered to bear substantial loads, handling the high bending moments generated when the leg swings down for landing. Typically, a series actuator mechanism enables the hip joint to achieve three degrees of freedom with a wide range of movement. For the knee joint with one degrees of freedom, actuators and reducers are usually positioned near the hip joint on the thigh to optimize mass distribution, concentrating mass around the waist. These joints are typically driven using belts, cranks, or linkages. The ankle joint often employs two parallel actuators to drive a spherical joint, allowing for two degrees of freedom while keeping the foot's inertia low to facilitate stable, agile movements [68].

4.1.2 | Sensor System

Sensors are integral to humanoid robots, providing them with the sensory input necessary for interaction with the environment. These systems help the robot perceive its surroundings, detect objects, and respond to external stimuli.

4.1.2.1 | Vision. Vision sensors, as the earliest and most developed sensor system, provide foundational data for environmental perception in humanoid robots. Most of the current advanced humanoid robots, such as Unitree G1 [69], Boston Dynamic Atlas [4], Tesla Optimus Gen2 [6], Xiaomi CyberOne [21] and Figure 02 [7], are equipped with light detection and ranging (LIDAR), RGB(-D) camera or a combination.

1. RGB(-D) Camera

Common RGB camera provide pictures for computer vision to extract dense information. Engineered Art Ameca uses two monocular cameras to form binocular vision [30]. Due to the absence of *Z*-axis data, a single monocular camera has limited depth perception. Consequently, RGB-D cameras are commonly used in humanoid robots, providing both visual and real-time depth information. However, the accuracy and measurement range of RGB-D cameras are limited compared to LIDAR, and they are easily affected in low-light or strong-light environments, resulting in inaccuracy.

2. LIDAR

LIDAR provides highly accurate depth data, usually down to the centimeter level, and is suitable for precise measurements over long range. In addition, LIDAR is unaffected by ambient light and can work stably in dark or bright light environments, making it very reliable in complex environments. However, LIDAR often generates a huge amount of data, which requires a high level of computational power to process.

4.1.2.2 | Haptics.

Haptics is another important sense for humanoid robots to understand complex scenes, for example, shaded or dimly lit environment. Additionally, the dexterous hand can be manipulated in a more refined way with haptic feedback. However, due to the difficulty of integrating and unstable signals, tactile sensors are still under research.

1. Flexible Arrayed Sensor

Arrayed sensor is the first tactile sensor to be developed and applied to humanoid robots. Common arrayed sensors are piezoelectric, capacitive and resistive. From micro-electro mechanics (MEMS) to flexible electronics, they become more integrable and flexible. Every sensor unit on the base is like a tactile neuron under skin, which can perceive contact information independently. Currently, most dexterous hands with tactile sensing capabilities are equipped with arrayed sensors, such as Optimus Gen2.

2. Vision-based Sensor

Vision-based tactile sensor (VBTs) is a recently developed high spatial resolution and multimodal haptic sensor. These sensors use a camera to catch the deformation of gel and compute the haptic information through computer vision approach. Although VBTs usually has higher resolution and denser tactile information, it suffers from large size that prevent it from integrated with dexterous hand. A bunch of miniaturization studies are proposed recent years, making VBTs the size of a finger. In the meanwhile, accuracy and topicality are reduced because of the miniaturization and more complex algorithm.

4.1.2.3 | Audition.

Microphone is essential for humanoid robots to communicate with environment and humans [4]. With the development of microphone arrays and sound localization technology, humanoid robots are allowed to accurately identify and respond to specific voice, even in noisy environment [21]. 2, 4 or 6 microphone arrays are currently the most commonly used in robots. As the number of microphones increases, localization and recognition accuracy improve, but at the same time the computational cost rises substantially. Trade-off should be considered between accuracy and running in real-time.

4.1.2.4 | Control.

In addition to the sensory sensors above, some other sensors are needed in the basic control of the robot to realize closed-loop control, and next two kinds of common sensors will be introduced.

1. Force/Torque Sensor

Force/Torque sensor for humanoid robots is usually made of strain gauges. According to Tesla AI Day [72], Single-axis sensors measure force or torque along one direction, suitable for simpler tasks like single DOF joint. Multi-axis sensors can measure and decouple forces and torques in multiple directions (typically three axes for force and three for torque), which can be fitted at the end of the actuator or the ankle joint, where the force is complex. It is worth mentioning that humanoid robots that emphasize human interaction as well as performance, force torque sensors are vital important.

2. Inertial Measurement Unit

Inertial Measurement Unit (IMU) integrates multiple inertial sensors such as accelerometers, gyroscopes, and sometimes magnetometers, for measuring linear acceleration, angular velocity, and angular change of an object. IMUs are capable of accurately sensing the robots' motion and attitude in three-dimensional space and are widely used for navigation and attitude control. On the downside, IMU data can generate cumulative errors over time, and in particular, small errors in accelerometers and gyroscopes can be superimposed over time, leading to a gradual loss of accuracy in the system. Therefore, IMUs usually need to be combined with vision sensors for error correction.

4.1.3 | Power Systems

The power system is crucial for humanoid robots to achieve precise movement and efficient operation, with its selection and configuration directly affecting the robot's agility, stability, accuracy, and overall operational efficiency. In the field of humanoid robotics, power systems must not only provide sufficient driving force to support complex dynamic movements but also maintain high control precision and responsiveness across diverse environments. This adaptability is essential to meet varied task requirements and operational settings. Currently, common drive methods include servo motors, hydraulic systems, and pneumatic systems, each with distinct advantages in terms of energy consumption, accuracy, response speed, and suitability for specific applications.

4.1.3.1 | Servo Motors.

Servo motors are highly integrated drive systems widely used in service-oriented humanoid robots [7, 69, 73]. They are characterized by high precision, fast response speed, and excellent control performance, typically serving as drives for smaller joints and hand components. Through closed-loop control with real-time feedback from encoders, servo motors enable highly accurate positioning, often down to the nanometer level. This makes them ideal for applications requiring high positional accuracy, such as finger and wrist joint movements. Their rapid start-stop response times also make them suitable for highly dynamic scenarios, such as walking on complex terrains or quick postural adjustments during interactive environments. However, servo motors exhibit high energy consumption during high-speed operations and frequent start-stop cycles, making them less suitable for long-duration, high-load applications. High-power servo motors tend to generate significant heat under heavy loads, necessitating additional cooling mechanisms. Furthermore, the precision and responsiveness of servo motors contribute to their higher manufacturing costs, aligning them with applications that can accommodate a higher budget.

4.1.3.2 | Hydraulic Systems.

Hydraulic systems, which use fluid pressure to transmit force, are primarily applied in scenarios requiring high load-bearing capacity and power [74–76]. In humanoid robots, hydraulic systems are typically used in joints that demand substantial output force, such as the lower limbs. These systems can deliver higher output forces than servo motors, making them ideal for applications with significant load-bearing requirements. By adjusting the pressure and flow of hydraulic fluid, hydraulic systems achieve precise force control, effectively simulating muscle tension. Unlike electric drives, hydraulic systems retain durability under high-load conditions and are relatively resistant to environmental influences. However, they involve complex networks of pipes and valves, require regular maintenance, and are vulnerable to issues such as hydraulic fluid leakage. The energy conversion efficiency of hydraulic systems is relatively low, with substantial energy consumption, especially under continuous high-power output. Additionally, the noise produced by hydraulic pumps and pipelines limits their applicability in noise-sensitive environments. Consequently, hydraulic systems are more suitable for high-load-bearing components in humanoid robots, such as the lower limbs. They perform exceptionally well in tasks demanding substantial power support, including heavy lifting and climbing. Hydraulic systems are also advantageous in extreme environments, such as high-temperature and high-humidity conditions.

4.1.3.3 | Pneumatic System. Pneumatic systems operate by compressing air within bladders to drive actuators, making them ideal for medium to low-load applications. Due to their flexibility and quick response characteristics, pneumatic systems are widely used in scenarios demanding lightweight and high flexibility [77–79]. Their structure is relatively simple, with fewer components, which makes maintenance easier and enables fast start-stop response times, making them suitable for applications with frequent activation cycles. Additionally, pneumatic systems provide excellent flexibility control, effectively simulating the elasticity and cushioning effect of muscles, which is beneficial for biomimetic and safe human-robot interactions. However, precise control of pressure is challenging in pneumatic systems, resulting in limited positional accuracy. Compared to hydraulic systems and servo motors, pneumatic systems offer lower output power, rendering them unsuitable for high-load applications. They are also sensitive to environmental factors such as temperature and humidity, making them more dependent on specific operating conditions. Consequently, pneumatic systems are best suited for lightweight and flexible-control components, such as fingers and arms where gentle contact is necessary. These systems perform well in clean indoor environments, making them ideal for educational or display-oriented lightweight humanoid robots.

4.2 | Software Architecture

As shown in Figure 5, the software architecture of the humanoid robot body is divided into two parts: the operating system and the communication scheme. The operating system includes the real-time operating system and the Robot Operating System (ROS). The operating system is responsible for event processing, decision analysis, and motion planning for the entire robot. It receives sensor information through communication protocols and controls the actuators to enable the robot's movement.

4.2.1 | Operating System

4.2.1.1 | Real-Time Operating System. Humanoid robots require precise, stable, and timely control of motion. They need to respond to high-frequency sensory input, process events in real time, and coordinate synchronized motion across multiple joints, placing a high demand on the system's real-time control capabilities. Real-time operating systems (RTOS) like QNX and VxWorks are generally suited to applications with stringent real-time requirements, where reliability and precision are paramount. On the other hand, Xenomai and PREEMPT_RT are open-source, affordable, and more accessible for early-stage technical validation, making them suitable choices for preliminary testing phases. These RTOS platforms provide essential support for handling complex tasks and maintaining control stability within a humanoid robot's mechanical system.

4.2.1.2 | Robot Operating System. ROS (Robot Operating System) is an open-source framework that provides a collection

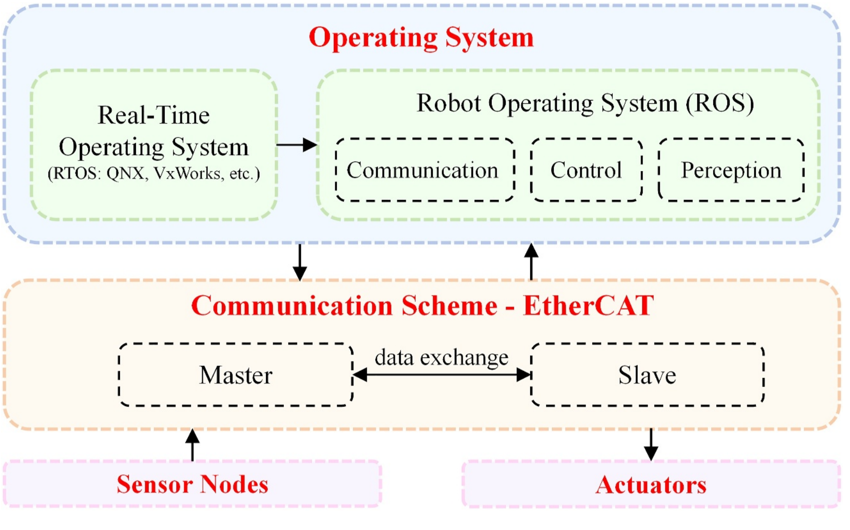

FIGURE 5 | Software architecture workflow diagram of a humanoid robot.

of tools, libraries, and conventions for building robotic applications. It offers a middleware layer for communication between robot components and facilitates the development of complex, modular robotic systems. ROS includes tools for hardware abstraction, device drivers, message-passing between nodes, and algorithm libraries for tasks like perception, motion planning, and control. It supports a wide range of robots and hardware platforms, making it a popular choice for both research and industrial applications. Recently, ROS has evolved into ROS 2, which improves performance, security, and scalability for modern robotics.

4.2.2 | Communication Scheme

EtherCAT is a high-performance industrial Ethernet protocol widely used in humanoid robots for its real-time capabilities and efficiency. In these systems, an EtherCAT master-slave scheme enables real-time data transmission between the controller (master) and multiple servo drives and sensors (slaves). The master sends a continuous data frame through the network, with each slave extracting relevant data and adding its feedback. This low-latency scheme achieves microsecond-level refresh rates, ideal for complex motion control. EtherCAT also supports Distributed Clocks for synchronized, coordinated movements and offers strong scalability, allowing hundreds of connected devices, which enhances response speed and stability in humanoid robot control systems.

5 | Key Technologies for Humanoid Robots

Humanoid robots, as complex systems designed to emulate human capabilities, rely on a variety of key technologies to enable them to operate autonomously and interact effectively with their environment. These technologies span multiple domains, from locomotion control to autonomous navigation, environmental perception, and intelligent manipulation. Each of these areas presents its own set of challenges and innovations, which contribute to the robot's overall functionality and versatility. In this section, we explore the fundamental technologies that empower humanoid robots to perceive, navigate, move and manipulate their surroundings with human-like efficiency.

5.1 | Environmental Perception

In addition to movement, the ability to perceive and understand the surrounding environment is vital for humanoid robots. Perception algorithms are designed to model and understand a robot's own state and its surrounding environment using signals from various onboard sensors. These algorithms can be categorized into internal state estimation [80] and external environment perception [81, 82]. The former includes pose estimation and robust localization, while the latter involves object recognition and map building. As illustrated in Figure 6, the vision-based humanoid robot perception system takes images from (multi-view) cameras as input, which undergoes preprocessing through a data association module. Subsequently, the system proceeds to state estimation, localization, and occupancy map construction. The data association module can be divided into short-term data association and long-term data association, with the latter referring to the problem of place recognition. Short-term data association is primarily employed to establish odometry constraints, facilitating the robot's short-term state estimation, while place recognition serves to enhance further localization algorithms.

In recent years, with the rapid advancements in deep learning and computer vision, lightweight and low-power visual sensors have emerged as the dominant technology for humanoid robots to achieve intelligent perception and autonomous navigation across their entire lifecycle. As a result, model-based robotic perception algorithms are progressively evolving toward learning-based approaches. This section offers a concise review of the key algorithms in state estimation, robust localization, and 3D

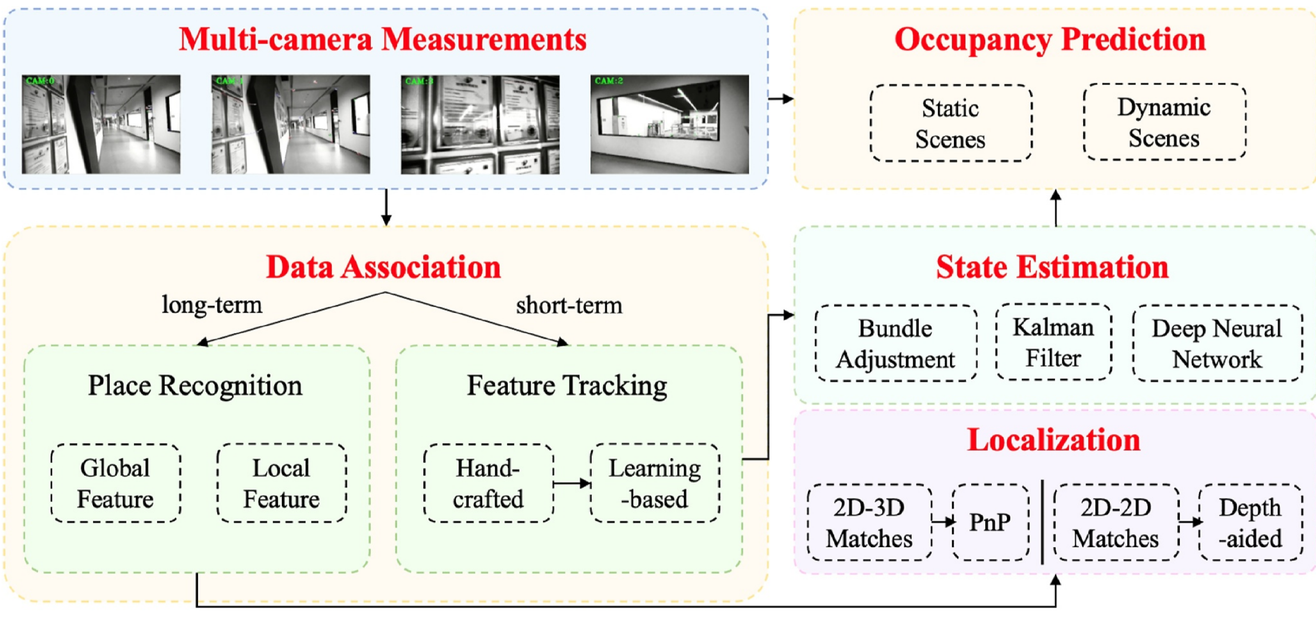

FIGURE 6 | The environment perception system of humanoid robot.

occupancy prediction, highlighting mainstream methods, development trends, and the associated research challenges.

5.1.1 | State Estimation

High-precision, low-latency pose estimation is foundational for achieving functions such as autonomous navigation and human-robot interaction [83, 84]. It is also essential for robust control on platforms with high degrees of freedom and strong nonlinear dynamics. To attain centimeter-level localization accuracy and sub-millisecond latency (i.e., a pose output frequency above 1 kHz), pose estimation modules typically integrate observations from various complementary sensors, including multi-camera vision [85, 86], inertial measurement units (IMUs) [85, 87, 88], and legged odometry [89]. These measurements are combined to form a maximum a posteriori (MAP) estimation problem regarding the robot's pose [90], which is then solved using Gaussian inference, extended Kalman filtering, pose graph optimization, and deep learning methods [80]. The solution process of the MAP estimation problem is typically referred to as the "back-end" of the pose estimation algorithm, while the modeling of sensor observations and data association is called the "front-end."

5.1.1.1 | Visual Front-End. Pose estimation front-ends based on vision can be divided into direct and indirect methods. Direct methods perform data association by matching images with a dense map in 2D-3D correspondences, optimizing for minimum photometric error at each pixel [91]. Indirect methods achieve data association by matching feature descriptors between image frames in 2D-2D correspondences, aiming to minimize the reprojection error of sparse map points [87, 88, 92]. Although indirect methods add steps for feature extraction and feature matching, they generally offer higher pose estimation accuracy and are more robust to changes in lighting conditions, making them the mainstream approach for visual pose estimation [90, 93].

