Review | Received 21 August 2026; Revised 15 October 2026; Accepted 30 October 2026; Published 13 November 2026
https://doi.org/10.55092/rl20260027

A systematic review of embodied intelligence: concepts, technologies, applications, evaluation and challenges

Yixiang Liang[1], He Ma[1], Yong Chen[1], Xuan Huang[1], Bichun Li[1], Linglong Lin[1], Lei Sun[1], Shengchao Zhen[2], Shuai Zhang[3], Biao Yu[1,*] and Chilai Chen[1,*]

[1] Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China
[2] School of Mechanical Engineering, Hefei University of Technology, Hefei, China
[3] UCL Hawkes Institute and the Department of Computer Science, University College London, London, UK

* Correspondence authors; E-mails: byu@hfcas.ac.cn (B.Y.); chlchen@iim.ac.cn (C.C.).

Highlights:

- Clarifying the conceptual boundaries between embodied intelligence, teleoperation, and automation, as well as the relationships among disembodied intelligence, spatial intelligence, physical intelligence, and embodied intelligence.
- Providing a systematic review of the key technological components of embodied intelligence, including embodiment, cerebellum, brain architectures, data acquisition, and training methodologies.
- Connecting technological advances with real-world deployment requirements through representative application scenarios and evaluation frameworks involving data-based, simulation-based and real-world testing.
- Identifying key challenges to the deployment of embodied intelligence systems in real-world scenarios, particularly for complex and precise manipulation tasks.

Abstract: Embodied intelligence extends artificial intelligence from information processing in the digital domain to autonomous interaction with the physical world. Despite rapid progress, the field still faces challenges such as ambiguous conceptual boundaries, fragmented technical frameworks, and insufficient alignment between application requirements and evaluation standards. This review presents a systematic framework for understanding embodied intelligence from five perspectives: conceptual foundations, key technologies, representative applications, evaluation methodologies, and future challenges. We first distinguish embodied intelligence from teleoperation and conventional automation, and establish an evolutionary framework spanning disembodied intelligence, spatial intelligence, physical intelligence, and embodied intelligence. Then we summarize key technologies in embodiment, the cerebellum, brain architectures, data acquisition, and training approaches, with particular emphasis on the evolution from traditional hierarchical modular architecture to large-model-driven modular architecture, Vision-Language-Action (VLA) model architecture, world model architecture, and world action model architecture. Representative applications are categorized into structured-environment

Copyright©2026 by the authors. Published by ELSP. This work is licensed under Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

manipulation, unstructured-environment interaction, and high-risk open-domain exploration. To bridge technological development and real-world deployment, we further review three evaluation paradigms which include data-based evaluation, simulation-based testing and real-world testing. Finally, we identify eight major challenges which cover data quality, long-horizon planning, multimodal fusion, edge deployment, safety and trustworthiness, continual learning, unified navigation and manipulation, and multi-agent collaboration. This review provides a systematic perspective on embodied intelligence and highlights key directions for advancing embodied systems toward reliable deployment in complex real-world environments.

Keywords: embodied intelligence; data acquisition; VLA; world model; world action model; evaluation and testing; future challenges

1. Introduction

Embodied intelligence refers to a paradigm in which an intelligent agent equipped with a physical body forms a closed-loop process of perception, prediction, decision-making, and action under environmental constraints, continuously learning and adapting through interactions with the physical world. Unlike disembodied intelligence, which primarily processes symbols, images, or text, embodied intelligence does not merely aim to understand the world; rather, it enables intelligent agents to enter, interact with and modify the physical world through their bodies. Therefore, understanding the scientific significance of embodied intelligence requires first revisiting the evolution process of artificial intelligence from digital information processing toward real-world interaction.

Historically, artificial intelligence has undergone three major stages: computational intelligence, perceptual intelligence, and embodied intelligence. These stages should not be viewed as a simple chronological progression, but rather as an expansion of intelligent capabilities. Computational intelligence focuses on symbolic reasoning, perceptual intelligence emphasizes pattern recognition from large-scale data, while embodied intelligence further requires intelligent agents to establish closed-loop interactions involving perception, cognition, action, and feedback in open environments.