5.1.1.2 | Back-End. The MAP estimation problem in the back-end of pose estimation algorithms can be approached using either filter-based methods, such as the extended Kalman filter (EKF) and its variants [88, 94, 95], or optimization-based methods [87, 91, 92], such as pose graph optimization [96] and bundle adjustment [97–99]. Depending on the MAP problem formulation, approaches that integrate observations from multiple sensors into a unified MAP estimation problem are termed "tightly-coupled" methods [87, 88, 92]. In contrast, "loosely-coupled" methods separately solve the MAP estimation problems for each sensor's observations, then fuse the results [100, 101]. Tightly-coupled methods better handle system noise, thus achieving higher pose estimation accuracy. However, since they jointly solve the robot's pose using all sensor signals, the failure of any single sensor can lead to a system breakdown. In comparison, loosely-coupled methods, by independently estimating pose for each sensor, may slightly sacrifice accuracy but are more robust to individual sensor failures [90, 93].

Lifelong navigation is a key objective for achieving embodied intelligence, which places higher demands and new challenges on pose estimation. At the same time, rapid advancements in artificial intelligence are invigorating pose estimation research. Traditional issues in robust visual feature extraction and feature matching under varying lighting conditions are now being addressed through deep learning methods, such as feature extraction using SuperPoint [102] and feature matching with SuperGlue [103] and LightGlue [104], which serve as front-ends in robot pose estimation algorithms [105–108]. Additionally, integrating deep learning into back-end information fusion holds the potential for continuous online optimization of robot pose estimation algorithms, enabling lifelong training and improvement [109].

5.1.2 | Robust Localization

The visual localization problem in robotics refers to estimating the robot's position and orientation within a pre-existing map by matching the robot's visual observations with the map information. The prior map is typically constructed through techniques such as Simultaneous Localization and Mapping (SLAM) or Bundle Adjustment (BA). Visual localization algorithms are commonly used to solve issues such as the robot's kidnapping problem, long-term data association in SLAM, and to provide pose priors and fault recovery for the robot's pose estimation algorithm. These algorithms are one of the key technologies for achieving autonomous navigation throughout the robot's entire lifecycle. Common visual localization algorithms use historical observation frames with known poses (geo-tagged frames) and a large number of visual landmarks to represent the map. The localization process is typically divided into two steps:

1. Place Recognition: Match the current observation frame with historical frames to find the most similar reference frame, achieving place recognition.
2. Pose Estimation: Use methods like Perspective-n-Point (PnP) or multi-view geometry to solve for the relative pose between the current frame and the reference frame, thereby estimating the pose of the current frame.

Due to the complexity and variability of real-world environments, the main challenges for visual localization algorithms are typically encountered during the place recognition stage:

1. Visual Appearance Variations: Under the influence of environmental factors such as weather, lighting, and seasons, images captured from the same location and angle can differ significantly from historical reference frames, leading to mismatches.
2. Viewpoint Changes: Since cameras used in visual localization are typically not omnidirectional, images captured from different viewpoints of the same location can exhibit large differences, causing place recognition failure.
3. Perception Aliasing: In environments with similar scenes or repetitive textures, erroneous matches are likely to occur.
4. Scalability: Real-world environments are vast and boundaryless, while computational resources are typically

limited. Achieving efficient and robust localization in large-scale environments remains a challenging task.

To achieve robust place recognition, existing algorithms represent images as vectors or matrices, using cosine distances between image descriptors to measure image similarity. Early global feature representation methods, such as Gist [110], VLAD [111], BoW [112], ASMK [113], and HOG [114], used manually designed feature extraction methods. With the rapid development of deep learning, deep learning-based global feature extraction algorithms have gradually replaced traditional methods, offering more robust place recognition performance. Examples include NetVLAD [115], TransVPR [116], MixVPR [117], and AnyLoc [118].

5.1.3 | 3D Occupancy Prediction

3D scene understanding is a challenging yet highly significant task, with the primary problem being how to effectively interpret 3D scenes. Since Tesla introduced Occupancy Networks at the 2022 CVPR Workshop on Autonomous Driving, occupancy prediction has gained widespread attention. The task of semantic occupancy prediction can be traced back to semantic scene completion, which primarily focuses on static scenes. In contrast, semantic occupancy prediction is not limited to static scenes and has potential to predict to motion of objects within the scene. 3D occupancy prediction outputs voxel-wise status and semantic labels for more comprehensive understandings of 3D scenes, which is crucial for robots to make efficient decisions and path planning in constantly changing environments.

The pioneering MonoScene [119] utilized a monocular camera as input for scene completion, which has driven the development of a series of vision-based occupancy perception algorithms. VoxFormer [120], on the other hand, utilized a two-stage Transformer architecture to achieve temporal single-view voxel prediction. VoxFormer outperforms MonoScene with a large gap. However, it relys on an offline depth estimation process, which means it is not yet suitable for practical deployment.

TPVFormer [121] proposed a tri-perspective view (TPV) representation which accompanies BEV with two additional perpendicular planes and expanded occupancy prediction from multi-view images. It utilized a cross-attention mechanism to extract TPV (Tri-View Point) features from image features and employed a cross-view hybrid attention mechanism to enable queries to exchange their information across different views. Under the task of 3D semantic occupancy prediction, the model was trained using sparse semantic labels (LiDAR points). However, as no benchmark was provided before this work, TPVFormer only performed a qualitative analysis. Besides, the qualitative results showed noticeable holes, indicating that it is not yet suitable for practical deployment.

With the continuous advancement of 3D occupancy prediction, researchers have realized the im-portance of constructing benchmarks. X. Wang et al. [122] extended the large-scale nuScenes dataset with dense semantic occupancy annotations, which is the first surrounding semantic occupancy perception benchmark. Unlike SemanticKITTI [123], which has a relatively small scale and limited diversity, nuScenes offers a larger, more diverse dataset and evaluates surrounding perception, making it more suitable for developing and testing occupancy perception algorithms for safe driving. During the process, an Augmenting And Purifying (AAP) pipeline was introduced to efficiently annotate and densify the occupancy labels. However, it is worth noting that the manual annotation process requires approximately 4000 human hours, making it quite costly.

Another similar work SurroundOcc [124], also builds a benchmark for semantic occupancy prediction based on the nuScenes dataset. Unlike OpenOccupancy, SurroundOcc generated ground truth without manual annotations. It used Poisson reconstruction to densify the point clouds aggregated from multiple frames and employed a nearest-neighbor algorithm to densify the point cloud labels, resulting in a dense occupancy ground truth. However, there are still noticeable holes in the results.

Occ3D [125] developed a label generation pipeline that produces dense, visibility-aware labels for any given scene. It proposed a pipeline called image-guided voxel refinement to improve the shape at object boundaries and generated two 3D occupancy prediction datasets, Occ3D-nuScenes and Occ3D-Waymo. After the establishment of benchmarks, the study of 3D occupancy prediction has entered a new phase of development. However, the scenarios of real world are much more complicated than those in the occupancy benchmarks based on autonomous driving datasets. In practical applications, we inevitably face the challenges of data collection and annotation, which brings the issue of expensive 3D annotation costs. In contrast to 3D annotations, 2D annotations are more cost-effective, efficient, and scalable. As a result, some researchers focus on training models using 2D supervision. RenderOcc [126] extracts a NeRF representation from multi-view images, obtains 2D renderings using volume rendering techniques, and then uses 2D semantic labels and depthmaps for supervision. OccNeRF [127] proposes a self-supervised multi-camera occupancy prediction method that reconstructs 2D images from 3D representations using volume rendering, specifically guiding the model to recover the 3D structure of the scene through photometric losses between adjacent frame images. Similar self-supervised approaches can be seen in works [128, 129].

Inference speed is another essential factor that cannot be ignored. Some researchers have attempted to explore ways to accelerate the inference process by optimizing the architecture of network. GaussianOcc [130] replaces volume rendering with Gaussian splatting, achieving more accurate and efficient rendering from occupancy voxels to depth maps. FastOcc [131] compresses 3D voxel features into 2D BEV (Bird's Eye View) features, compensates for missing z-axis information by interpolating voxel features derived from images, and replaced the time-consuming 3D convolution network with a novel residual-like architecture. This approach achieves an inference delay of 221 ms and demonstrates strong performance on the Occ3D-nuScenes dataset.

Finally, understanding the trends in the surrounding environment's changes is a crucial aspect for robots to safely and reliably perform downstream tasks. The aforementioned works are mostly limited to representing the current 3D space and do not consider the future state of surrounding objects along the time axis. Simply stacking the output results can lead to ghosting effects caused by dynamic obstacles. Cam4DOcc [132] builds a 4D occupancy prediction benchmark based on image inputs, using the motion state of voxels represented as a flow on the nuScenes, nuScenes-Occupancy, and Lyft-Level5 datasets, pushing the development of occupancy prediction tasks. Similar works include [133, 134].

It is worth noting that most research on occupancy prediction has been conducted using autonomous driving datasets. However, the motion environment of humanoid robots is much more complex than that of autonomous vehicles. The height of humanoid robots can vary significantly during movement, which requires higher demands on the real-time performance, accuracy, and robustness of perception models. Furthermore, the task goals of humanoid robots are usually more complex, involving a wide range of uncertain objects. The robot needs not only to recognize these objects but also to interact with and manipulate them. Existing research mainly focuses on the semantic labels of occupied areas and does not explicitly distinguish the occupancy boundaries between different instances, which to some extent affects the accuracy and efficiency of downstream tasks.

5.2 | Autonomous Navigation

Navigation of mobile robots is one of the core problems in the field of robotics, which has been extensively researched and applied. Unlike the traditional navigation frameworks for wheeled robots, mainstream navigation planning frameworks for humanoid robots generally consist of three parts: global planning, local planning, and footstep planning, which together enable multi-level motion planning.

Global planning is responsible for solving feasible paths based on the environmental descriptions provided by different map representations, providing directional guidance for local planning. Local planning handles obstacle avoidance and trajectory tracking in the local environment, while footstep planning ultimately determines suitable footholds to guide the robot in following the commanded velocity.

This section introduces the current research status of global planning, local planning, and footstep planning, respectively, highlighting the research progress and development trends in the field of mobile robot navigation planning, as shown in Figure 7.

5.2.1 | Global Planning

Traditional global path planning algorithms primarily focus on position targets in static environments, relying on pre-built maps and using graph search (e.g., A* [135]) or sampling-based methods (e.g., RRT* [136]) to generate collision-free paths. While effective in such settings, these methods are less applicable to unknown environments due to the lack of complete mapping [137–140]. Researchers have proposed using SLAM for real-time mapping with re-planning, but SLAM's complexity and resource demands limit flexible deployment.

Actually, in new environments, full coverage mapping is unnecessary; instead, focus on key areas near the target and along critical paths is more efficient [141, 142]. Effective planning in unknown environments involves two main aspects: environmental representation and decision-making frameworks. Environmental representation, through dense maps, topological maps, or scene graphs, provides foundational information, while decision-making uses this to select suitable path planning strategies. These two components work together to achieve navigation objectives.

While human navigation, unlike traditional planning, is based on partial perceptions, sparse memory, and target awareness, leading to more adaptive behavior. Inspired by this, sparse topological maps have been used to replace dense maps, reducing computational costs while retaining essential information [143–145].

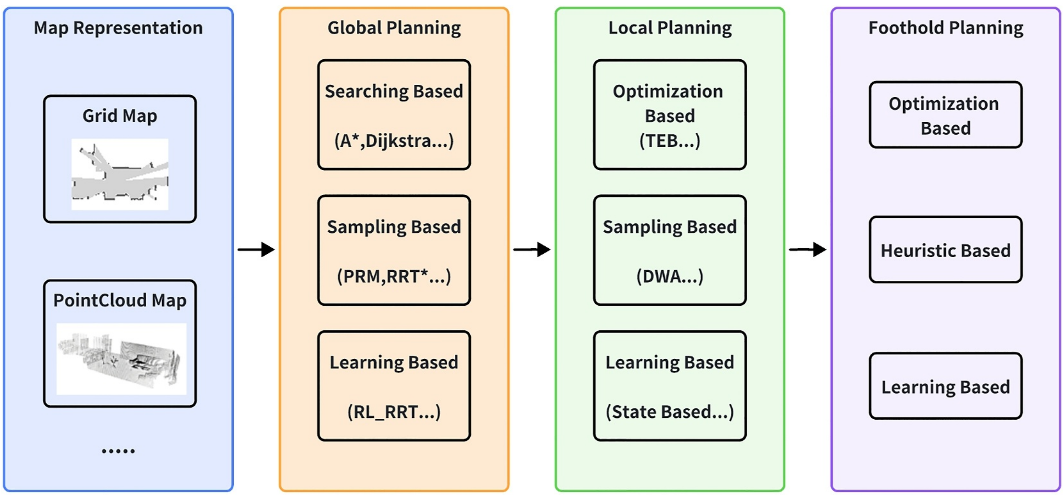

FIGURE 7 | The autonomous navigation framework of humanoid robots.

Those methods also improve generalization and adaptability using fewer nodes and edges [146–148]. Recent research has also incorporated deep learning to better understand topological maps, enhancing their utility in complex environments [149–151].

Recently, the need for multimodal navigation has driven robots to use visual and linguistic cues. Compared to traditional navigation, those navigation tasks involve interpreting multiple modality information, which poses additional challenges. Early methods used dense maps but were vulnerable to sensor noise [152–154]. Deep learning has improved robustness and performance [155, 156], and topological maps have shown promise due to scalability [157, 158], though integration challenges remain. Hierarchical planning using sub-goals has improved adaptability but struggles with generalization [159]. A combined approach of traditional planning and deep learning could leverage the strengths of both, enhancing navigation in unknown environments. Overall, multimodal navigation is in its early stages, with significant research potential, especially with the integration of foundation models.

5.2.2 | Local Planning

Rule-based local planning algorithms use modular design to handle dynamic obstacle avoidance. Traditionally, these methods employ sensors and dynamic analysis to assess the environment and generate safe paths. The process involves decomposing dynamic obstacle avoidance into static subproblems through modules like obstacle detection and tracking, trajectory prediction, and path planning. Detection and tracking modules use handcrafted features or deep learning to identify dynamic obstacles, while prediction modules forecast future positions based on obstacles' historical trajectories and distributions. Models such as RVO (Reciprocal Velocity Obstacles) [160] and ORCA [161] take inter-obstacle interactions into account to optimize avoidance paths. Rule-based planning is computationally efficient and interpretable but limited in adaptability for complex environments and sensitive to the accuracy of detection and prediction modules, resulting in instability in highly dynamic settings [162].

Supervised learning-based local planning, on the other hand, learns from expert demonstration data through an end-to-end network architecture, achieving good performance in dynamic obstacle detection and path planning. NVIDIA [163] applied supervised learning in autonomous driving, using cameras to directly output control commands, which avoids complex multi-module designs and improves system efficiency. Other studies have enhanced environmental understanding by adding modules for depth information and semantic inference, increasing network transferability and interpretability [164, 165]. This method depends heavily on the distribution of the dataset, limiting its adaptability in new scenarios, where data diversity and annotation quality greatly affect performance.

Reinforcement learning-based local planning learns strategies through trial and error in the environment, commonly using model-free and model-based reinforcement learning methods. Model-free reinforcement learning, like the value function network proposed by Y. F. Chen et al. optimizes robot motion by assessing the safety and comfort of obstacle interactions [166]. And some detection-free methods directly utilize sensor data to achieve mapless navigation [167]. Model-based reinforcement learning accelerates strategy training by establishing a world transition model, especially useful for tasks that demand high sampling efficiency [168]. However, reinforcement learning methods tend to lack interpretability and safety guarantee, prompting recent research to improve safety through methods like Control Barrier Functions (CBF) [169], which support real-world deployment.

Each of these three approaches has distinct advantages: rule-based methods are highly interpretable and suitable for relatively simple dynamic environments; supervised learning can quickly learn expert strategies but relies on high-quality datasets; and reinforcement learning offers flexibility and adaptability, though safety and interpretability require further enhancement.

5.2.3 | Foothold Planning

The selection of footholds is pivotal for enhancing robot mobility across diverse terrains. Overall, this task is a precise control challenge aimed at bolstering the robustness of robots on various types of terrain. Heuristic methods are highly effective on flat surfaces [170]. For more complex terrains, such as stairs and stepping stones, some other strategies use heuristic formulas to select the nearest terrain segment [171]. However, these strategies often neglect the critical interaction between the robot's four legs and its Center of Mass (COM). To address this, optimization-based approaches have grown more popular than heuristic and sampling methods. This trend is due to their ability to identify optimal solutions based on specific criteria, rather than focusing solely on feasibility [172]. For instance, some researchers describe a foothold generation method involving batch searches on grid maps [173, 174]. Moreover, researchers enhance foothold optimization by treating footholds as continuous variables within the trajectory optimization process, leveraging gradient-based methods [175].

With the recent development of large-scale parallel reinforcement learning environments, researchers have integrated the precise trajectories generated from optimization-based methods with robust control strategies from learning-based methods [176]. This integration has enabled quadrupedal locomotion on outdoor terrains, significantly improving the off-road performance of legged robots.

5.3 | Locomotion Control

Locomotion control is a cornerstone of humanoid robotics, as it enables robots to navigate complex environments using their limbs. Unlike wheeled robots, humanoid robots face the additional challenge of maintaining balance while walking, running, or navigating uneven terrain. Achieving stable and efficient locomotion requires integrating multiple control strategies, considering the robot's dynamic balance, kinematics, and interaction with the environment. The field of locomotion control for humanoid robots can be divided into model-based methods and data-driven approaches, each offering distinct advantages and addressing different aspects of the control problem.

Model-based methods, grounded in classical control theory, rely on a detailed mathematical representation of the robot's dynamics to compute motion commands. These methods are particularly advantageous in environments where precise control is needed, but they require a high level of accuracy in the system's model. Alternatively, learning-based approaches seek to overcome the limitations of model-based control by enabling robots to adapt to varying conditions through experience.

5.3.1 | Model-Based Methods

The main part of model-based locomotion control method is balance controller, as illustrated in Figure 8. The Balance Controller receives the state of the robot and the commands of users to control the balance of robot. And the Motor Controller transfer the balance controller's outputs actions into position or torque to control the robot.

Early research results mainly started from model-based approaches, studying the stability criteria and planning control methods grounded in kinematic and dynamic modeling. Stability criteria, such as the zero moment point (ZMP), center of pressure (CoP), and capture point (Capture Point), have been widely adopted in humanoid robot control. Kajita et al. [177] developed a widely recognized ZMP-based control framework that has become foundational in achieving balance for bipedal robots under dynamic conditions. Furthermore, Pratt et al. [178] introduced the capture point concept to predict a robot's balance recovery ability, a crucial advancement in real-time stability control. Dariush et al. [179] presented an online task-space control retargeting formulation that satisfies balance and kinematic constraints with a ZMP controller to maintain balance under changing task requirements.

Center of mass (CoM) motion planning methods and preview control techniques have been widely studied as stability criteria-based approaches for humanoid robots. Yamane et al. [180] proposed a balance controller based on a simplified robot model combined with an optimization process to compute joint torques, enabling a force-controlled robot to maintain balance while tracking a given trajectory without contact switching.

Another effective method for achieving stable bipedal locomotion is the hybrid zero dynamics (HZD) approach. By leveraging zero dynamic characteristics, this approach maintains closed-loop stability along the zero dynamics manifold, facilitating robust and natural gait patterns. As initially proposed by Westervelt et al. [181], and later expanded by Ames et al. [182] and Grizzle et al. [183], HZD has proven effective for various walking conditions, providing closed-loop control for more complex bipedal walking models.

In addition, decoupling walking control into separate components—such as attitude control, speed control, and foothold control—and managing these via virtual models is a common practice in humanoid control. Model predictive control (MPC) has emerged as a widely adopted approach in this area due to its robustness and adaptability. Romualdi et al. [184] proposed an online centroidal MPC to adjust foot placement, achieving improved robustness and precise trajectory tracking. Similarly, Dai et al. [185] developed a whole-body MPC framework that enhances dynamic stability over rough terrains by enabling simultaneous foot placement and posture control. Di Carlo et al. [186] further demonstrated MPC's adaptability in challenging environments, enabling precise foot placement and balance maintenance even on uneven surfaces.

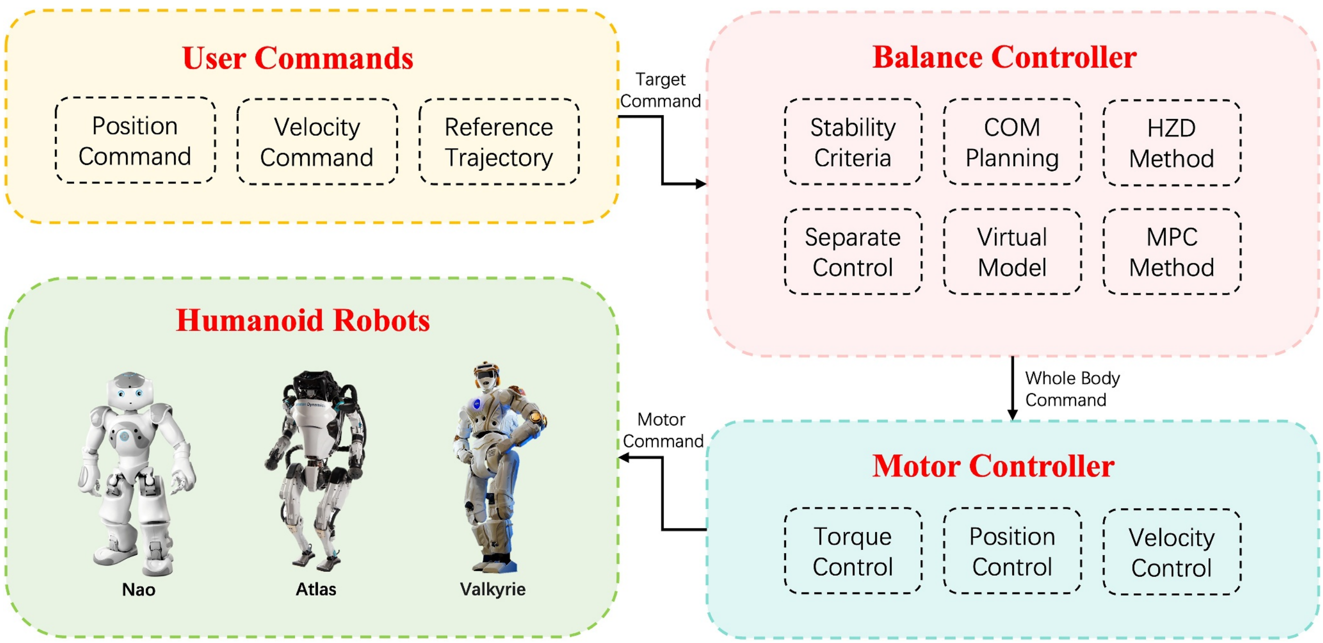

FIGURE 8 | The framework of model-based locomotion method.

While each of these approaches offers unique advantages, such as high interpretability, they also have limitations. Primarily, these methods depend heavily on accurate modeling and manual parameter tuning, which complicates optimization in multi-rigid body planning and control.

5.3.2 | Learning-Based Methods

Unlike traditional model-based methods that rely on explicit mathematical models of the robot's dynamics, learning-based approaches leverage data-driven techniques to enable robots to learn locomotion strategies from experience. Learning-based methods are particularly advantageous in environments where traditional models may struggle to capture the complexities of interaction with unstructured or unpredictable terrains. The RL policy receives the state of the robot from observation and proprioception of the robot's sensors and the reference trajectory to control the action of the robot. The reference trajectories are given by motion retargeting, navigation or motion planning and generation. Domain Randomization added in training process helps building robust RL policy for sim-to-real deployment. RL policy outputs actions and a PD controller transfer them into the torque to control the robot.

Motion retargeting is a technique that facilitates the transfer of motion data from a source character to a target character. In the realm of animation, motion retargeting is typically achieved through optimization-based methods [187–192] and learning-based methods [192–197]. The learning-based approaches tend to yield more efficient results, particularly in enabling motion transfer across various skeletal structures [195]. In the context of humanoid motion retargeting, Holden et al. [196] employ shared latent variable models to adapt motions between different humanoid figures. Jang et al. [197] introduce a self-supervised shared latent embedding technique that generates humanoid motions from human motion capture (MoCap) data or video sources. Delhaisse et al. [198] utilize a wearable device to capture human movement, implementing an inverse kinematics (IK) control scheme combined with a quadratic programming (QP) optimization solver for motion retargeting. Additionally, Choi et al. [199] reconstruct human motion while adhering to the physical constraints dictated by humanoid dynamics and provide a precise morphing function for accommodating various human body dimensions.

Before implementing reinforcement learning (RL)-based controllers in real-world humanoid robots, RL techniques have often been applied in the realm of physics-based animation control [201–211]. However, humanoid avatars typically possess a high degree of freedom, with few constraints on joint positions, torques, and sometimes additional auxiliary forces [212]. In contrast, realistic humanoids are governed by complex dynamic models, making it challenging to obtain privileged states, such as base velocity and height, from onboard sensors [213]. This disparity hinders the direct transfer of RL models designed for animated characters to physical humanoid robots. Z. Li et al. [214] introduced an end-to-end RL approach that incorporates task randomization to develop a robust dynamic locomotion controller for bipedal robots. Radosavovic et al. [215] created a causal Transformer that is, trained using autoregressive prediction of future actions based on a history of observations and actions to facilitate locomotion in full-sized humanoids. Additionally, Cheng et al. [213] employed stair-like terrain randomization to design an RL controller specifically for humanoids navigating stair-like environments.

After RL controllers being widely used in humanoid control, researchers have begun to focus on the application of reinforcement learning in action imitation. T. He et al. [216] utilized a privileged policy to select an executable motion dataset, which aids in training a robust RL policy suitable for sim-to-real deployment. Fu et al. [217] developed a task-agnostic low-level policy designed to track retargeted humanoid poses. These approaches have collectively advanced the capabilities of whole-body teleoperation for humanoids. By leveraging privileged information and effective policy training, they have enhanced the performance and versatility of humanoid robots in various operational contexts.

5.4 | Intelligent Manipulation

Intelligent manipulation is one of the key capabilities that distinguishes humanoid robots from simpler robotic systems. It involves the ability to interact with, manipulate, and modify objects within the environment in a human-like manner. Achieving intelligent manipulation requires an intricate integration of task planning, perception, and motor control. Unlike traditional robotic manipulation, which may rely on predefined actions, intelligent manipulation demands adaptability, flexibility, and the ability to respond dynamically to environmental changes.

The process of intelligent manipulation can be broken down into two main areas: task planning and skill learning. Task planning enables the robot to determine the sequence of actions needed to achieve specific objectives, while skill learning allows the robot to improve its performance through experience and practice. Together, these components enable humanoid robots to autonomously perform complex tasks, from simple object grasping to sophisticated assembly operations.

This section introduces the core technologies behind intelligent manipulation, exploring the strategies used for task planning, including symbolic reasoning, large language models, and closed-loop task planning. It also delves into skill learning techniques that allow robots to learn both single and multi-task manipulation skills, as well as long-horizon tasks requiring extended sequences of actions. The overview list of intelligent manipulation method is illustrated in Figure 9.

5.4.1 | Task Planning for Robotic Manipulation

Task planning for robotic manipulation is the process by which robots determine the sequence of actions needed to achieve a desired outcome. This process is essential for enabling robots to perform high-level tasks that require multi-step reasoning and execution. Effective task planning must not only account for the robot's physical capabilities but also the constraints of the environment and the properties of the objects being manipulated.

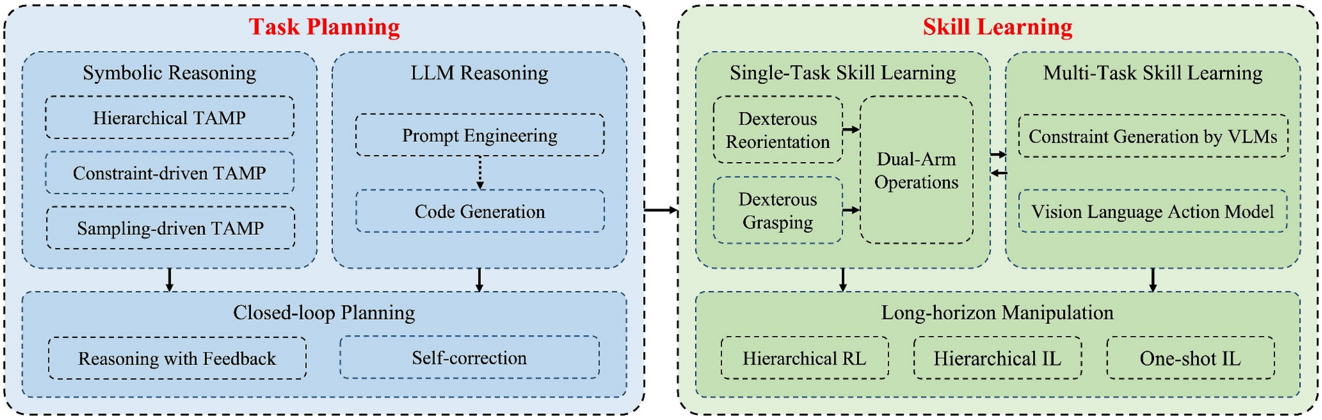

FIGURE 9 | The overview list of intelligent manipulation method.

TABLE 3 | Comparison of different task plan generation methods.

Method name	Year	Method type	Function
HTN with geometric feasibility [218]	2018	Hierarchical TAMP	Dual-arm manipulation with complex geometric constraints
IDTMP [219]	2016	Constraint-driven TAMP	Incremental action feasibility checking in task planning
FFRob [220]	2018	Sampling-driven TAMP	Efficient task and motion planning with probabilistic completeness
Co-optimized TAMP [221]	2016	Optimization-based TAMP	Multi-layer optimization to minimize task costs in robotic operations
Say can [222]	2023	Large language model (LLM)	Long-horizon task planning using real-world grounded semantic knowledge
ProgPrompt [223]	2023	Programmatic LLM prompt	Generates feasible task plans by combining context, robot capabilities, and task specs
HuggingGPT [224]	2023	LLM task controller	Cross-modal task planning and coordination of AI models
ReAct [225]	2022	LLM reasoning + action generation	Enhanced task tracking and plan updating through interaction with external resources
Inner monologue [226]	2022	Closed-loop LLM	Task planning with feedback for improved decision-making in various scenarios
Reflexion [227]	2023	Closed-loop LLM with memory	Self-reflection using linguistic feedback for goal-driven task planning
REFLECT [228]	2023	Closed-loop LLM with multi-sensory feedback	Failure explanation and correction in task completion

In the context of humanoid robots, task planning becomes increasingly complex due to the need to reason about physical interactions with the environment and coordinate multiple degrees of freedom. Recent advancements in task planning for robotic manipulation include symbolic reasoning, the use of large language models, and closed-loop planning systems, each offering unique advantages in terms of flexibility, adaptability, and real-time performance. The comparison of different task plan generation methods is shown in Table 3.

5.4.1.1 | Symbolic Reasoning-Based Task Planning. A prevailing trend in symbolic task planning is the use of hierarchical planning to address complex problems or to integrate it with geometric motion planning, commonly known as Task and Motion Planning (TAMP) [229]. Within hierarchical TAMP approaches, Suárez-Hernández et al. [218] introduced a framework that integrates hierarchical task networks (HTN) with geometric feasibility checks for complex geometric constraints in dual-arm operations, iteratively decomposing compound tasks until a feasible solution is found. In constraint-driven TAMP, Dantam et al. [219] proposed IDTMP, a method that leverages SAT solvers to dynamically execute action feasibility checks within task planning, incrementally incorporating geometric constraint information to enhance scalability and planning efficiency. In sampling-driven TAMP methods, Garrett et al. [220] developed the FFRob algorithm, which utilizes Extended Action Specification (EAS) to combine geometric and kinematic constraints, thereby increasing planning efficiency and probabilistic completeness. They also introduced PDDLStream as an open-source TAMP framework, facilitating sampling generation in PDDL language. Lastly, in optimization-based

TAMP, C. Zhang and Shah [221] proposed a hierarchical optimization approach that uses multi-layer optimization and heuristic algorithms to solve large-scale robotic operations by explicitly modeling task transition costs and minimizing planning costs.

5.4.1.2 | Large Language Model-Enabled Task Planning. The advent of large language models (LLMs) has led to significant advancements in task planning, with LLMs demonstrating notable intelligence in reasoning, tool use, planning, and instruction following. Ahn et al. [222] utilized pre-trained skills to ground LLMs in real-world contexts, enabling them to incorporate semantic knowledge and environmental constraints to guide robots through complex, long-horizon tasks. Singh et al. [223] proposed a programmatic LLM prompting structure that combines contextual environment, robot capabilities, and task specifications, enabling LLMs to generate feasible action sequences in task planning. Shen et al. [224] utilized LLMs as task controllers, using language interfaces to handle task planning, model selection, and result synthesis, integrating diverse AI models from the Hugging Face platform to collaboratively solve cross-modal and cross-domain tasks. A key component of their HuggingGPT approach involves an initial task decomposition, where the LLM explicitly breaks down a given task into subtasks and defines dependencies among them. Additionally, Yao et al. [225] combined LLM reasoning traces with task-specific action generation, enabling the reasoning component to help the model track and update task plans effectively while interacting with external resources, such as knowledge bases, for supplementary information.

5.4.1.3 | Closed-Loop Task Planning and Self-Correction. Recent studies have further demonstrated the potential of LLMs in closed-loop task planning and self-reflection/self-correction based on environmental feedback. W. Huang et al. [226] proposed an intrinsic reasoning mechanism based on feedback, where LLMs use feedback signals—such as success detection, scene description, and human interaction—to enhance their reasoning and planning abilities, significantly improving high-level task completion across varied scenarios. Shinn et al. [227] introduced the Reflexion framework, where LLMs engage in linguistic reflection on task feedback signals and store these reflections in an episodic memory buffer, thereby improving decision-making in subsequent trials and reinforcing LLMs' performance in goal-driven task planning. Similarly, Z. Liu et al. [228] proposed the REFLECT framework, which generates hierarchical summaries from multi-sensory observations to query LLMs for failure reasoning, utilizing a language-based planner to correct failures and enhance task completion rates.

5.4.2 | Skill Learning for Robotic Manipulation

Skill learning for robotic manipulation involves training robots to acquire and refine the motor skills necessary for performing tasks. Rather than relying on pre-programmed sequences of actions, skill learning allows robots to adapt and improve through experience. This capability is essential for enabling robots to perform a wide variety of tasks in unstructured environments, where traditional programming may be insufficient.

Skill learning is typically divided into two primary categories: single-task skill learning and multi-task skill learning. While single-task learning focuses on mastering specific actions, multi-task learning allows robots to generalize across a range of tasks, improving their ability to handle novel situations.

5.4.2.1 | Single-Task Skill Learning. The single skill learning of humanoid robots is a widely studied topic. Dexterous hand manipulation and dual-arm coordinated operations are crucial components of this area. Traditionally, these skills were programmed using rule-based systems and expert knowledge, which required extensive manual tuning and were often inflexible to environmental changes. However, recent advances have shifted toward learning-based methods, particularly deep learning and reinforcement learning, to improve the adaptability and robustness of these skills. The comparison of different single skill learning approaches is shown in Table 4.

For in-hand manipulation with dexterous hands, OpenAI et al. [230, 231] learned a controller to reorient a block or a Rubik's cube, but it requires accurate acquisition of the object's state information. Qi et al. [232, 233] respectively designed learning models based on motor fast adaptation and vision-based tactile sensors. These models infer features of the expert policy for policy learning. T. Chen, Xu, et al. [234] developed a model for dexterous hand object reorientation using teacher-student learning and an end-to-end vision model. In the recent work by T. Chen, Tippur, et al. [235], they developed a general object reorientation controller to dynamically reorient over 150 different objects in real-time.

For dexterous hand grasping, some works [236, 237] have leveraged analytical methods to model the kinematics and dynamics of both hands and objects. However, these methods typically require simplifications. Q. Liu et al. [238] proposed DexRep, a geometric and spatial hand-object interaction representation combined with PointNet [248], to achieve a generalizable grasping policy. Y. Xu et al. [239] introduce UniDexGrasp, an innovative framework that attains versatile robotic dexterous manipulation by merging a probabilistic model for generating a variety of grasp hypotheses with a goal-conditioned policy for their execution. Wan et al. [240] optimized UniDexGrasp by leveraging the geometry features of tasks, significantly improving generalizability and demonstrating universal dexterous grasping on thousands of objects.