The early stage of computational intelligence was dominated by symbolic approaches and logical reasoning. Allen Newell and Herbert Simon [1] proposed the Symbol System Hypothesis, arguing that intelligent behavior could be achieved through symbolic representations and rule-based operations. This symbolic paradigm attempted to solve mathematical theorem proving and complex reasoning problems using formal logic and heuristic search. John McCarthy [2] and other researchers further believed that human cognitive processes could eventually be formalized and implemented through computational systems. During this period, artificial intelligence achieved considerable success in well-defined abstract domains.

However, symbolic methods encountered fundamental limitations when dealing with open-ended perceptual tasks, leading artificial intelligence into the era of perceptual intelligence dominated by connectionism. Geoffrey Hinton and others [3] continuously explored neural networks and backpropagation algorithms from the 1980s to the early 2000s. In 2012, Alex Krizhevsky *et al.* [4] demonstrated the remarkable capability of deep convolutional neural networks in the ImageNet Large Scale Visual Recognition Challenge, significantly reducing image recognition errors and accelerating the adoption of deep learning in computer vision. Subsequently, deep learning achieved substantial breakthroughs in speech recognition and natural language processing, gradually replacing traditional rule-based approaches

in multimodal data analysis involving images, speech, and text. In 2017, Vaswani *et al.* [5] introduced the Transformer architecture, whose self-attention mechanism replaced recurrent structures and greatly improved sequence modeling performance, becoming the fundamental architecture underlying large language models and multimodal foundation models.

Nevertheless, the rapid advancement of perception models and large language models has revealed a critical gap, although these models can generate sophisticated responses in digital environments, they still struggle to reliably perform long-horizon, dynamic, and safety-critical tasks in physical environments. Moravec's Paradox [6] highlights the underlying reason for this limitation: sensorimotor abilities that appear intuitive for humans are often much more difficult to engineer than abstract reasoning capabilities. Embodied intelligence has emerged in response to this challenge, aiming to couple cognitive capabilities, physical bodies, and environmental feedback into a unified intelligent system. The evolution of artificial intelligence technologies is illustrated in Figure 1.

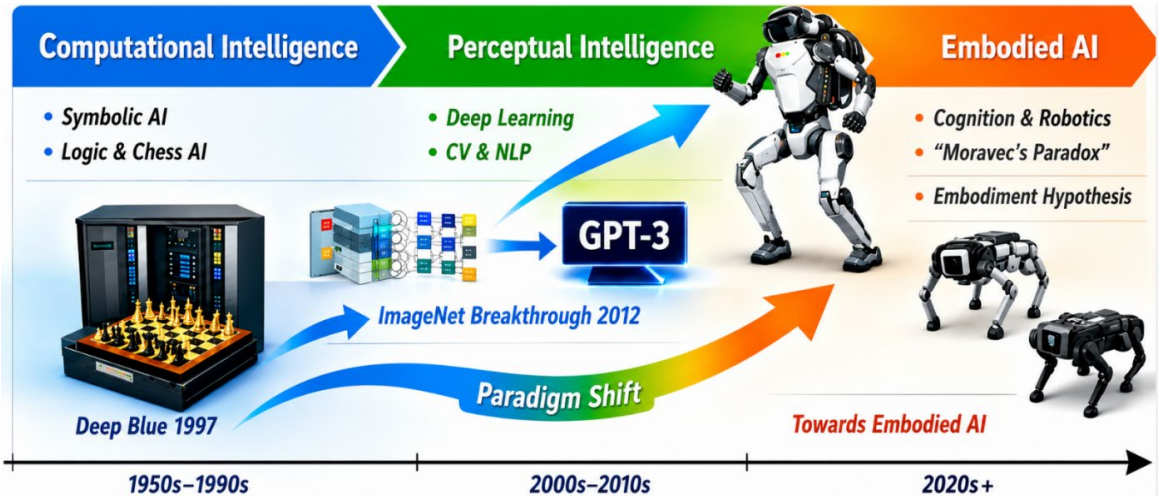


Figure 1. The evolution process of AI technologies. The conceptual figure was created with the assistance of ChatGPT and subsequently reviewed, edited, and refined by the authors.