Compared to tasks performed by dexterous hands, dual-arm robots offer a more diverse range of task scenarios. Dual-arm tasks are characterized by collaborative constraints, which can be divided into two categories: unity collaboration and task-based collaboration. Task-based collaboration involves each arm handling different objects independently, such as in shaft-hole assembly [241] or cooking [242]. Conversely, unity collaboration involves both arms working on a single object simultaneously, as demonstrated in activities like folding clothes [243] or performing hand-over tasks [244, 245]. W. Wang et al. [246] focuses on enhancing the robustness of dual-

TABLE 4 | Comparison of different single skill learning approaches.

Skill name	Year	Methodology	Reference
Dexterous hand reorientation	2020	Deep reinforcement learning	[230]
	2019	Deep reinforcement learning	[231]
	2023	Rapid motor adaptation, extrinsic vector	[232]
	2023	Vision-based tactile sensors, extrinsic vector	[233]
	2022	End-to-end vision model, imitation learning	[234]
	2023	End-to-end vision model, imitation learning	[235]
Dexterous hand grasping	2014	Empirical model	[236]
	2010	Push-grasping, motion planning algorithm	[237]
	2023	Geometric and spatial interaction representation	[238]
	2023	Probabilistic model of grasp pose conditioned, imitation learning	[239]
	2023	End-to-end vision model, imitation learning	[240]
Dual-arm operations	2022	Model-free RL, modular approach	[241]
	2022	Motion planning algorithm, motion sequence graph	[242]
	2018	Stereo vision system	[243]
	2020	Regrasp graph, multimodal planning	[244]
	2020	Conservative estimation planning, optimistic estimation planning	[245]
	2024	Shared-control teleoperation, coupled LPV dynamics	[246]
	2024	Learning-based perception, optimization-based control	[247]

arm object-moving tasks through imitation learning, employing a shared teleoperation strategy and a coupled linear parameter-varying dynamical system to ensure precise and stable task execution, even in the presence of disturbances. Meanwhile, Ren et al. [247] introduces a comprehensive framework that enhances the dexterity and versatility of dual-arm manipulators through a learning-based perception module and an optimization-based control module, enabling effective bimanual grasping and manipulation of unseen objects.

These studies collectively advance the field of single skill learning in humanoid robots, providing robust and versatile solutions for a wide range of tasks. They lay a solid foundation for future research on multi-skill learning and the execution of more complex tasks.

5.4.2.2 | Multi-Task Skill Learning.

Rather than training a unique policy for each task, multi-task learning in robotic manipulation seeks to develop a unified policy capable of handling diverse tasks. Noticing the significant advances in Natural Language Processing (NLP) and Computer Vision (CV), numerous studies [249–252] leveraged large vision and/or language models to extract structured task and environment information. This information supports low-level planning frameworks to generate robot trajectories. For example, VoxPoser [245] incorporates affordances and constraints derived from vision-language models (VLMs) into a model-based planning framework. Similarly, ReKep [252] utilizes large vision models (LVMs) and VLMs to generate keypoints and constraints, formulating the manipulation task as a constrained optimization problem. Beyond simply applying existing foundation models in robotics, we also see considerable potential for the development of models tailored specifically to robotics, notably Vision Language Action (VLA) models. These models, trained on demonstration data, generate robot actions based on visual and language input. Examples include RT-1 [253], RT-2 [254], RoboFlamingo [255], and Octo [256], all of which have shown enhanced performance in real-world robotic manipulation tasks. To improve 3D spatial reasoning, many works [257–264] focused on advancing 3D representations. PerAct [265], for example, took input voxelized point clouds, while ChainedDiffuser [257] and Act3D [258] represented scenes as multi-scale 3D feature clouds. However, VLA models require extensive demonstration data, prompting the collection of large-scale robotic manipulation datasets, such as Open X-Embodiment [266] and DROID [267]. Additional studies explore how to synthesize large-scale demonstrations automatically, either through human-designed rules [268] or foundation models [269, 270]. Another line of works [271, 272] attempted to replicate success in the field of CV, where large datasets pretrain a base model later fine-tuned on specific tasks with fewer data. And Lin et al. [273] explore whether data scaling laws observed in NLP could similarly impact robotics.

Despite these advancements, substantial challenges remain. A generalist robotic model, akin to ChatGPT in NLP, has yet to be realized. Fully exploiting web-scale, action-less human video data and enabling models for active, incremental learning are ongoing challenges. Integrating additional modalities, such as force data, to achieve high-precision tasks also presents a promising research direction.

5.4.2.3 | Long-Horizon Manipulation.

Long-horizon manipulation involves tasks that require a series of actions over an extended period of time, such as assembling a product or cooking a meal. These tasks demand the robot to plan and execute multiple steps, coordinating its actions over a long timeframe while continually adjusting to feedback. Long-horizon manipulation is one of the most complex areas of robotic skill learning, as it requires not only the execution of sequential actions but also the ability to handle uncertainty and adapt to changing conditions.

Learning motor strategies directly from long-term demonstrations has always been a challenging task in robotics [274–277]. In a home scenario, the robot must have the ability to perform multiple tasks, such as cooking, through visual observation. However, problems such as compound errors [277–279] in the execution of long sequence tasks and huge visual observation space increase the difficulty of vision-based imitation learning. A common solution is to break a long task into multiple subtasks and use a hierarchical learning architecture to reduce the difficulty of imitation learning. Two main research areas are presented here.

Hierarchical reinforcement learning [280, 281] is one of the important methods to solve the long sequence problem, which mainly realizes the discovery of sub-skills and the planning of skills through self-exploration. Options framework [282, 283] and MAXQ Method [284] are two common methods in this field. Options are used to interrupt the planning process, and the concept of subgoals is used to optimize options itself, to achieve better performance. Some work uses the time-optimization method to solve the generalization problem of new tasks, using the skill-chaining functions [283, 285–290] to determine the relationship between the termination condition of the current execution skill and the start condition of the next skill. The operations are performed in the link mode until the final target condition is reached. For example, Bagaria and Konidaris [283] combined skill-chaining with deep neural networks to achieve the effect that the execution of the previous skill satisfies the conditional properties of the subsequent skill execution. However, this approach requires learning sub-skills and planning strategies from scratch, which is computationally expensive.

Hierarchical Imitation is another area of research that addresses the long sequence problem. Hierarchical behavior cloning is a common method [270, 277, 291–298] that has demonstrated strong capabilities in the field of robotics. The approach utilizes a high-level planner to predict skills and uses a low-level control module to execute to sub-goals set by a high-level planner. Some studies learn from human demo videos and mimic human behavior. The method needs to be able to understand the task representation semantically. For example, extracting human priors from human videos [294, 298], or combining robot presentations with human presentations to align task representations [296], the embedding of these continuous skills achieves goal-oriented control and can achieve better results with a small amount of demonstration data in previously unseen tasks. Some other works learn the potential skill representation of subtasks in an unsupervised way. For example, Zhu et al. [291] uses hierarchical clustering to cluster a complete presentation data into several different skills and uses a meta controller to combine skills and provide target guidance for skills.

One-shot imitation learning is another common approach to Hierarchical Imitation, which adopts a meta-learning framework. Learning a new task [293, 299–302] based on a complete or partial trajectory demonstration video [292, 293], this method can quickly adapt to the optimal strategy of the new task. For example, S. Wu et al. [303] combines DMP with the meta-learning framework. DMP is used to encode a skill repertoire consisting of movement primitives, and the upper-level meta-learning framework provides the task parameters for each sub-task.

6 | Potential Application Scenarios

The potential applications of humanoid robots span a wide array of domains, each offering unique opportunities and challenges. As the capabilities of humanoid robots continue to evolve, their integration into various sectors is becoming more feasible, allowing them to perform tasks traditionally handled by humans. From enhancing workplace productivity to providing companionship and service, humanoid robots are poised to have a profound impact on society. Understanding the various application scenarios where humanoid robots can be deployed is essential for evaluating their long-term potential and societal impact.

In this section, we will explore several promising application domains for humanoid robots, including their role as ordinary workers, educational companions, interactive service providers, domestic attendants, and explorers or rescuers. Each of these scenarios highlights the versatility of humanoid robots and the diverse needs they can address in both professional and personal settings.

6.1 | Ordinary Workers

As the population in industrialized countries continues to age, labor costs in production have been steadily rising. Simultaneously, the new generation of young people is increasingly reluctant to enter the manufacturing sector, leading to soaring expenses associated with training skilled workers. In this landscape, robots have emerged as a vital alternative to traditional labor. Over the past 2 decades, industrial robots have firmly established their roles, while collaborative robots enable seamless interaction between machines and humans in shared workspaces.

Despite these advancements, a significant portion of factory work still relies on human labor. Many of these tasks are repetitive and tedious, yet their inherent unpredictability poses challenges for existing robotic systems. In response to this need, humanoid robots have been proposed as a solution to undertake such tasks. For instance, the HRP-2 project in Japan [304] has been documented for its application in aircraft assembly, while Tesla's humanoid robot, Optimus [305], has been showcased during tours of the company's robot factory.

6.2 | Interactive Service Providers

Humanoid robots can also provide convenience to people through interaction. They can serve as staff in various settings such as in restaurants, or as tour guides at attractions. They can also act as receptionist in companies. For example, the humanoid service robot Pepper [306], which is designed to help people in business, can display various cues with its eyes and interact with human in an intuitive way. The robot tour guide Promobot [307] can function not only as a museum worker but also be its own exhibit at the same time.

6.3 | Domestic Attendant

Home robots will get more reliable in the future with the advances in mechanical and AI technologies [308]. In addition to doing basic household chores, humanoid robots can also be used at home to assist in caring for the elderly and patients, replacing some of the responsibilities of human caregivers. They can not only help people walk and retrieve items, but also take care of the elderly or aid patients in rehabilitation training. Additionally, they can monitor their living conditions and promptly detect any issues that arise.

6.4 | Explorers and Rescuers

Humanoid robots can replace exploration workers in performing tasks in hazardous environments like underground mines and nuclear power stations, and replace firefighters in fire scenes [309]. They can even access areas that are unreachable for humans. For example, NASA's Valkyrie robot [310] is envisioned for use in space exploration missions, particularly for tasks on other planets or moons where human presence may be limited.

7 | Open Challenges and General Trends

As humanoid robots continue to evolve and move toward broader application, several open challenges must be addressed to fully realize their potential. These challenges span a variety of domains. For humanoid robot head, it's from the appearance design to natural emotional interaction. For humanoid robot body, it's from ensuring the security and robustness of robots in dynamic environments to overcoming economic and technical limitations for large-scale deployment. Addressing these challenges is crucial for enabling humanoid robots to operate safely, effectively, and efficiently in real-world scenarios.

In addition to these challenges, there are several emerging trends in humanoid robotics that promise to shape the future of the field. These trends focus on enhancing the capabilities of humanoid robots, improving their versatility and adaptability, and making them more accessible and sustainable. Trends such as cost-effective robot platforms, modularization, interface standardization, and the development of embodied intelligence are gaining significant attention from researchers and developers, as they can drastically improve both the functionality and practicality of humanoid robots.

This section will first discuss the current open challenges that are hindering the progress of humanoid robots, followed by an exploration of the key trends shaping the future of humanoid robotics. Together, these challenges and trends provide a comprehensive view of the ongoing developments in the field, highlighting both the obstacles that need to be overcome and the innovations that are driving the next generation of humanoid robots, as illustrated in Figure 10.

7.1 | Appearance Design

Designing the appearance of humanoid robot heads raises unique considerations for both non-anthropomorphic and anthropomorphic humanoid robots. Non-anthropomorphic heads benefit from human-like contours with simplified features, such as cartoon-like shapes and eye style, to enhance comfort and approachability. However, achieving the balance between human-like qualities and simplification, while avoiding the uncanny valley, remains a key research focus. For anthropomorphic humanoid robot heads, crossing the uncanny valley is essential. While these robots may feature lifelike skin and hair, they currently lack the nuanced skin textures, color variability, and tactile simulation of human faces. High-fidelity skin color changes, bionic skin, and advanced tactile materials will likely be major research. Developing more refined skin texture simulation materials, introducing flexible materials (such as soft skin), will enable robots to adaptively adjust their facial expressions and body movements in response to environmental changes. Additionally, integrating an intelligent haptic feedback system will allow robots to engage in multi-level, multi-modal emotional communication with humans.

7.2 | Emotional Interaction

Enabling natural emotional expressions is crucial for humanoid robot. High degrees of freedom are necessary for natural micro-expressions, especially in eye, lip, and eyebrow movements. Future innovations must increase actuator freedom, incorporate miniaturized mechanical designs, and employ soft and flexible materials for refined expression control. Natural emotional interaction also requires multimodal emotion perception—robots must read human emotions through visual, auditory, and tactile cues. Current emotion recognition lacks the accuracy and speed needed in complex situations, leading to imprecise emotional judgments. Additionally, emotional systems should automatically adjust and gradually express emotions in response to changing environments and interactions. Integrating existing emotion recognition technologies to enhance the robot's ability to understand human emotions through visual, auditory, and tactile signals will improve the real-time performance and accuracy of multimodal emotion perception systems. This will narrow the gap with human emotional responses. Additionally, introducing a more flexible facial expression generation module will enable the robot to express emotions more naturally through facial and body language.

7.3 | Security and Robustness

One of the critical challenges facing humanoid robots is ensuring their security and robustness, particularly as these

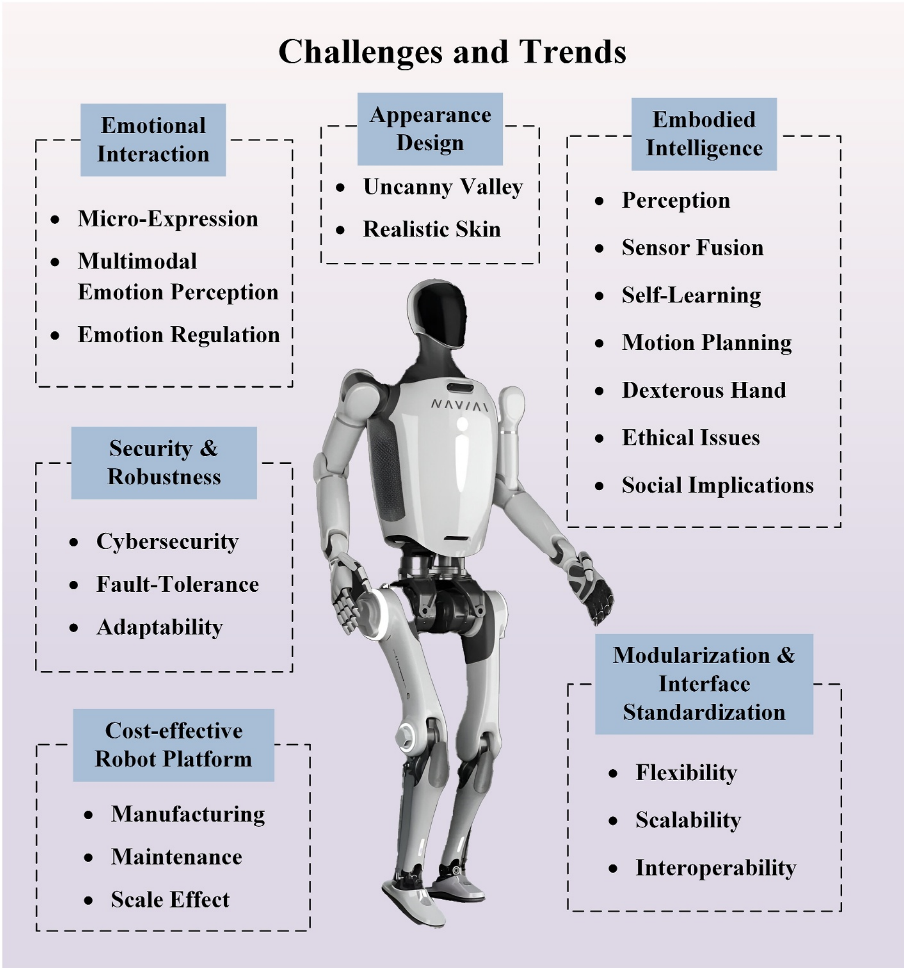

FIGURE 10 | Challenges and trends of humanoid robots.

robots become more integrated into human environments and perform complex tasks. As robots gain greater autonomy and interconnectivity, they must be protected from cyberattacks, malfunctions, and unpredictable environmental conditions. Ensuring that humanoid robots are secure and resilient to errors is essential for their deployment in sensitive areas like healthcare, public service, and industrial settings.

7.4 | Cost-Effective Robot Platform

The high research and manufacturing costs limit the widespread adoption of humanoid robots in the consumer market. Therefore, finding ways to reduce the maintenance and operating costs of robots is essential to make them more viable for commercial applications. As their costs decrease, they'll get commercialized and enter large-scale production, and the prices are expected to decrease, leading to wider adoption in people's daily lives.

7.5 | Modularization and Interface Standardization

The trend toward modularization and interface standardization is becoming increasingly important as humanoid robots are designed for greater flexibility and scalability. By developing robots with standardized, interchangeable components, manufacturers can create more adaptable systems that are easier to maintain, upgrade, and customize for different tasks. This modular approach reduces development time, lowers costs, and promotes innovation by enabling a wider range of configurations and integrations. Additionally, standardizing interfaces ensures compatibility across different robotic systems, facilitating interoperability in multi-robot environments.

7.6 | Embodied Intelligence

Humanoid robots need to accurately perceive their surrounding environment, including objects, obstacles, and human emotions and intentions. These requirements necessitate that robots possess the ability of self-learning and adapt in various environments and tasks, ensuring safety during interactions with humans while also considering ethical and social implications. At the same time, based on a more dexterous hand structure with higher degrees of freedom and in-hand grasping operation strategies, humanoid robots will achieve a level of dexterity and versatility that is, closer to that of natural hands. By addressing these challenges, humanoid robots can advance toward embodied intelligence, achieving higher levels of autonomy, flexibility, and human-robot collaboration. This will facilitate the widespread application of humanoid robots in various fields, including home care, healthcare, and services.

Author Contributions

Zhongxiang Zhou, Rong Xiong, Jun Zou, and Huayong Yang proposed and supervised the project. Xiangyu Mi, Pingyu Xiang, Zhenghan Chen, Haocheng Xu, Xiyang Wu, Shuhao Ye, Jiyu Yu, Yuhan Du, Shichao Zhai, Yifei Yang, Zhichen Lou, Zherui Song, Zikang Yin, and Yu Sun designed the investigation. Qincheng Sheng, Zhongxiang Zhou, Jinhao Li, Shenhan Jia, Yuxiang Cui, and Kechun Xu wrote the manuscript with input from all authors. Qincheng Sheng and Zhongxiang Zhou revised the manuscript.

Acknowledgments

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. T. Geller, "Overcoming the Uncanny Valley," *IEEE Computer Graphics and Applications* 28, no. 4 (2008): 11–17, https://doi.org/10.1109/MCG.2008.79.

2. J. Fink, "Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction," *Social Robotics* 7621 (2012): 199–208, https://doi.org/10.1007/978-3-642-34103-8_20.

3. Unitree Robotics, "Unitree H1: Humanoid Agent AI Avatar," accessed November 21, 2024, https://www.unitree.com/h1.

4. Boston Dynamics, "An Electric New Era for Atlas," accessed November 21, 2024, https://bostondynamics.com/blog/electric-new-era-for-atlas/. Image courtesy of Boston Dynamics, Inc. © 2025. All Rights Reserved.

5. G. Ficht and S. Behnke, "Bipedal Humanoid Hardware Design: A Technology Review," *Current Robotics Reports* 2, no. 2 (2021): 201–210, https://doi.org/10.1007/s43154-021-00050-9.

6. Telsa, "Optimus - Gen 2," accessed November 21, 2024, https://www.youtube.com/watch?v=cpraXaw7dyc.

7. Figure, "Introducing Figure 02: The World's First Commercially-Viable Autonomous Humanoid Robot," accessed November 21, 2024, https://www.figure.ai/.

8. "Humanoid Robot Navigator 2 NAVIAI," accessed November 21, 2024, https://www.worldrobotconference.com/ex/product/810.html.

9. UBTECH, "Walker: Humanoid Service Robot Walk Into the Future," accessed November 21, 2024, https://www.ubtrobot.com/en/humanoid/products/Walker.

10. K. Berns and A. Ashok, "'You Scare Me': The Effects of Humanoid Robot Appearance, Emotion, and Interaction Skills on Uncanny Valley Phenomenon," *Actuators* 13, no. 10 (2024): 419, https://doi.org/10.3390/act13100419.

11. "1X Technologies: The Smarter Working Android," accessed November 21, 2024, https://www.1x.tech/.

12. N. Lorenzo, B. Chiara, P. Daniele, W. Agnieszka, and M. Giorgio, "iCub: The Not-Yet-Finished Story of Building a Robot Child," *Science Robotics* 2, no. 13 (2017): eaaq1026, https://doi.org/10.1126/scirobotics.aaq1026.

13. D. Gouaillier, H. Vincent, P. Blazevic, et al., "Mechatronic Design of NAO Humanoid," in *IEEE International Conference on Robotics and Automation* (2009), 769–774, https://doi.org/10.1109/ROBOT.2009.5152516.

14. B. G. Draghici, A. Elena Dobre, M. Misaros, and P. S. Ovidiu, "Development of a Human Service Robot Application Using Pepper Robot as a Museum Guide," in *IEEE International Conference on Automation, Quality and Testing, Robotics* (2022), 1–5, https://doi.org/10.1109/AQTR55203.2022.9802037.

15. A. Kalouguine, V. De-León-Gómez, C. Chevallereau, S. Dalibard, and Y. Aoustin, "A New Human-Like Walking for the Humanoid Robot Romeo," *Multibody System Dynamics* 53, no. 4 (2021): 411–434, https://doi.org/10.1007/s11044-021-09805-w.

16. E. Guizzo, "The Little Robot That Could...Maybe," *IEEE Spectrum* 53, no. 1 (2016): 58–62, https://doi.org/10.1109/MSPEC.2016.7367471.

17. F. Hegel, F. Eyssel, and B. Wrede, "The Social Robot 'Flobi': Key Concepts of Industrial Design," in *IEEE International Workshop on Robot and Human Interactive Communication* (2010), 107–112, https://doi.org/10.1109/ROMAN.2010.5598691.

18. F. Faber, M. Bennewitz, C. Eppner, et al., "The Humanoid Museum Tour Guide Robotinho," in *The 18th IEEE International Symposium on Robot and Human Interactive Communication* (2009), 891–896, https://doi.org/10.1109/ROMAN.2009.5326326.

19. T. Kishi, T. Otani, N. Endo, et al., "Development of Expressive Robotic Head for Bipedal Humanoid Robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2012), 4584–4589, https://doi.org/10.1109/IROS.2012.6386050.

20. InMoov, "Open Source 3D Printed Life-Size Robot," accessed November 21, 2024, https://inmoov.fr/.

21. Xiaomi, "Xiaomi Unveils CyberOne — Humanoid Robot Exploring Frontiers of Connected Living," accessed November 21, 2024, https://www.mi.com/global/discover/article?id=2754.

22. 1X Technologies, "NEO by 1X Technologies: Your Intelligent Android Assistant," accessed November 21, 2024, https://www.1x.tech/neo.

23. UBTECH, "Walker-S Industrial Humanoid Service Robot Walk Into the Future," accessed November 21, 2024, https://www.ubtrobot.com/en/humanoid/products/WalkerS.

24. A. Meghdari, A. Shariati, M. Alemi, A. A. Nobaveh, M. Khamooshi, and B. Mozaffari, "Design Performance Characteristics of a Social Robot Companion 'Arash' for Pediatric Hospitals," *International Journal of Humanoid Robotics* 15, no. 05 (2018): 1850019, https://doi.org/10.1142/S0219843618500196.

25. E. Ferrari, B. Robins, and K. Dautenhahn, "Therapeutic and Educational Objectives in Robot Assisted Play for Children With Autism," in *IEEE International Symposium on Robot and Human Interactive Communication* (2009), 108–114, https://doi.org/10.1109/ROMAN.2009.5326251.

26. ROBOPRENEUR, "Mirokai: Harnessing the Future of Robotics Today," accessed November 21, 2024, https://www.robopreneur.com/mirokai.

27. S. Al Moubayed, G. Skantze, and J. Beskow, "The Furhat Back-Projected Humanoid Head–Lip Reading, Gaze and Multi-Party Interaction," *International Journal of Humanoid Robotics* 10, no. 01 (2013): 1350005, https://doi.org/10.1142/S0219843613500059.

28. Y. Nakata, S. Yagi, S. Yu, et al., "Development of 'Ibuki' an Electrically Actuated Childlike Android With Mobility and Its Potential in the Future Society," *Robotica* 40, no. 4 (2022): 933–950, https://doi.org/10.1017/S0263574721000898.

29. Zhejiang University, "Meet Xiao Yao, The Fourth Generation Interactive Face Robot," accessed November 21, 2024, https://www.youtube.com/watch?v=aijydBb1JHw.

30. Engineered Arts, "Ameca: The Future Face of Robotics," accessed November 21, 2024, https://engineeredarts.co.uk/robot/ameca/.

31. Z. Yu, G. Ma, and Q. Huang, "Modeling and Design of a Humanoid Robotic Face Based on an Active Drive Points Model," *Advanced Robotics* 28, no. 6 (2014): 379–388, https://doi.org/10.1080/01691864.2013.867290.

32. Hanson Robotics, "Sophia," accessed November 21, 2024, https://www.hansonrobotics.com/sophia/.

33. S. Kajita, K. Kaneko, F. Kaneiro, et al., "Cybernetic Human HRP-4C: A Humanoid Robot With Human-Like Proportions," *Robotics Research* 70 (2011): 301–314, https://doi.org/10.1007/978-3-642-19457-3_18.

34. M. Riaz, I. Nawaz, and M. Sohail, "Novel Concepts of Soft Multi Rough Sets With MCGDM for Selection of Humanoid Robot," *Punjab University Journal of Mathematics* 52, no. 2 (2020): 111–137.

35. D. F. Glas, T. Minato, C. T. Ishi, T. Kawahara, and H. Ishiguro, "ERICA: The ERATO Intelligent Conversational Android," in *IEEE International Symposium on Robot and Human Interactive Communication* (2016), 22–29, https://doi.org/10.1109/ROMAN.2016.7745086.

36. H. Kang, M. Ben Moussa, and N. M. Thalmann, "Nadine: A Large Language Model-Driven Intelligent Social Robot With Affective Capabilities and Human-Like Memory," *Computer Animation and Virtual Worlds* 35, no. 4 (2024): e2290, https://doi.org/10.1002/cav.2290.

37. Z. Faraj, M. Selamet, C. Morales, et al., "Facially Expressive Humanoid Robotic Face," *HardwareX* 9 (2021): e00117, https://doi.org/10.1016/j.ohx.2020.e00117.

38. Y. Hu, B. Chen, J. Lin, et al., "Human-Robot Facial Coexpression," *Science Robotics* 9, no. 88 (2024): eadi4724, https://doi.org/10.1126/scirobotics.adi4724.

39. B. Li, H. Li, and H. Liu, "Driving Animatronic Robot Facial Expression From Speech," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2024), 7012–7019, https://doi.org/10.1109/IROS58592.2024.10801970.

40. M.-J. Han, C.-H. Lin, and K.-T. Song, "Robotic Emotional Expression Generation Based on Mood Transition and Personality Model," *IEEE Transactions on Cybernetics* 43, no. 4 (2013): 1290–1303, https://doi.org/10.1109/TSMCB.2012.2228851.

41. Y. Li, L. Zhu, Z. Zhang, et al., "Humanoid Robot Heads for Human-Robot Interaction: A Review," *Science China Technological Sciences* 67, no. 2 (2023): 357–379, https://doi.org/10.1007/s11431-023-2493-y.

42. J. Zhang, J. Zheng, and N. Magnenat-Thalmann, "PCMD: Personality-Characterized Mood Dynamics Model Toward Personalized Virtual Characters," *Computer Animation and Virtual Worlds* 26, no. 3–4 (2015): 237–245, https://doi.org/10.1002/cav.1660.

43. H. Ishihara, Y. Yoshikawa, and M. Asada, "Realistic Child Robot 'Affetto' for Understanding the Caregiver-Child Attachment Relationship That Guides the Child Development," *IEEE International Conference on Development and Learning* 2 (2011): 1–5, https://doi.org/10.1109/DEVLRN.2011.6037346.

44. T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, "Development of the Face Robot SAYA for Rich Facial Expressions," in *SICE-ICASE International Joint Conference* (2006), 5423–5428, https://doi.org/10.1109/SICE.2006.315537.

45. F. H. David and W. Victor, "Converging the Capabilities of EAP Artificial Muscles and the Requirements of Bio-Inspired Robotics," *Smart Structures and Materials* (2004), https://doi.org/10.1117/12.543095.

46. P. Ekman, V. F. Wallace, and J. C. Hager, *Facial Action Coding System: The Manual* (Research Nexus, 2002), 512.

47. D. Yang, W. Sato, Q. Liu, T. Minato, S. Namba, and S. Nishida, "Optimizing Facial Expressions of an Android Robot Effectively: A Bayesian Optimization Approach," in *IEEE-RAS International Conference on Humanoid Robots* (2022), 542–549, https://doi.org/10.1109/Humanoids53995.2022.10000154.

48. Y. Tadesse, D. Hong, and S. Priya, "Twelve Degree of Freedom Baby Humanoid Head Using Shape Memory Alloy Actuators," *Journal of Mechanisms and Robotics* 3, no. 1 (2011): 011008, https://doi.org/10.1115/1.4003005.

49. G. Pioggia, R. Igliozzi, M. Ferro, A. Ahluwalia, F. Muratori, and D. De Rossi, "An Android for Enhancing Social Skills and Emotion Recognition in People With Autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 13, no. 4 (2005): 507–515, https://doi.org/10.1109/TNSRE.2005.856076.

50. Y. Li, L. Zhu, Z. Zhang, et al., "Modular Multilayered PVC Gel Soft Actuators for Novel Lightweight Humanoid Facial Robot," *Sensors and Actuators A: Physical* 378 (2024): 115825, https://doi.org/10.1016/j.sna.2024.115825.

51. H. Kiilavuori, V. Sariola, M. J. Peltola, and J. K. Hietanen, "Making Eye Contact With a Robot: Psychophysiological Responses to Eye Contact With a Human and With a Humanoid Robot," *Biological Psychology* 158 (2021): 107989, https://doi.org/10.1016/j.biopsycho.2020.107989.

52. N. Pateromichelakis, A. Mazel, M. A. Hache, et al., "Head-Eyes System and Gaze Analysis of the Humanoid Robot Romeo," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2014), 1374–1379, https://doi.org/10.1109/IROS.2014.6942736.

53. M. Raković, D. N. Ferreira, J. Marques, A. Billard, and J. Santos-Victor, "The Gaze Dialogue Model: Nonverbal Communication in HHI and HRI," *IEEE Transactions on Cybernetics* 54, no. 4 (2024): 2026–2039, https://doi.org/10.1109/TCYB.2022.3222077.

54. A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot," *IEEE Transactions on Human-Machine Systems* 44, no. 2 (2014): 157–168, https://doi.org/10.1109/THMS.2014.2303083.

55. H. McGurk and J. Macdonald, "Hearing Lips and Seeing Voices," *Nature* 264, no. 5588 (1976): 746–748, https://doi.org/10.1038/264746a0.

56. C. Strathearn and E. M. Ma, "A Novel Speech to Mouth Articulation System for Realistic Humanoid Robots," *Journal of Intelligent and Robotic Systems* 101, no. 3 (2021): 54, https://doi.org/10.1007/s10846-021-01332-2.

57. M. Heisler, S. Kopp, and C. Becker-Asano, "Making an Android Robot Head Talk," in *IEEE International Conference on Robot and Human Interactive Communication* (2023), 1837–1842, https://doi.org/10.1109/RO-MAN57019.2023.10309532.

58. U. Martinez-Hernandez, A. Rubio-Solis, and T. J. Prescott, "Bayesian Perception of Touch for Control of Robot Emotion," in *International Joint Conference on Neural Networks* (2016), 4927–4933, https://doi.org/10.1109/IJCNN.2016.7727848.

59. P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion* 6, no. 3–4 (1992): 169–200, https://doi.org/10.1080/02699939208411068.

60. J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology* 39, no. 6 (1980): 1161–1178, https://doi.org/10.1037/h0077714.

61. A. Mehrabian, "Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament," *Current Psychology* 14, no. 4 (1996): 261–292, https://doi.org/10.1007/BF02686918.

62. R. Plutchik, "A Psychoevolutionary Theory of Emotions," *Social Science Information/sur les Sciences Sociales* 21, no. 4–5 (1982): 529–553, https://doi.org/10.1177/053901882021004003.

63. A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions* (Cambridge University Press, 1988), 218.

64. R. W. Picard, *Affective Computing* (MIT Press, 1997), 292.

65. R. R. McCrae and P. T. Costa, "Validation of the Five-Factor Model of Personality Across Instruments and Observers," *Journal of Personality and Social Psychology* 52, no. 1 (1987): 81–90, https://doi.org/10.1037/0022-3514.52.1.81.

66. O. Stasse and F. Thomas, "An Overview of Humanoid Robots Technologies," *Biomechanics of Anthropomorphic Systems* 124 (2019): 281–310, https://doi.org/10.1007/978-3-319-93870-7_13.

67. S. Lohmeier, *Design and Realization of a Humanoid Robot for Fast and Autonomous Bipedal Locomotion* (Technische Universität München, 2010), 214.

68. K. Hashimoto, "Mechanics of Humanoid Robot," *Advanced Robotics* 34, no. 21–22 (2020): 1390–1397, https://doi.org/10.1080/01691864.2020.1813624.

69. Unitree Robotics, "Unitree G1: Humanoid Agent AI Avatar," accessed November 21, 2024, https://www.unitree.com/g1.

70. U. Kim, D. Jung, H. Jeong, et al., "Integrated Linkage-Driven Dexterous Anthropomorphic Robotic Hand," *Nature Communications* 12, no. 1 (2021): 1–13, https://doi.org/10.1038/s41467-021-27261-0.

71. S. R. Kashef, S. Amini, and A. Akbarzadeh, "Robotic Hand: A Review on Linkage-Driven Finger Mechanisms of Prosthetic Hands and Evaluation of the Performance Criteria," *Mechanism and Machine Theory* 145 (2020): 103677, https://doi.org/10.1016/j.mechmachtheory.2019.103677.

72. Tesla, "Tesla AI Day 2022," accessed November 21, 2024, https://www.youtube.com/watch?v=ODSJsviD_SU.

73. ASIMO by Honda, "The World's Most Advanced Humanoid Robot," accessed November 21, 2024, https://asimo.honda.com.

74. Boston Dynamics, "Sick Tricks and Tricky Grips," accessed November 21, 2024, https://bostondynamics.com/blog/sick-tricks-and-tricky-grips/.

75. S.-H. Hyon, D. Suewaka, Y. Torii, and N. Oku, "Design and Experimental Evaluation of a Fast Torque-Controlled Hydraulic Humanoid Robot," *IEEE* 22, no. 2 (April 2017): 623–634, https://doi.org/10.1109/TMECH.2016.2628870.

76. J. Morimoto, C. G. Atkeson, G. Endo, and G. Cheng, "Improving Humanoid Locomotive Performance With Learnt Approximated Dynamics via Gaussian Processes for Regression," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, October* (IEEE, 2007), 4234–4240, https://doi.org/10.1109/IROS.2007.4399485.

77. C. M. Best, J. P. Wilson, and M. D. Killpack, "Control of a Pneumatically Actuated, Fully Inflatable, Fabric-Based, Humanoid Robot," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), Seoul, South Korea, November* (IEEE, 2015), 1133–1140, https://doi.org/10.1109/HUMANOIDS.2015.7363495.

78. FESTO, "Festo Fluidic Muscle Gives Robot Natural Movement," accessed November 21, 2024, https://www.roboticsupdate.com/2016/02/festo-fluidic-muscle-gives-robot-natural-movement/.