Existing surveys on embodied intelligence [7−11] provided important foundations for research on VLA models, world models, robotic manipulation, and safety-aware control. However, several limitations remain. First, many studies focus on individual technical directions while providing insufficient discussion of the evolutionary relationships among disembodied intelligence, spatial intelligence, physical intelligence, and embodied intelligence, which may lead to conceptual confusion. Second, embodiment, motion control, cognitive decision-making, data acquisition, and training methodologies are often discussed separately, lacking a unified full-stack perspective that connects hardware embodiment, motor control, cognitive reasoning, and data-driven learning. Third, the relationship among application scenarios, evaluation benchmarks, and technical challenges remains unclear, making it difficult to answer critical questions regarding how capabilities should be validated and how deployment bottlenecks can be overcome in real-world environments. Table 1 summarizes the main content of recent related review articles and provides a comparison with other reviews.

Table 1. Comparison of related research surveys.

literature	Comparison of intelligence	Cerebellum	VLA	WM	WAM	Data Collection	Model Training	Application	Evaluation Benchmarks
[7]	×	×	✓	✓	×	×	✓	✓	×
[8]	×	×	✓	✓	×	✓	×	✓	✓
[9]	×	×	✓	✓	×	×	✓	✓	×
[10]	×	×	✓	×	×	×	✓	✓	×
[11]	×	×	×	×	×	✓	✓	✓	×
This paper	✓	✓	✓	✓	✓	✓	✓	✓	✓

To address these issues, this review establishes a unified analytical framework covering conceptual boundaries, key technologies, representative applications, and evaluation pipelines. The main contributions of this review are summarized as follows:

(1) We propose an evolutionary perspective of embodied intelligence by positioning disembodied intelligence, spatial intelligence, physical intelligence, and embodied intelligence within a unified developmental trajectory, while clarifying the fundamental differences between embodied intelligence, teleoperation, and automation.

(2) We construct a full-stack technological framework spanning embodiment, cerebellum, brain, data acquisition, and training, integrating hardware platforms, motion control, cognitive decision-making, data collection, and model learning into a unified architecture.

(3) We systematically review the evolution of embodied intelligence brain architectures, ranging from hierarchical modular systems to world action models, highlighting the transition toward predictive and action-coupled autonomous control.

(4) We connect technological progress with real-world deployment requirements through representative applications and evaluation benchmarks, and further summarize future challenges.

To ensure broad and up-to-date coverage, relevant literature was identified primarily through Google Scholar and arXiv based on the research topics and technical contents discussed in each section. Recent studies published since 2022 were prioritized, particularly for rapidly developing areas such as embodied foundation models, vision-language-action models, world models, robot learning, and embodied applications. Earlier publications were retained when they provided fundamental definitions, pioneering methods, or essential conceptual foundations. The retrieved literature was screened based on its relevance to the scope and technical focus of this review, while duplicate and unrelated studies were excluded.

2. The concept of embodied intelligence

2.1. Perspectives on concept of embodied intelligence from different disciplines

From a philosophical perspective, embodied intelligence emphasizes that intelligence does not exist as an isolated process of abstract reasoning or symbolic computation detached from the body [12]. Instead, intelligence emerges from the interaction between an organism's physical body and its surrounding environment through continuous perception and action. From this viewpoint, the body is not merely a carrier of intelligence but an essential foundation for the emergence of cognition.

From the perspective of computer science, embodied intelligence represents an artificial intelligence paradigm oriented toward the physical world. Its core objective is not simply to equip intelligent systems with additional sensors and actuators, but to enable them to establish a learnable closed-loop interaction

mechanism among bodily constraints, environmental feedback, and task objectives. Through continuous interaction with the environment, embodied agents can progressively improve their behavioral capabilities and achieve adaptation in complex and unstructured scenarios [13]. Cognitive science provides a bridge between philosophical theories of embodiment and computational models. The theory of embodied cognition demonstrates through psychological and neuroscience studies that bodily states and physical experiences directly influence higher-level cognitive processes. It further suggests that advanced cognitive abilities, including memory, emotion, and reasoning, are fundamentally grounded in the interaction between the body and the environment [14].

Overall, the central idea of embodied intelligence is that genuine intelligence cannot exist independently of the body, physical interaction, and environmental context. Intelligent agents must perceive the world through a physical body and generate actions through continuous perception-action loops, allowing cognition to emerge through interaction. Embodied intelligence therefore represents an evolutionary pathway of intelligence: philosophy establishes the fundamental role of embodiment, cognitive science provides empirical evidence for body-dependent cognition, and computer science ultimately aims to construct intelligent systems capable of real-time interaction and learning in physical environments.

2.2. Comparison of teleoperation, automation and embodied intelligence

By comparing embodied intelligence with traditional teleoperation and automation, the conceptual boundaries and intelligent characteristics of these technologies can be more clearly distinguished from the perspectives of autonomy level, physical interaction mode, and environmental adaptation mechanism.

Teleoperation refers to a system in which human operators transmit control commands to remote agents through dedicated control devices across spatial distances, directly controlling their actions, environmental interactions, and task execution. The operation process of teleoperation follows the following loop: remote environmental information is transmitted to the operator terminal through multimodal sensors [15]; the operator generates commands based on the returned information; the commands are transmitted to the agent to perform corresponding actions; and the execution results are fed back through sensors to support dynamic adjustment of commands. In this process, human decision-making runs through the entire workflow. The intelligent agent does not participate in autonomous decision-making, and all actions strictly follow human instructions, ensuring high precision and controllability.

Automation refers to the use of programming or physical constraints to convert specific task execution processes into predefined logic, allowing devices to complete repetitive production or daily tasks according to programmed rules. The autonomy of automation is mainly limited to the predefined rule boundaries. Once environmental conditions exceed the predefined range, human intervention is still required. A typical automated system operates through a closed-loop process consisting of rule definition, condition matching, autonomous execution, and result feedback. For example, engineers can define action sequences and triggering conditions through code, PLC configuration, or mechanical constraints [16]. The system collects environmental and internal state information, matches them with predefined rules, and executes corresponding actions. When abnormal situations occur, such as sensor failures or action timeouts, the system triggers alarms and waits for human inspection. Therefore, the core of automation is the stable reproduction of predefined procedures rather than understanding task semantics or autonomous learning in unknown environments. Taking current humanoid robots performing dancing

or boxing as an example, their behaviors are essentially high-precision trajectory tracking and dynamic compensation systems, which reproduce predefined joint angle sequences rather than demonstrating open-world understanding capabilities. Neither teleoperation nor traditional automation possesses intrinsic cognitive, reasoning, or learning capabilities. They only accurately reproduce assigned commands or trajectories without understanding task meanings or handling unexpected environmental changes.

Embodied intelligence, in contrast, targets complex, open, and unstructured physical environments. It emphasizes that intelligent agents can autonomously perceive the environment, understand task semantics, reason about action consequences, generate control strategies, and continuously adjust behaviors through feedback. Rather than simply reproducing instructions or trajectories, embodied intelligence extracts semantic and physical properties from high-dimensional sensory streams through multimodal perception, learnable world models, reinforcement learning, and foundation models. It can perform online task planning and motion generation and adjust strategies in previously unseen scenarios. The comparison among the three paradigms is shown in Table 2.

Table 2. Comparison of teleoperation, automation and embodied intelligence.

Comparison	Basic Definition	Core Objective	Autonomy	Environmental Interaction Mode	Task Structure Level	Learning and Generalization Ability
Teleoperation	Remote control of a distant device through communication links	Enable humans to observe or operate remote or dangerous environments	No autonomy	Humans interact with the environment indirectly through machines	Mainly handles unstructured environments but highly depends on human real-time judgment and flexibility	No learning capability or low learning capability
Automation	Enables machines to automatically complete specific tasks according to predefined programs	Replace humans in repetitive and high-precision tasks	Autonomous within predefined ranges	Interaction follows fixed logic	Highly dependent on structured, repeatable environments and sensitive to dynamic changes	No learning capability or low learning capability
Embodied Intelligence	Enables intelligent agents with physical bodies to interact and learn in real environments	Autonomously complete complex and changing tasks in unstructured environments	Highly autonomous with adaptive learning and generalization ability	Active perception-cognition-action loop through physical interaction with environments	Designed for unstructured, dynamic, and uncertain real-world scenarios	Continuously learns from experience and transfers skills to new tasks and environments