79. FESTO, "Airic's_arm Robot Arm With Fluidic Muscle," accessed November 21, 2024, https://www.festo.com/hk/en/e/about-festo/research-and-development/bionic-learning-network/highlights-from-2006-to-2009/airic-s-arm-id_33870/.

80. T. D. Barfoot, *State Estimation for Robotics* (Cambridge University Press, 2017). [2023-11-09].

81. J. Miao, K. Jiang, T. Wen, et al., "A Survey on Monocular Re-Localization: From the Perspective of Scene Map Representation," *IEEE Transactions on Intelligent Vehicles* (2024): 1–33, https://doi.org/10.1109/TIV.2024.3378716.

82. Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE* 111, no. 3 (2023): 257–276, https://doi.org/10.1109/JPROC.2023.3238524.

83. A. Roychoudhury, S. Khorshidi, S. Agrawal, and M. Bennewitz, "Perception for Humanoid Robots," *Current Robotics Reports* 4, no. 4 (2023): 127–140, https://doi.org/10.1007/s43154-023-00107-x.

84. A. Bonci, C. Cheng, P. David, I. Marina, N. Giacomo, and F. Sibona, "Human-Robot Perception in Industrial Environments: A Survey," *Sensors* 21, no. 5 (2021): 1571, https://doi.org/10.3390/s21051571.

85. Y. Yang, P. Geneva, and G. Huang, "Multi-Visual-Inertial System: Analysis, Calibration, and Estimation," *International Journal of Robotics Research* 43, no. 13 (2024): 1995–2026, https://doi.org/10.1177/02783649241245726.

86. Y. He, H. Yu, Y. Wen, and S. Scherer, "Toward Efficient and Robust Multiple Camera Visual-Inertial Odometry," *arXiv* (2021), accessed November 12, 2024, http://arxiv.org/abs/2109.12030.

87. T. Qin, P. Li, and S. Shen, "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator," *IEEE Transactions on Robotics* 34, no. 4 (2018): 1004–1020, https://doi.org/10.1109/TRO.2018.2853729.

88. P. Geneva, E. Kevin, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A Research Platform for Visual-Inertial Estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), 4666–4672, https://doi.org/10.1109/ICRA40945.2020.9196524.

89. R. Hartley, M. Ghaffari, R. M. Eustice, and J. W. Grizzle, "Contact-Aided Invariant Extended Kalman Filtering for Legged Robot State Estimation," *International Journal of Robotics Research* 39, no. 4 (2020): 402–430, https://doi.org/10.1177/0278364919894385.

90. C. Cadena, L. Carlone, H. Carrillo, et al., "Past, Present, and Future of Simultaneous Localization & Mapping: Towards the Robust-Perception Age," *IEEE Transactions on Robotics* 32, no. 6 (2016): 1309–1332, https://doi.org/10.1109/TRO.2016.2624754.

91. J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, no. 3 (2018): 611–625, https://doi.org/10.1109/TPAMI.2017.2658577.

92. C. Campos, R. Elvira, J. J. Gómez Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM," *IEEE Transactions on Robotics* 37, no. 6 (2021): 1874–1890, https://doi.org/10.1109/TRO.2021.3075644.

93. G. Huang, "Visual-Inertial Navigation: A Concise Review," in *2019 International Conference on Robotics and Automation (ICRA)* (2019), 9572–9582, https://doi.org/10.1109/ICRA.2019.8793604.

94. A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-Aided Inertial Navigation," in *Proceedings 2007 IEEE International Conference on Robotics and Automation* (2007), 3565–3572, https://doi.org/10.1109/ROBOT.2007.364024.

95. A. Barrau and B. Silvere, "The Invariant Extended Kalman Filter as a Stable Observer," *IEEE Transactions on Automatic Control* 62, no. 4 (2016): 1797–1812, https://doi.org/10.1109/TAC.2016.2594085.

96. N. Sünderhauf and P. Protzel, "Towards a Robust Back-End for Pose Graph SLAM," in *2012 IEEE International Conference on Robotics and Automation* (2012), 1254–1261, https://doi.org/10.1109/ICRA.2012.6224709.

97. H. Alismail, B. Browning, and L. Simon, "Photometric Bundle Adjustment for Vision-Based SLAM," in *Computer Vision–ACCV 2016*, Vol. 10114 (2017), 324–341, https://doi.org/10.1007/978-3-319-54190-7_20.

98. S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, "Bundle Adjustment in the Large," in *Computer Vision - ECCV 2010*, Vol. 6312 (2010), 29–42, https://doi.org/10.1007/978-3-642-15552-9_3.

99. B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon, "Bundle Adjustment - A Modern Synthesis. International Workshop on Vision Algorithms," *Theory and Practice* (1999): 298–372, https://doi.org/10.1007/3-540-44480-7_21.

100. A. M. Elsergany, M. F. Abdel-Hafez, and M. A. Jaradat, "Novel Augmented Quaternion UKF for Enhanced Loosely Coupled GPS/INS Integration," *IEEE Transactions on Control Systems Technology* 32, no. 6 (2024): 2321–2331, https://doi.org/10.1109/TCST.2024.3425211.

101. A. Wang, R. Zhong, K. Zheng, and H. Fang, "LL-SLAM: Lightweight Loosely-Coupled Visual-Inertial SLAM," in *42nd Chinese Control Conference* (2023), 3632–3637, https://doi.org/10.23919/CCC58697.2023.10239955.

102. D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," *arXiv* (2018). abs/1712.07629, https://doi.org/10.48550/arXiv.1712.07629.

103. P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning Feature Matching With Graph Neural Networks," *arXiv* (2019). abs/1911.11763, https://doi.org/10.48550/arXiv.1911.11763.

104. P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local Feature Matching at Light Speed," in *IEEE/CVF International Conference on Computer Vision* (2023), 17581–17592, https://doi.org/10.1109/ICCV51070.2023.01616.

105. K. Xu, Y. Hao, S. Yuan, C. Wang, and L. Xie, "AirSLAM: An Efficient and Illumination-Robust Point-Line Visual SLAM System," *arXiv* (2024). abs/2408.03520, https://doi.org/10.48550/arXiv.2408.03520.

106. H. M. S. Bruno and E. L. Colombini, "LIFT-SLAM: A Deep-Learning Feature-Based Monocular Visual SLAM Method," *arXiv* (2021). abs/2104.00099, https://doi.org/10.48550/arXiv.2104.00099.

107. Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," *arXiv* (2021). abs/2108.10869, https://doi.org/10.48550/arXiv.2108.10869.

108. X. Wang, N. Wang, and G. Zhang, "XRDSLAM: A Flexible and Modular Framework for Deep Learning Based SLAM," *arXiv* (2024). abs/2410.23690, https://doi.org/10.48550/arXiv.2410.23690.

109. Z. C. Guo, V. Korotkine, J. R. Forbes, and T. D. Barfoot, "Koopman Linearization for Data-Driven Batch State Estimation of Control-Affine Systems," *IEEE Robotics and Automation Letters* 7, no. 2 (2022): 866–873, https://doi.org/10.1109/LRA.2021.3133587.

110. C. Siagian and L. Itti, "Biologically Inspired Mobile Robot Vision Localization," *IEEE Transactions on Robotics* 25, no. 4 (2009): 861–873, https://doi.org/10.1109/TRO.2009.2022424.

111. H. Jégou, D. Matthijs, C. Schmid, and P. Pérez, "Aggregating Local Descriptors Into a Compact Image Representation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), 3304–3311, https://doi.org/10.1109/CVPR.2010.5540039.

112. D. Galvez-López and J. D. Tardos, "Bags of Binary Words for Fast Place Recognition in Image Sequences," *IEEE Transactions on Robotics* 28, no. 5 (2012): 1188–1197, https://doi.org/10.1109/TRO.2012.2197158.

113. G. Tolias, Y. Avrithis, and H. Jegou, "To Aggregate or Not to Aggregate: Selective Match Kernels for Image Search," in *IEEE International Conference on Computer Vision* (2013), 1401–1408, https://doi.org/10.1109/ICCV.2013.177.

114. N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2005), 886–893, https://doi.org/10.1109/CVPR.2005.177.

115. R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, no. 6 (2017): 1437–1451, https://doi.org/10.1109/TPAMI.2017.2711011.

116. R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "TransVPR: Transformer-Based Place Recognition With Multi-Level Attention Aggregation," *arXiv* (2022): 13638–13647. abs/2201.02001, https://doi.org/10.48550/arXiv.2201.02001.

117. A. Ali-bey, B. Chaib-draa, and P. Giguère, "MixVPR: Feature Mixing for Visual Place Recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2023), 2998–3007.

118. N. Keetha, A. Mishra, J. Karhade, et al., "AnyLoc: Towards Universal Visual Place Recognition," *IEEE Robotics and Automation Letters* 9, no. 2 (2024): 1286–1293, https://doi.org/10.1109/LRA.2023.3343602.

119. A.-Q. Cao and R. de Charette, "MonoScene: Monocular 3D Semantic Scene Completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 3991–4001.

120. Y. Li, Z. Yu, C. Choy, et al., "VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 9087–9098.

121. Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 9223–9232.

122. X. Wang, Z. Zhu, W. Xu, et al., "OpenOccupancy: A Large Scale Benchmark for Surrounding Semantic Occupancy Perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 17850–17859.

123. J. Behley, M. Garbade, A. Milioto, et al., "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

124. Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 21729–21740.

125. X. Tian, T. Jiang, L. Yun, et al., "Occ3D: A Large-Scale 3D Occupancy Prediction Benchmark for Autonomous Driving," *Advances in Neural Information Processing Systems* 36 (2023): 64318–64330, https://proceedings.neurips.cc/paper_files/paper/2023/file/cabfaeecaae7d6540ee797a66f0130b0-Paper-Datasets_and_Benchmarks.pdf.

126. M. Pan, J. Liu, R. Zhang, et al., "RenderOcc: Vision-Centric 3D Occupancy Prediction With 2D Rendering Supervision," in *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)* (2024), 12404–12411, https://doi.org/10.1109/ICRA57147.2024.10611537.

127. C. Zhang, J. Yan, W. Yi, et al., "OccNeRF: Advancing 3D Occupancy Prediction in LiDAR-Free Environments," *arXiv* (2023). abs/2312.09243, https://doi.org/10.48550/arXiv.2312.09243.

128. Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, "SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 19946–19956.

129. W. Gan, N. Mo, H. Xu, and N. Yokoya, "A Simple Framework for 3D Occupancy Estimation in Autonomous Driving," *arXiv* (2023). abs/2303.10076, https://doi.org/10.48550/arXiv.2303.10076.

130. W. Gan, F. Liu, H. Xu, N. Mo, and N. Yokoya, "GaussianOcc: Fully Self-Supervised and Efficient 3D Occupancy Estimation With Gaussian Splatting," *arXiv* (2024). abs/2408.11447, https://doi.org/10.48550/arXiv.2408.11447.

131. J. Hou, X. Li, W. Guan, et al., "FastOcc: Accelerating 3D Occupancy Prediction by Fusing the 2D Bird's-Eye View and Perspective View," *arXiv* (2024). abs/2403.02710, https://doi.org/10.48550/arXiv.2403.02710.

132. J. Ma, X. Chen, J. Huang, et al., "Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), 21486–21495.

133. J. Li, X. He, C. Zhou, X. Cheng, Y. Wen, and D. Zhang, "ViewFormer: Exploring Spatiotemporal Modeling for Multi-View 3D Occupancy Perception via View-Guided Transformers," in *Computer Vision–ECCV 2024*, Vol. 15101 (2025), 90–106, https://doi.org/10.1007/978-3-031-72775-7_6.

134. C. Sima, W. Tong, T. Wang, et al., "Scene as Occupancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 8406–8415.

135. P. E. Hart, N. J. Nilsson, and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," *IEEE Transactions on Systems Science and Cybernetics* 4, no. 2 (1968): 100–107, https://doi.org/10.1109/TSSC.1968.300136.

136. S. Karaman and F. Emilio, "Sampling-Based Algorithms for Optimal Motion Planning," *International Journal of Robotics Research* 30, no. 7 (2011): 846–894, https://doi.org/10.1177/0278364911406761.

137. F. Niroui, K. Zhang, Z. Kashino, and G. Nejat, "Deep Reinforcement Learning Robot for Search and Rescue Applications: Exploration in Unknown Cluttered Environments," *IEEE Robotics and Automation Letters* 4, no. 2 (2019): 610–617, https://doi.org/10.1109/LRA.2019.2891991.

138. T. Fan, X. Cheng, J. Pan, et al., "Getting Robots Unfrozen and Unlost in Dense Pedestrian Crowds," *IEEE Robotics and Automation Letters* 4, no. 2 (2019): 1178–1185, https://doi.org/10.1109/LRA.2019.2891491.

139. J. A. Placed, J. Strader, H. Carrillo, et al., "A Survey on Active Simultaneous Localization and Mapping: State of the Art and New Frontiers," *IEEE Transactions on Robotics* 39, no. 3 (2023): 1686–1705, https://doi.org/10.1109/TRO.2023.3248510.

140. F. Chen, S. Bai, T. Shan, and B. Englot, "Self-Learning Exploration and Mapping for Mobile Robots via Deep Reinforcement Learning," in *AIAA Scitech 2019 Forum* (2019), 0396, https://doi.org/10.2514/6.2019-0396.

141. R. Cimurs, I. H. Suh, and H. L. Jin, "Goal-Driven Autonomous Exploration Through Deep Reinforcement Learning," *IEEE Robotics and Automation Letters* 7, no. 2 (2021): 730–737, https://doi.org/10.1109/LRA.2021.3133591.

142. Z. Ravichandran, L. Peng, N. Hughes, J. D. Griffith, and L. Carlone, "Hierarchical Representations and Explicit Memory: Learning Effective Navigation Policies on 3D Scene Graphs Using Graph Neural Networks," in *2022 International Conference on Robotics and Automation (ICRA)* (2022), 9272–9279, https://doi.org/10.1109/ICRA46639.2022.9812179.

143. W.-C. Lee, M. C. Lim, and H.-L. Choi, "Extendable Navigation Network Based Reinforcement Learning for Indoor Robot Exploration," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), 11508–11514, https://doi.org/10.1109/ICRA48506.2021.9561040.

144. E. Tolstaya, F. Gama, J. Paulos, P. George, V. Kumar, and &A. Ribeiro, "Learning Decentralized Controllers for Robot Swarms With Graph Neural Networks," in *Conference on Robot Learning* (2020), 671–682.

145. Q. Li, F. Gama, A. Ribeiro, and A. Prorok, "Graph Neural Networks for Decentralized Multi-Robot Path Planning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), 11785–11792, https://doi.org/10.1109/IROS45743.2020.9341668.

146. T. Dang, F. Mascarich, S. Khattak, C. Papachristos, and K. Alexis, "Graph-Based Path Planning for Autonomous Robotic Exploration in Subterranean Environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019), 3105–3112, https://doi.org/10.1109/IROS40897.2019.8968151.

147. R. Reinhart, D. Tung, E. Hand, C. Papachristos, and K. Alexis, "Learning-Based Path Planning for Autonomous Exploration of Subterranean Environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), 1215–1221, https://doi.org/10.1109/ICRA40945.2020.9196662.

148. D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine, "Rapid Exploration for Open-World Navigation With Latent Goal Models," *arXiv Preprint* (2021), https://doi.org/10.48550/arXiv.2104.05859.

149. H. Zhang, J. Cheng, Z. Lin, Y. Li, and W. Zhang, "H2GNN: Hierarchical-Hops Graph Neural Networks for Multi-Robot Exploration in Unknown Environments," *IEEE Robotics and Automation Letters* 7, no. 2 (2022): 3435–3442, https://doi.org/10.1109/LRA.2022.3146912.

150. F. Chen, J. D. Martin, Y. Huang, J. Wang, and B. Englot, "Autonomous Exploration Under Uncertainty via Deep Reinforcement Learning on Graphs," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA* (2020), 6140–6147, https://doi.org/10.1109/IROS45743.2020.9341657.

151. F. Chen, P. Szenher, Y. Huang, et al., "Zero-Shot Reinforcement Learning on Graphs for Autonomous Exploration Under Uncertainty," in *IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China* (2021), 5193–5199, https://doi.org/10.1109/ICRA48506.2021.9561917.

152. J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó, "The SLAM Problem: A Survey," in *Artificial Intelligence Research and Development* (2008), 363–371, https://doi.org/10.3233/978-1-58603-925-7-363.

153. H. Li, Q. Zhang, and D. Zhao, "Deep Reinforcement Learning-Based Automatic Exploration for Navigation in Unknown Environment," *IEEE Transactions on Neural Networks and Learning Systems* 31, no. 6 (2020): 2064–2076, https://doi.org/10.1109/TNNLS.2019.2927869.

154. M. F. Ahmed, K. Masood, V. Fremont, and I. Fantoni, "Active SLAM: A Review on Last Decade," *Sensors* 23, no. 19 (2023): 8097, https://doi.org/10.3390/s23198097.

155. D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to Explore Using Active Neural SLAM," *arXiv Preprint* (2020), https://doi.org/10.48550/arXiv.2004.05155.

156. S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive Mapping and Planning for Visual Navigation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 2616–2625, https://doi.org/10.48550/arXiv.1702.03920.

157. D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural Topological SLAM for Visual Navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 12875–12884, https://doi.org/10.48550/arXiv.2005.12256.

158. D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine, "ViNG: Learning Open-World Navigation With Visual Goal," in *IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China* (2021), 13215–13222, https://doi.org/10.1109/ICRA48506.2021.9561936.

159. Q. Wu, J. Wang, J. Liang, X. Gong, and D. Manocha, "Image-Goal Navigation in Complex Environments via Modular Learning," *IEEE Robotics and Automation Letters* 7, no. 3 (July 2022): 6902–6909, https://doi.org/10.1109/LRA.2022.3178810.

160. J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal Velocity Obstacles for Real-Time Multi-Agent Navigation," in *IEEE International Conference on Robotics and Automation, Pasadena, CA, USA* (2008), 1928–1935, https://doi.org/10.1109/ROBOT.2008.4543489.