It should be noted that teleoperation, automation, and embodied intelligence are not mutually exclusive categories but represent different degrees and configurations of human involvement, autonomy, and learning. Teleoperation primarily emphasizes human control of the robot, but may incorporate shared autonomy or learned assistance. Automation generally emphasizes the execution of predefined or algorithmically generated procedures, while modern automated systems can also incorporate learning-based perception, planning, and control. Embodied intelligence places greater emphasis on autonomous perception, decision-making, interaction, and adaptation through the physical body, but human intervention can still be retained when appropriate. Therefore, these concepts are better understood as overlapping system configurations rather than strictly separated categories.

2.3. Comparison of disembodied intelligence, spatial intelligence, physical intelligence and embodied intelligence

With the development of foundation models, multimodal perception, and robotics technologies, concepts such as disembodied intelligence [17] , spatial intelligence [18], physical intelligence [19] and embodied intelligence [20] have been frequently used to describe different levels of intelligent capabilities. These concepts respectively emphasize symbolic and semantic reasoning, three-dimensional spatial representation,

physical law modeling, and physical interaction. This section compares these concepts from four dimensions: cognitive foundation, spatial representation, physical constraints, and closed-loop interaction.

Disembodied intelligence is an intelligence paradigm centered on cognitive computation without physical entities. It is based on multimodal foundation models such as text, vision, and speech models, which perform advanced cognitive activities including perception understanding, logical reasoning, content generation, and task planning in digital spaces. The operation process of disembodied intelligence follows a loop consisting of multimodal input, unified representation, cognitive reasoning, and digital output. Users provide information in the forms of text, images, speech, and videos. The model integrates multimodal information through vision encoders, language encoders, and Transformer architectures to achieve intent understanding. Based on large-scale pretrained knowledge and in-context learning capabilities, the model performs reasoning, planning, and generation. The outputs are provided in digital forms through interactive interfaces. Although the model can continuously improve through fine-tuning and alignment learning, the entire process does not involve physical perception, action execution, or real-world feedback. Disembodied intelligence is widely applied in digital information processing scenarios such as text generation [21] and video generation [22].

Spatial intelligence is an intelligence paradigm centered on spatial understanding and 3D scenario cognition [18]. Its essence is to establish structured representations of the real three-dimensional world based on two-dimensional perception, enabling intelligent systems to understand object positions, scales, orientations, topological relationships, and scene layouts. It represents the transition of AI from traditional 2D visual recognition toward 3D world modeling and interaction capabilities, and serves as an important foundation for AI entering physical environments. Compared with general visual semantic understanding, spatial intelligence places greater emphasis on geometric structures, accessibility, and object relationships, providing fundamental representations for navigation, manipulation, and world models.

Physical intelligence is a key paradigm for AI transitioning from information understanding to understanding the laws of the physical world [19]. Its core objective is to enable intelligent systems to model, predict, and make decisions based on real-world physical processes. Compared with disembodied intelligence, which mainly processes symbols, language, and statistical patterns, physical intelligence focuses more on physical laws such as gravity, friction, inertia, contact, material properties, and causal dynamics. Therefore, physical intelligence can be regarded as a capability for modeling and predicting physical laws. It does not necessarily require a physical entity and can often be validated through simulation environments. In contrast, embodied intelligence requires a complete closed-loop system consisting of a physical body, environmental interaction, and real-time feedback.

Embodied intelligence refers to an intelligence paradigm based on physical bodies for perception and action. Its core characteristics include three inseparable dimensions: physicality, interaction, and adaptability [20]. Physicality means that intelligence relies on physical entities such as sensors and actuators and is constrained by physical laws including gravity, dynamics, and friction. Interaction means that intelligence emerges from active interaction between agents and the real environment rather than static data processing. Adaptability means that systems can dynamically adjust behaviors and autonomously complete tasks in unstructured environments. In summary, embodied intelligence embeds cognition into a physical body and achieves a paradigm shift from data processing to physical interaction and autonomous operation through real-time interaction with the environment. Embodied intelligence operates through a closed loop in which the system autonomously perceives the environment,

interacts with it in real time, makes cognitive decisions, executes actions, and adaptively iterates based on feedback. Through visual, force, tactile, and inertial sensors, the system collects information about the environment and its own state. Based on multimodal foundation models and reinforcement learning, it performs environmental modeling, object recognition, and intention understanding. Through online reasoning and dynamic planning, it generates control strategies adapted to current scenarios without relying on human teleoperation commands or fixed rules. The system drives mechanical structures and mobile platforms to perform movement, manipulation, and collaboration tasks, while continuously receiving environmental feedback during execution. According to interaction results, it updates behavioral strategies and improves task performance in unknown and dynamic environments.

Overall, disembodied intelligence, spatial intelligence, physical intelligence, and embodied intelligence are not independent concepts, but rather represent a continuous capability evolution from digital space toward the physical world. Disembodied intelligence solves problems related to symbols, language, knowledge, and content generation, but lacks physical embodiment and real-world feedback. Spatial intelligence further establishes understanding of 3D structures, scene layouts, and object relationships. Physical intelligence focuses on modeling physical laws such as gravity, contact, friction, materials, and causal dynamics, providing predictive capabilities regarding how the world evolves. Embodied intelligence integrates these capabilities into systems with physical constraints and action abilities, forming a perception-cognition-action closed loop. Therefore, embodied intelligence is not simply an accumulation of spatial intelligence and physical intelligence. Instead, it integrates representation, prediction, decision-making, and execution into a unified interaction cycle. The comparison among the four concepts is shown in Table 3.

Table 3. Comparison of disembodied, spatial, physical and embodied intelligence.

Intelligence Type	Core Definition	Physical Entity	Interaction with the Environment	Typical Capabilities
Disembodied Intelligence	Intelligence that performs computation and reasoning in a purely informational space	No physical entity; exists purely in software form	Does not interact with the physical world; only processes symbols, data, and digital signals	Language understanding, logical reasoning, image recognition, knowledge question answering, content creation
Spatial Intelligence	Intelligence capable of perceiving, understanding, reasoning about, and generating three-dimensional spatial structures and relationships	Does not require a physical entity; can exist in virtual space	Perceives, navigates, and manipulates spatial information in virtual or digital space without physical interaction	3D perception, spatial navigation, scene understanding, 3D content generation, visual localization
Physical Intelligence	Intelligence capable of understanding the laws of the physical world and performing physical reasoning	Does not require a physical entity; often relies on simulation environments	Learns and predicts the consequences of physical interactions; can apply actions in simulation but not necessarily in the real world	Physical commonsense prediction, object motion estimation, stability judgment, intuitive physical reasoning
Embodied Intelligence	Intelligence with a physical body that perceives and acts through interactions with the environment	Must rely on a physical body	Continuously applies physical actions and receives real-time physical feedback in the real world through the body	Fine manipulation, autonomous locomotion, tool use, dexterous manipulation, adaptation to dynamic environments

The evolutionary path of the above concepts reflects the continuous shift in the focus of artificial intelligence research: from information processing to interaction with the physical world, and from understanding the world to changing the world. Spatial intelligence answers the question of what the world is like, physical intelligence answers how the world works, and embodied intelligence further answers how to act in the world. Therefore, this study regards embodied intelligence as an integrated paradigm that connects foundation models, physical laws, and robotic systems. Its innovation potential

lies not only in the improvement of individual model capabilities, but also in the closed-loop synergy among the body, the environment, and learning mechanisms, as shown in Figure 2.