161. J. Van den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal N-Body Collision Avoidance," in *Robotics Research. Springer Tracts in Advanced Robotics*, ed. C. Pradalier, R. Siegwart, and G. Hirzinger, Vol. 70 (Springer, 2011), 3–19, https://doi.org/10.1007/978-3-642-19457-3_1.

162. A. J. Sathyamoorthy, U. Patel, T. Guan, and D. Manocha, "Frozone: Freezing-Free, Pedestrian-Friendly Navigation in Human Crowds," *IEEE Robotics and Automation Letters* 5, no. 3 (July 2020): 4352–4359, https://doi.org/10.1109/LRA.2020.2996593.

163. M. Bojarski, "End to End Learning for Self-Driving Cars," *arXiv Preprint* (2016). arXiv:1604.07316.

164. Z. Li, T. Motoyoshi, K. Sasaki, T. Ogata, and S. Sugano, "Rethinking Self-Driving: Multi-Task Knowledge for Better Generalization and Accident Explanation Ability," *arXiv Preprint* (2018). arXiv:1809.11100.

165. N. Rhinehart, R. McAllister, and S. Levine, "Deep Imitative Models for Flexible Inference, Planning, and Control," *arXiv Preprint* (2018). arXiv:1810.06544.

166. Y. F. Chen, M. Liu, M. Everett, and J. P. How, "Decentralized Non-Communicating Multiagent Collision Avoidance With Deep Reinforcement Learning," in *IEEE International Conference on Robotics and Automation (ICRA), Singapore* (2017), 285–292, https://doi.org/10.1109/ICRA.2017.7989037.

167. L. Tai, P. Giuseppe, and M. Liu, "Virtual-to-Real Deep Reinforcement Learning: Continuous Control of Mobile Robots for Mapless Navigation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2017), 31–36, https://doi.org/10.1109/IROS.2017.8202134.

168. L. Kaiser, B. Mohammad, M. Piotr, et al., "Model-Based Reinforcement Learning for Atari," *arXiv* (2019). abs/1903.00374, https://doi.org/10.48550/arXiv.1903.00374.

169. A. D. Ames, C. Samuel, E. Magnus, N. Gennaro, S. Koushil, and P. Tabuada, "Control Barrier Functions: Theory and Applications," in *European Control Conference (ECC)* (2019), 3420–3431, https://doi.org/10.23919/ECC.2019.8796030.

170. J. Norby, Y. Yang, T. Ardalan, et al., "Quad-SDK: Full Stack Software Framework for Agile Quadrupedal Locomotion," in *ICRA Workshop on Legged Robot* (2022).

171. R. Grandia, J. Fabian, S. Yang, F. Farbod, and H. Marco, "Perceptive Locomotion Through Nonlinear Model Predictive Control," *IEEE Transactions on Robotics* 39, no. 5 (2023): 3402–3421, https://doi.org/10.1109/TRO.2023.3275384.

172. T. Corbères, M. Carlos, M. Wolfgang, et al., "Perceptive Locomotion Through Whole-Body MPC and Optimal Region Selection," *arXiv* (2023). abs/2305.08926, https://doi.org/10.48550/arXiv.2305.08926.

173. F. Jenelten, T. Miki, A. E. Vijayan, B. Marko, and H. Marco, "Perceptive Locomotion in Rough Terrain–Online Foothold Optimization," *IEEE Robotics and Automation Letters* 5, no. 4 (2020): 5370–5376, https://doi.org/10.1109/LRA.2020.3007427.

174. P. Fankhauser and M. Hutter, "A Universal Grid Map Library: Implementation and Use Case for Rough Terrain Navigation," in *Robot Operating System (ROS) the Complete Reference*, ed. A. Koubaa, Vol. 1 (Springer, 2016), 99–120. ch. 5. [Online], https://doi.org/10.1007/978-3-319-26054-9_5.

175. C. Mastalli, H. Ioannis, F. Michele, D. G. Caldwell, and C. Semini, "Motion Planning for Quadrupedal Locomotion: Coupled Planning, Terrain Mapping, and Whole-Body Control," *IEEE Transactions on Robotics* 36, no. 6 (2020): 1635–1648, https://doi.org/10.1109/TRO.2020.3003464.

176. F. Jenelten, J. He, F. Farbod, and H. Marco, "DTC: Deep Tracking Control," *Science Robotics* 9, no. 86 (2024): eadh5401, https://doi.org/10.1126/scirobotics.adh5401.

177. S. Kajita, F. Kanehiro, K. Kaneko, et al., "Biped Walking Pattern Generation by Using Preview Control of Zero-Moment Point," in *IEEE International Conference on Robotics and Automation* (2003), 1620–1626, https://doi.org/10.1109/ROBOT.2003.1241826.

178. J. Pratt, J. Carff, S. Drakunov, and A. Goswami, "Capture Point: A Step Toward Humanoid Push Recovery," in *IEEE-RAS International Conference on Humanoid Robots* (2006), 200–207, https://doi.org/10.1109/ICHR.2006.321385.

179. B. Dariush, M. Gienger, H. Janssen, M. Mistry, R. Tellez, and C. Goerick, "Online Task Space Control Under Balance Constraints for Compliant Humanoid Robots," in *IEEE International Conference on Robotics and Automation* (2008), 218–223.

180. K. Yamane, J. K. Hodgins, and S. Brown, "Controlling a Humanoid Robot via Simplified Model and Optimization," in *IEEE International Conference on Robotics and Automation* (2004), 2517–2524.

181. E. R. Westervelt, J. W. Grizzle, and D. E. Koditschek, "Hybrid Zero Dynamics of Planar Biped Walkers," *IEEE Transactions on Automatic Control* 48, no. 1 (2003): 42–56, https://doi.org/10.1109/tac.2002.806653.

182. A. D. Ames, J. W. Grizzle, and P. Tabuada, "Control of Planar Bipedal Walking Robots With Hybrid Zero Dynamics," *IEEE Conference on Decision and Control* 48, no. 1 (2006): 42–56. Jan. 2003, https://doi.org/10.1109/TAC.2002.806653.

183. J. W. Grizzle, C. Chevallereau, R. W. Sinnet, and A. D. Ames, "Models, Feedback Control, and Open Problems of 3D Bipedal Robotic Walking," *Automatica* 50, no. 8 (2014): 1955–1988, https://doi.org/10.1016/j.automatica.2014.04.021.

184. G. Romualdi, L. Baccelliere, M. Kamedula, and D. Pucci, "An Online Centroidal Model Predictive Control for Legged Locomotion," *IEEE Robotics and Automation Letters* 5, no. 2 (2020): 3573–3580.

185. H. Dai, A. Valenzuela, and T. Russ, "Whole-body Motion Planning With Centroidal Dynamics and Full Kinematics," *IEEE-RAS International Conference on Humanoid Robots* (2016): 295–302, https://doi.org/10.1109/HUMANOIDS.2014.7041375.

186. J. Di Carlo, P. M. Wensing, B. Katz, G. Bledt, and S. Kim, "Dynamic Locomotion in the MIT Cheetah 3 Through Model Predictive Control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2018), 1–9, https://doi.org/10.1109/IROS.2018.8594448.

187. S. Tak and H.-S. Ko, "A Physically-Based Motion Retargeting Filter," *ACM Transactions on Graphics* 24, no. 1 (2005): 98–117, https://doi.org/10.1145/1037957.1037963.

188. E. Lyard and M.-T. Nadia, "Motion Adaptation Based on Character Shape," *Computer Animation and Virtual Worlds* 19, no. 3–4 (2008): 189–198, https://doi.org/10.1002/cav.233.

189. J. Lee and S. Y. Shin, "A Hierarchical Approach to Interactive Motion Editing for Human-Like Figures," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (1999), 39–48, https://doi.org/10.1145/311535.311539.

190. M. Gleicher, "Retargetting Motion to New Characters," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (1998), 33–42, https://doi.org/10.1145/280814.280837.

191. Z. Popović and A. Witkin, "Physically Based Motion Transformation," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques* (1999), 11–20, https://doi.org/10.1145/311535.311536.

192. R. Rim, M. Marsot, A.-H. Olivier, J.-S. Franco, and S. Wuhrer, "Correspondence-Free Online Human Motion Retargeting," *arXiv Preprint* (2023). arXiv:2302.00556.

193. J. Zhang, J. Weng, D. Kang, et al., "Skinned Motion Retargeting With Residual Perception of Motion Semantics & Geometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 13864–13872, https://doi.org/10.1109/CVPR48813.2023.00138.

194. R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural Kinematic Networks for Unsupervised Motion Retargetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 8639–8648, https://doi.org/10.1109/CVPR.2018.00887.

195. K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, and B. Chen, "Skeleton-Aware Networks for Deep Motion Retargeting," *ACM Transactions on Graphics* 39, no. 4 (2020): 62:1–62:14, https://doi.org/10.1145/3386569.3392462.

196. D. Holden, J. Saito, and T. Komura, "A Deep Learning Framework for Character Motion Synthesis and Editing," *ACM Transactions on Graphics* 35, no. 4 (2016): 1–11, https://doi.org/10.1145/2897824.292597.

197. H. Jang, B. Kwon, M. Yu, S. U. Kim, and J. Kim, "A Variational U-Net for Motion Retargeting," in *SIGGRAPH Asia 2018 Posters* (2018), 1–2, https://doi.org/10.1145/3283289.3283316.

198. B. Delhaisse, D. Esteban, L. Rozo, and D. Caldwell, "Transfer Learning of Shared Latent Spaces Between Robots With Similar Kinematic Structure," in *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), 4142–4149, https://doi.org/10.1109/IJCNN.2017.7966379.

199. S. Choi, M. J. Song, H. Ahn, and J. Kim, "Self-Supervised Motion Retargeting With Safety Guarantee," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), 8097–8103, https://doi.org/10.1109/ICRA48506.2021.9560860.

200. X. B. Peng, Y. Guo, L. Halper, S. Levine, and F. Sanja, "Ase: Large-Scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters," *ACM Transactions on Graphics* 41, no. 4 (2022): 1–17, https://doi.org/10.1145/3528223.3530110.

201. X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial Motion Priors for Stylized Physics-Based Character Control," *ACM Transactions on Graphics* 40, no. 4 (2021): 1–20, https://doi.org/10.1145/3450626.3459670.

202. X. Xu, Y. Zhang, Y.-L. Li, L. Han, and C. Lu, "HumanVLA: Towards Vision-Language Directed Object Rearrangement by Physical Humanoid," *arXiv Preprint* (2024). arXiv:2406.19972.

203. M. Hassan, Y. Guo, T. Wang, M. Black, S. Fidler, and X. B. Peng, "Synthesizing Physical Character-Scene Interactions," in *ACM SIGGRAPH 2023 Conference Proceedings* (2023), 1–9, https://doi.org/10.1145/3588432.3591525.

204. A. Winkler, J. Won, and Y. Ye, "QuestSim: Human Motion Tracking From Sparse Sensors With Simulated Avatars," in *SIGGRAPH Asia 2022 Conference Papers* (2022), 1–8, https://doi.org/10.1145/3550469.3555411.

205. Z. Luo, J. Cao, J. Merel, et al., "Universal Humanoid Motion Representations for Physics-Based Control," in *The Twelfth International Conference on Learning Representations* (2023), https://openreview.net/forum?id=OrOd8PxOO2.

206. X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills," *ACM Transactions on Graphics* 37, no. 4 (2018): 1–14, https://doi.org/10.1145/3197517.3201311.

207. J. Won, D. Gopinath, and J. Hodgins, "A Scalable Approach to Control Diverse Behaviors for Physically Simulated Characters," *ACM Transactions on Graphics* 39, no. 4 (2020): 33–41, https://doi.org/10.1145/3386569.3392381.

208. Z. Luo, J. Cao, K. Kitani, and W. Xu, "Perpetual Humanoid Control for Real-Time Simulated Avatars," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 10895–10904, https://doi.org/10.1109/ICCV51070.2023.01000.

209. X. B. Peng, G. Berseth, K.K. Yin, and V. D. P. Michiel, "DeepLoco: Dynamic Locomotion Skills Using Hierarchical Deep Reinforcement Learning," *ACM Transactions on Graphics* 36, no. 4 (2017): 1–13, https://doi.org/10.1145/3072959.3073602.

210. T. Wang, Y. Guo, M. Shugrina, and F. Sanja, "Unicon: Universal Neural Controller for Physics-Based Character Motion," *arXiv Preprint* (2020). arXiv:2011.15119, https://arxiv.org/abs/2011.15119.

211. L. Fussell, K. Bergamin, and D. Holden, "Supertrack: Motion Tracking for Physically Simulated Characters Using Supervised Learning," *ACM Transactions on Graphics* 40, no. 6 (2021): 1–13, https://doi.org/10.1145/3478513.3480527.

212. Y. Yuan and K. Kitani, "Residual Force Control for Agile Human Behavior Imitation and Extended Motion Synthesis," *Advances in Neural Information Processing Systems* 33 (2020): 21763–21774, https://doi.org/10.5555/3495724.3497550.

213. X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive Whole-Body Control for Humanoid Robots," *arXiv Preprint* (2024). arXiv:2402.16796, https://arxiv.org/abs/2402.16796.

214. Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement Learning for Versatile, Dynamic, and Robust Bipedal Locomotion Control," *International Journal of Robotics Research* (2024): 02783649241285161, https://doi.org/10.1177/02783649241285161.

215. I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-World Humanoid Locomotion With Reinforcement Learning," *Science Robotics* 9, no. 89 (2024): eadi9579, https://doi.org/10.1126/scirobotics.adi9579.

216. T. He, Z. Luo, W. Xiao, et al., "Learning Human-To-Humanoid Real-Time Whole-Body Teleoperation," *arXiv Preprint* (2024). arXiv:2403.04436, https://arxiv.org/abs/2403.04436.

217. Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "HumanPlus: Humanoid Shadowing and Imitation From Humans," *arXiv Preprint* (2024). arXiv:2406.10454, https://arxiv.org/abs/2406.10454.

218. A. Suárez-Hernández, A. Guillem, and T. Carme, "Interleaving Hierarchical Task Planning and Motion Constraint Testing for Dual-Arm Manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), 4061–4066, https://doi.org/10.1109/IROS.2018.8593847.

219. N. T. Dantam, Z. K. Kingston, S. Chaudhuri, and L. E. Kavraki, "Incremental Task and Motion Planning: A Constraint-Based Approach," *Robotics: Science and Systems* 12 (2016): 00052, https://doi.org/10.15607/RSS.2016.XII.002.

220. C. R. Garrett, L.-P. Tomás, and K. Leslie Pack, "FFRob: Leveraging Symbolic Planning for Efficient Task and Motion Planning," *International Journal of Robotics Research* 37, no. 1 (2018): 104–136, https://doi.org/10.1177/0278364917739114.

221. C. Zhang and J. A. Shah, "Co-Optimizing Task and Motion Planning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2016), 4750–4756, https://doi.org/10.1109/IROS.2016.7759698.

222. M. Ahn, A. Brohan, N. Brown, et al., "Do As I Can, Not as I Say: Grounding Language in Robotic Affordances," in *Conference on Robot Learning* (2023), 287–318.

223. I. Singh, B. Valts, M. Arsalan, et al., "ProgPrompt: Generating Situated Robot Task Plans Using Large Language Models," in *IEEE International Conference on Robotics and Automation (ICRA)* (2023), 11523–11530, https://doi.org/10.1109/ICRA.2023.10161317.

224. Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI Tasks With ChatGPT and Its Friends in HuggingFace," *Journal of Artificial Intelligence Research* 65 (2023): 102–117, https://doi.org/10.48550/arXiv.2303.17580.

225. S. Yao, J. Zhao, D. Yu, et al., "ReAct: Synergizing Reasoning and Acting in Language Models," *Journal of Machine Learning Research* 22, no. 1 (2022): 212–236, https://doi.org/10.48550/arXiv.2210.03629.

226. W. Huang, F. Xia, T. Xiao, et al., "Inner Monologue: Embodied Reasoning Through Planning With Language Models," *Journal of Robotics and Automation* 42, no. 7 (2022): 345–367, https://doi.org/10.48550/arXiv.2207.05608.

227. N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: An Autonomous Agent With Dynamic Memory and Self-Reflection," *Journal of Autonomous Agents and Multi-Agent Systems* 31, no. 4 (2023): 233–247, https://doi.org/10.48550/arXiv.2303.11366.

228. Z. Liu, A. Bahety, and S. Song, "Reflect: Summarizing Robot Experiences for Failure Explanation and Correction," in *Conference on Robot Learning, PMLR* (2023), 3468–3484, https://doi.org/10.48550/arXiv.2306.15724.

229. C. R. Garrett, C. Rohan, H. Rachel, et al., "Integrated Task and Motion Planning," *Annual Review of Control, Robotics, and Autonomous Systems* 4, no. 1 (2021): 265–293, https://doi.org/10.1146/annurev-control-091420-084139.

230. M. Andrychowicz, B. Baker, M. Chociej, et al., "Learning Dexterous In-Hand Manipulation," *International Journal of Robotics Research* 39, no. 1 (2020): 3–20, https://doi.org/10.1177/0278364919887447.

231. OpenAI, I. Akkaya, M. Andrychowicz, et al., "Solving Rubik's Cube With a Robot Hand," *Journal of Machine Learning Research* 20, no. 81 (2019): 1–12, https://doi.org/10.48550/arXiv.1910.07113.

232. H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik, "In-Hand Object Rotation via Rapid Motor Adaptation (Version 1)," *arXiv* (2022), https://doi.org/10.48550/ARXIV.2210.04887.

233. H. Qi, B. Yi, S. Suresh, et al., "General In-Hand Object Rotation With Vision and Touch," *IEEE Transactions on Robotics* 39, no. 4 (2023): 1–12, https://doi.org/10.48550/arXiv.2309.09979.

234. T. Chen, J. Xu, and P. Agrawal, "A System for General In-Hand Object Re-Orientation," *Journal of Robotics and Automation* 36, no. 5 (2022): 455–467, https://doi.org/10.48550/arXiv.2111.03043.

235. T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal, "Visual Dexterity: In-Hand Reorientation of Novel and Complex Object Shapes," *Science Robotics* 8, no. 84 (2023): eadc9244, https://doi.org/10.1126/scirobotics.adc9244.