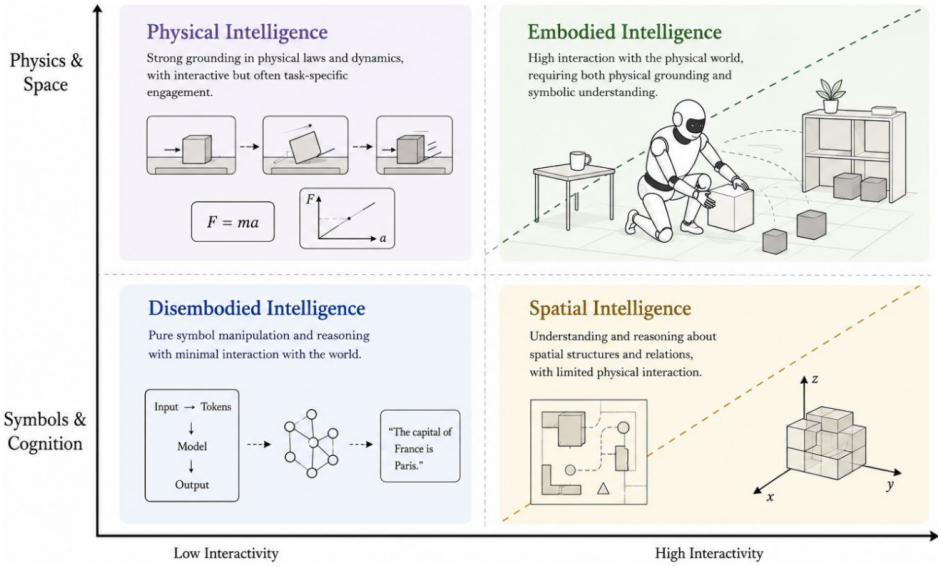


Figure 2. The framework for disembodied, spatial, physical and embodied intelligence. The conceptual figure was created with the assistance of ChatGPT and subsequently reviewed, edited, and refined by the authors.

3. Key technologies

The key technologies of embodied intelligence can be regarded as a full-stack technical system jointly composed of embodiment, motion control, cognitive decision-making and the data loop. This paper presents a technical framework for embodied intelligence organized around five key components: embodiment, cerebellum, brain, data collection, and training. The embodiment determines how an intelligent agent enters the physical world, the cerebellum is responsible for real-time motion coordination and stable execution, the brain undertakes task understanding, world modeling, and policy generation, while the data and training framework determine whether the system can acquire generalizable capabilities from limited experience.

3.1. Embodiment

The embodiment form of embodied intelligence directly determines the robot's motion capability and application scenarios. At present, it mainly evolves along four directions: humanoid, quadruped, wheeled, and hybrid embodiments [20]. Humanoid embodiments imitate the human body structure and motion mechanism, and possess bipedal locomotion and dexterous dual-arm manipulation capabilities. Their structure is naturally compatible with human environments, making them an ideal platform for general-purpose task execution and deep human–robot interaction. They impose high requirements on whole-body coordinated control, dynamic balance, and multimodal perception. Quadruped embodiments adopt bio-inspired leg structures. Through multi-point support and highly dynamic gaits, they achieve excellent mobility and stability in unstructured terrain. Compared with wheeled robots, they are better able to traverse stairs, slopes, and rubble, and are widely used in inspection, exploration, and load transportation.

They serve as highly adaptable mobile platforms for complex outdoor and industrial environments. Wheeled embodiments are based on differential-drive or omnidirectional mobile platforms. They feature high motion efficiency, low energy consumption, and simple control, enabling high-speed and accurate autonomous navigation in flat and structured environments. They are currently the most widely deployed embodiment form and are extensively used in logistics, commercial cleaning, and indoor security. Hybrid embodiments typically integrate an autonomous mobile platform with a multi-degree-of-freedom robotic arm to form a unified mobile manipulation system. They overcome the limitations of fixed workstations by performing grasping, assembly, and material handling while moving, making them the core embodied platform for intelligent manufacturing and flexible production lines. Typical embodiment forms of embodied intelligence are shown in Table 4.

Table 4. Representative embodied intelligence robots.