236. Y. Bai and C. Karen Liu, "Dexterous Manipulation Using Both Palm and Fingers," in *IEEE International Conference on Robotics and Automation* (2014), 1560–1565, https://doi.org/10.1109/ICRA.2014.6907059.

237. M. R. Dogar and S. S. Srinivasa, "Push-Grasping With Dexterous Hands: Mechanics and a Method," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2010), 2123–2130, https://doi.org/10.1109/IROS.2010.5652970.

238. Q. Liu, Y. Cui, Q. Ye, et al., "DexRepNet: Learning Dexterous Robotic Grasping Network With Geometric, Spatial Hand-Object Representations," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2023), 3153–3160, https://doi.org/10.1109/IROS55552.2023.10342334.

239. Y. Xu, W. Wan, J. Zhang, et al., "UniDexGrasp: Universal Robotic Dexterous Grasping via Learning Diverse Proposal Generation and Goal-Conditioned Policy," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 4737–4746, https://doi.org/10.1109/CVPR52729.2023.00459.

240. W. Wan, H. Geng, Y. Liu, et al., "UniDexGrasp++: Improving Dexterous Grasping Policy Learning via Geometry-Aware Curriculum and Iterative Generalist-Specialist Learning," in *IEEE/CVF International Conference on Computer Vision* (2023), 3868–3879, https://doi.org/10.1109/ICCV51070.2023.00360.

241. M. Alles and E. Aljalbout, "Learning to Centralize Dual-Arm Assembly," *Frontiers in Robotics* 9 (2022): 830007, https://doi.org/10.3389/frobt.2022.830007.

242. K. Takata, T. Kiyokawa, I. G. Ramirez-Alpizar, N. Yamanobe, W. Wan, and K. Harada, "Efficient Task/Movement Planning for a Dual-Arm Robot From Language Instructions and Cooking Images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2022), 12058–12065, https://doi.org/10.1109/IROS47612.2022.9981280.

243. L. Sun, G. Aragon-Camarasa, S. Rogers, and J. P. Siebert, "Autonomous Clothes Manipulation Using a Hierarchical Vision Architecture," *IEEE Access* 6 (2018): 76646–76662, https://doi.org/10.1109/ACCESS.2018.2883072.

244. W. Wan, K. Harada, and F. Kanehiro, "Preparatory Manipulation Planning Using Automatically Determined Single and Dual Arm," *IEEE Transactions on Industrial Informatics* 16, no. 1 (2019): 442–453, https://doi.org/10.1177/0278364919887447.

245. C. Mitash, R. Shome, B. Wen, A. Boularias, and K. Bekris, "Task-Driven Perception, Manipulation for Constrained Placement of Unknown Objects," *IEEE Robotics and Automation Letters* 5, no. 4 (2020): 5605–5612, https://doi.org/10.1109/LRA.2020.3006816.

246. W. Wang, C. Zeng, Z. Lu, and C. Yang, "A Novel Robust Imitation Learning Framework for Dual-Arm Object-Moving Tasks," *IEEE Transactions on Industrial Electronics* 71, no. 12 (2024): 16068–16076, https://doi.org/10.1109/TIE.2024.3387098.

247. Y. Ren, Z. Zhou, Z. Xu, et al., "Enabling Versatility and Dexterity of the Dual-Arm Manipulators: A General Framework Toward Universal Cooperative Manipulation," *IEEE Transactions on Robotics* 40 (2024): 2024–2045, https://doi.org/10.1109/TRO.2024.3370048.

248. C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition* (2017), 652–660, https://doi.org/10.1109/CVPR.2017.16.

249. W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "VoxPoser: Composable 3D Value Maps for Robotic Manipulation With Language Models," in *Conference on Robot Learning* (2023), https://doi.org/10.48550/arXiv.2307.05973.

250. F. Liu, K. Fang, P. Abbeel, and S. Levine, "MOKA: Open-Vocabulary Robotic Manipulation Through Mark-Based Visual Prompting," in *Proceedings of Robotics: Science and Systems* (2024), https://doi.org/10.15607/rss.2024.xx.062.

251. H. Huang, F. Lin, Y. Hu, S. Wang, and Y. Gao, "CoPa: General Robotic Manipulation Through Spatial Constraints of Parts With Foundation Models," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024* (2024), https://doi.org/10.48550/arXiv.2403.08248.

252. W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation," in *Conference on Robot Learning* (2024), https://doi.org/10.48550/arXiv.2409.01652.

253. A. Brohan, N. Brown, J. Carbajal, et al., "RT-1: Robotics Transformer for Real-World Control at Scale," in *Proceedings of Robotics: Science and Systems* (2022), https://doi.org/10.48550/arXiv.2212.06817.

254. B. Zitkovich, T. Yu, S. Xu, et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in *Conference on Robot Learning* (2023), https://doi.org/10.48550/arXiv.2307.15818.

255. X. Li, M. Liu, H. Zhang, et al., "Vision-Language Foundation Models as Effective Robot Imitators," *arXiv* (2023). arXiv:2311.01378, https://doi.org/10.48550/arXiv.2311.01378.

256. O. Mees, D. Ghosh, K. Pertsch, et al., "Octo: An Open-Source Generalist Robot Policy," in *Proceedings of Robotics: Science and Systems* (2024), https://doi.org/10.15607/rss.2024.xx.090.

257. Z. Xian, N. Gkanatsios, T. Gervet, T. W. Ke, and K. Fragkiadaki, "ChainedDiffuser: Unifying Trajectory Diffusion and Keypose Prediction for Robotic Manipulation," in *Conference on Robot Learning* (2023), https://proceedings.mlr.press/v229/xian23a/xian23a.pdf.

258. G. Theophile, X. Zhou, G. Nikolaos, and F. Katerina, "Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation," in *The 7th Annual Conference on Robot Learning*, Vol. 229 (2023), https://doi.org/10.48550/arXiv.2306.17817.

259. A. Goyal, J. Xu, Y. Guo, V. Blukis, Yu-W. Chao, and D. Fox, "RVT: Robotic View Transformer for 3D Object Manipulation," in *The 7th*

Annual Conference on Robot Learning, Vol. 229 (2023), 694–710, https://doi.org/10.48550/arXiv.2306.14896.

260. Y. Ze, G. Yan, Y.-H. Wu, et al., “GNFactor: Multi-Task Real Robot Learning With Generalizable Neural Feature Fields,” in *The 7th Annual Conference on Robot Learning*, Vol. 229 (2023), 284–301, https://doi.org/10.48550/arXiv.2308.16891.

261. T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, “3D Diffuser Actor: Policy Diffusion With 3D Scene Representations,” *arXiv* (2024). arXiv 2402.10885, https://doi.org/10.48550/arXiv.2402.10885.

262. G. Yan, Y.-H. Wu, and X. Wang, “DNAct: Diffusion Guided Multi-Task 3D Policy Learning,” *arXiv* (2024). arXiv:2403.04115, https://doi.org/10.48550/arXiv.2403.04115.

263. Z. Haoyu, Q. Xiaowen, C. Peihao, et al., “3D-VLA: A 3D Vision-Language-Action Generative World Model,” *arXiv* (2024). arXiv:2403.09631, https://doi.org/10.48550/arXiv.2403.09631.

264. G. Ankit, B. Valts, X. Jie, G. Yijie, C. Yu-Wei, and D. Fox, “RVT-2: Learning Precise Manipulation From Few Demonstrations,” *arXiv* (2024). arXiv:2406.08545, https://doi.org/10.48550/arXiv.2406.08545.

265. M. Shridhar, L. Manuelli, and D. Fox, “Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation,” in *Conference on Robot Learning* (2023), https://doi.org/10.48550/arXiv.2209.05451.

266. O. Abby, A. Neill, A. Rehman, et al., “Open X-Embodiment: Robotic Learning Datasets and RT-X Models,” *arXiv* (2023). arXiv: 2310.08864, https://doi.org/10.48550/arXiv.2310.08864.

267. K. Alexander, P. Karl, N. Suraj, et al., “DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset,” *arXiv* (2024). arXiv:2403.12945, https://doi.org/10.15607/arXiv:2403.12945.

268. M. Ajay, N. Soroush, W. Bowen, et al., “MimicGen: A Data Generation System for Scalable Robot Learning Using Human Demonstrations,” in *The 7th Annual Conference on Robot Learning*, Vol. 229 (2023), https://doi.org/10.48550/arXiv.2310.17596.

269. H. Ha, P. Florence, and S. Song, “Scaling up and Distilling Down: Language-Guided Robot Skill Acquisition,” in *The 7th Annual Conference on Robot Learning*, Vol. 229 (2023), 3766–3777, https://doi.org/10.48550/arXiv.2307.14535.

270. D. Jiafei, Y. Wentao, P. Wilbert, et al., “Manipulate-Anything: Automating Real-World Robots Using Vision-Language Models,” in *Conference on Robot Learning* (2024), https://doi.org/10.48550/arXiv.2406.18915.

271. L. Songming, W. Lingxuan, L. Bangguo, et al., “RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation,” *arXiv* (2024). arXiv:2410.07864, https://doi.org/10.48550/arXiv.2410.07864.

272. B. Kevin, B. Noah, D. Danny, et al., “π_0: A Vision-Language-Action Flow Model for General Robot Control,” *arXiv* (2024). arXiv:2410.24164, https://doi.org/10.48550/arXiv.2410.24164.

273. F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, “Data Scaling Laws in Imitation Learning for Robotic Manipulation,” *arXiv* (2024). arXiv:2410.18647, https://doi.org/10.48550/arXiv.2410.18647.

274. G. Abhishek, K. Vikash, L. Corey, L. Sergey, and H. Karol, “Relay Policy Learning: Solving Long-Horizon Tasks via Imitation and Reinforcement Learning,” *arXiv* (2019). arXiv:1910.11956, https://doi.org/10.48550/arXiv.1910.11956.

275. M. Heo, Y. Lee, D. Lee, and J. J. Lim, “FurnitureBench: Reproducible Real-World Benchmark for Long-Horizon Complex Manipulation,” *arXiv* (2023). arXiv:2305.12821, https://doi.org/10.48550/arXiv.2305.12821.

276. Y. Lee, E. S. Hu, and J. J. Lim, “IKEA Furniture Assembly Environment for Long-Horizon Complex Manipulation Tasks,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), 6343–6349, https://doi.org/10.1109/ICRA48506.2021.9562002.

277. A. Mandlekar, D. Xu, R. Martín-Martín, et al., “Learning to Generalize Across Long-Horizon Tasks From Human Demonstrations,” *arXiv Preprint* (2020). arXiv:2003.06085, accessed November 21, 2024, https://arxiv.org/abs/2003.06085.

278. E. Jang, A. Irpan, M. Khansari, et al., “BC-Z: Zero-Shot Task Generalization With Robotic Imitation Learning,” in *Conference on Robot Learning, PMLR* (2022), 991–1002, accessed November 21, 2024, https://proceedings.mlr.press/v164/jang22a.html.

279. F. Ni, J. Hao, Y. Mu, et al., “MetaDiffuser: Diffusion Model as Conditional Planner for Offline Meta-RL,” in *International Conference on Machine Learning, PMLR* (2023), 26087–26105, accessed November 21, 2024, https://proceedings.mlr.press/v202/ni23a.html.

280. L. P. Kaelbling and T. Lozano-Pérez, “Hierarchical Task and Motion Planning in the Now,” in *2011 IEEE International Conference on Robotics and Automation (ICRA)* (2011), 1470–1477, https://doi.org/10.1109/ICRA.2011.5980305.

281. T. G. Dietterich, “The MAXQ Method for Hierarchical Reinforcement Learning,” in *ICML* (1998), 118–126, https://icml.cc/Conferences/1998/proceedings/papers/098.pdf.

282. P.-L. Bacon, J. Harb, and D. Precup, “The Option-Critic Architecture,” *Proceedings of the AAAI Conference on Artificial Intelligence* 31, no. 1 (2017), https://doi.org/10.1609/aaai.v31i1.10916.

283. A. Bagaria and G. Konidaris, “Option Discovery Using Deep Skill Chaining,” in *International Conference on Learning Representations* (2019), accessed November 21, 2024, https://openreview.net/forum?id=H1g0Z3R5KQ.

284. R. S. Sutton, D. Precup, and S. Singh, “Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning,” *Artificial Intelligence* 112, no. 1–2 (1999): 181–211, https://doi.org/10.1016/S0004-3702(99)00052-1.

285. G. Konidaris and A. Barto, “Skill Discovery in Continuous Reinforcement Learning Domains Using Skill Chaining,” *Advances in Neural Information Processing Systems* 22 (2009), https://proceedings.neurips.cc/paper_files/paper/2009/file/338d4c64e05c9ba808de3854d0c3e3dd-Paper.pdf.

286. Y. Lee, J. J. Lim, A. Anandkumar, et al., “Adversarial Skill Chaining for Long-Horizon Robot Manipulation via Terminal State Regularization,” *arXiv* (2021). abs/2111.07999, https://doi.org/10.48550/arXiv.2111.07999.

287. Y. Lee, S. H. Sun, S. Somasundaram, et al., “Composing Complex Skills by Learning Transition Policies,” in *Proceedings of the International Conference on Learning Representations* (2019).

288. C. Agia, T. Migimatsu, J. Wu, et al., “Taps: Task-Agnostic Policy Sequencing,” *arXiv* (2022). abs/2210.12250 2022, https://doi.org/10.48550/arXiv.2210.12250.

289. G. Konidaris, S. Kuindersma, R. Grupen, and A. Barto, “Robot Learning From Demonstration by Constructing Skill Trees,” *International Journal of Robotics Research* 31, no. 3 (2012): 360–375, https://doi.org/10.1177/0278364911428653.

290. A. Clegg, W. Yu, J. Tan, C. K. Liu, and G. Turk, “Learning to Dress: Synthesizing Human Dressing Motion via Deep Reinforcement Learning,” *ACM Transactions on Graphics* 37, no. 6 (2018): 1–10, https://doi.org/10.1145/3272127.3275048.

291. Y. Zhu, P. Stone, and Y. Zhu, “Bottom-Up Skill Discovery From Unsegmented Demonstrations for Long-Horizon Robot Manipulation,” *IEEE Robotics and Automation Letters* 7, no. 2 (2022): 4126–4133, https://doi.org/10.1109/LRA.2022.3150115.

292. D.-A. Huang, S. Nair, D. Xu, et al., “Neural Task Graphs: Generalizing to Unseen Tasks From a Single Video Demonstration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 8565–8574, https://doi.org/10.1109/CVPR.2019.00877.

293. D. Xu, S. Nair, Y. Zhu, et al., "Neural Task Programming: Learning to Generalize Across Hierarchical Tasks," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), 3795–3802, https://doi.org/10.1109/ICRA.2018.8460871.

294. S. Bahl, A. Gupta, and D. Pathak, "Human-to-Robot Imitation in the Wild," *arXiv Preprint* (2022). arXiv:2207.09450, accessed November 21, 2024, https://arxiv.org/abs/2207.09450.

295. A. Mandlekar, F. Ramos, B. Boots, et al., "IRIS: Implicit Reinforcement Without Interaction at Scale for Learning Control From Offline Robot Manipulation Data," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), 4414–4420, https://doi.org/10.1109/ICRA40945.2020.9196906.

296. K. Pertsch, R. Desai, V. Kumar, et al., "Cross-Domain Transfer via Semantic Skill Imitation," *arXiv Preprint* (2022). arXiv:2212.07407, accessed November 21, 2024, https://arxiv.org/abs/2212.07407.

297. A. Tung, J. Wong, A. Mandlekar, et al., "Learning Multi-Arm Manipulation Through Collaborative Teleoperation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), 9212–9219, https://doi.org/10.1109/ICRA48506.2021.9561133.

298. C. Wang, L. Fan, J. Sun, et al., "MimicPlay: Long-Horizon Imitation Learning by Watching Human Play," *arXiv* (2023). abs/2302.12422, https://doi.org/10.48550/arXiv.2302.12422.

299. Y. Duan, M. Andrychowicz, B. Stadie, et al., "One-Shot Imitation Learning," *Advances in Neural Information Processing Systems* 30 (2017).

300. L. Smith, N. Dhawan, M. Zhang, et al., "Avid: Learning Multi-Stage Tasks via Pixel-Level Translation of Human Videos," *arXiv* (2019). abs/1912.04443, https://doi.org/10.48550/arXiv.1912.04443.

301. T. Yu, P. Abbeel, S. Levine, et al., "One-Shot Hierarchical Imitation Learning of Compound Visuomotor Tasks," *arXiv* (2018). abs/1810.11043, https://doi.org/10.48550/arXiv.1810.11043.

302. P. Wu, K. Hakhamaneshi, Y. Du, et al., "Semi-Supervised One-Shot Imitation Learning," *arXiv* (2024). abs/2408.05285, https://doi.org/10.48550/arXiv.2408.05285.

303. S. Wu, Y. Wang, and Y. Huang, "One-Shot Robust Imitation Learning for Long-Horizon Visuomotor Tasks From Unsegmented Demonstrations," *arXiv* (2024). abs/2410.01630, https://doi.org/10.48550/arXiv.2410.01630.

304. O. Stasse, F. Morsillo, M. Geisert, N. Mansard, M. Naveau, and C. Vassallo, "Airbus/future of Aircraft Factory HRP-2 as Universal Worker Proof of Concept," in *2014 IEEE-RAS International Conference on Humanoid Robots* (2014), 1014–1015, https://doi.org/10.1109/humanoids.2014.7041488.

305. Tesla, "AI & Robotics," https://www.tesla.com/en_eu/AI.

306. SoftBank Robotics America, "Meet Pepper: The Robot Built for People," https://us.softbankrobotics.com/pepper.

307. PROMOBOT, "Robot Tour Guide," https://promo-bot.ai/use-case/tour-guide.

308. P. Stone, R. Brooks, E. Brynjolfsson, et al., "Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence (Version 1)," *arXiv* (2022), https://doi.org/10.48550/ARXIV.2211.06318.

309. D. Lahr, V. Orekhov, B. Lee, and D. Hong, "Early Developments of a Parallelly Actuated Humanoid, SAFFiR," in *Proceedings of the International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 55942 (American Society of Mechanical Engineers, 2013), https://doi.org/10.1115/DETC2013-12590.

310. NASA, "Valkyrie-NASA," https://www.nasa.gov/podcasts/houston-we-have-a-podcast/valkyrie/.