Embodiment Type	Name	Height/Dimensions	Payload	Maximum Speed ($m \cdot s^{-1}$)	Degrees of Freedom (DOF)	Main Sensors
Humanoid	Atlas [23]	Approximately 1.5 m	Approximately 11 kg	Approximately 1.5	28+ (whole body)	Depth camera, LiDAR, IMU
	Figure 03 [24]	Approximately 1.70 m	Approximately 20 kg	1.2	41 (including dexterous hands)	Whole-body tactile sensors, multi-camera system
	NEO [25]	1.65 m	20 kg	1.1	20+ (tendon-driven)	Visual navigation, depth camera
	H2 [26]	Approximately 1.80 m	≥30 kg	3.3	27+ (estimated)	3D LiDAR, depth camera
Quadruped	GR-3 [27]	1.75 m	Approximately 50 kg	Approximately 1.4	40+ (whole body)	Surround-view vision, torque sensors
	Honey Badger 4.0 [28]	Approximately 80*50*40 cm	Approximately 20 kg	Approximately 0.8	Optional robotic arm	Camera, LiDAR
	ANYmal X [29]	Approximately 0.9 m	10–15 kg	1.3	12 (legs)	LiDAR, thermal imaging camera, explosion-proof enclosure
	Stretch 3 [30]	Approximately 1.4 m (standing)	2 kg	1.0	4+ (telescopic arm)	RGB-D camera, LiDAR
Hybrid	EVE [31]	Approximately 1.9 m	15 kg	2.2	Dual arms + dexterous hands	Surround-view vision, LiDAR
	AGIBOT G2 [32]	Approximately 1.7 m	5–7 kg per arm	Approximately 1.1	28+	Multimodal sensors, dexterous hands
	R1 Pro [33]	1.72 m	20 kg per arm	Approximately 1.4	≈ 15	Depth vision, force-controlled joints
	Galbot G1 [34]	1.73 m	Approximately 10 kg per arm	Approximately 1.5	Dual arms + dexterous hands	3D vision, tactile sensors
	RIVR [35]	Approximately 0.6 m	Approximately 6 kg	Approximately 4.0	12 (legs)	Wide-angle depth camera, LiDAR
	TRON 2 [36]	Approximately 0.70 × 0.50 m	Approximately 50 kg	2.0	Optional collaborative arm	LiDAR, IMU
	B2-W [37]	Approximately 1.0 m (standing)	Maximum 40 kg	Approximately 4.2 (wheeled)	Four wheels + leg joints	3D LiDAR, depth camera

3.2. Cerebellum

The cerebellum of embodied intelligence serves as the core hub of motion control. It is responsible for integrating multimodal sensory information in real time and generating precise motion commands to ensure agile response, coordinated movement, and stable execution in dynamic environments. In the motion control architecture of embodied intelligence, the low-level motion coordination layer represented by the cerebellum is generally responsible for real-time decision-making and whole-body coordination. Its main implementation paradigms can be summarized as Model Predictive Control (MPC) and Whole-Body Control (WBC).

MPC employs an internal predictive model to optimize motion trajectories online. Similar to the feedforward prediction and smooth regulation functions of the biological cerebellum, it enables robots to generate predictive motions in dynamic and unstructured environments [38]. In recent years, whole-body MPC has further unified centroidal dynamics and whole-body dynamics within a single optimization framework, explicitly handling multiple contacts and force constraints, thereby achieving integrated planning from locomotion to upper-limb manipulation [39]. WBC solves joint torques satisfying multiple task priorities and physical constraints in real time through hierarchical quadratic programming, decomposing high-level commands into coordinated whole-body joint motions [40]. In practical systems, MPC and WBC are often combined hierarchically: MPC generates reference trajectories for motion and contact forces, while WBC tracks these references at high frequency and compensates for model mismatch and unmodeled dynamics, thereby achieving both predictive capability and real-time robustness [41]. In addition, reinforcement learning has played an increasingly important role in the development of motion control algorithms for embodied intelligence. For example, deep reinforcement learning has been used to enable adaptive navigation and gait switching of quadruped robots on complex unstructured terrain [42], while adversarial motion priors have been employed to learn highly dynamic parkour skills and whole-body coordination for bipedal robots [43].

3.3. Brain

This paper classifies the brain architectures of embodied intelligence into five technical paradigms, *i.e.*, the traditional layered modular architecture, the large model-driven layered modular architecture, the Vision-Language-Action (VLA) model [44], the World Model (WM) [45] and the World Action Model (WAM) [46]. The traditional hierarchical modular architecture emphasizes engineering controllability; the large-model-driven architecture enhances open-ended semantic understanding; the VLA model aims at end-to-end mapping from perception to action; the World Model provides internal simulation capability; and the World Action Model further couples prediction and action generation within a unified computational framework. The typical architectures of the five paradigms are shown in Figure 3.

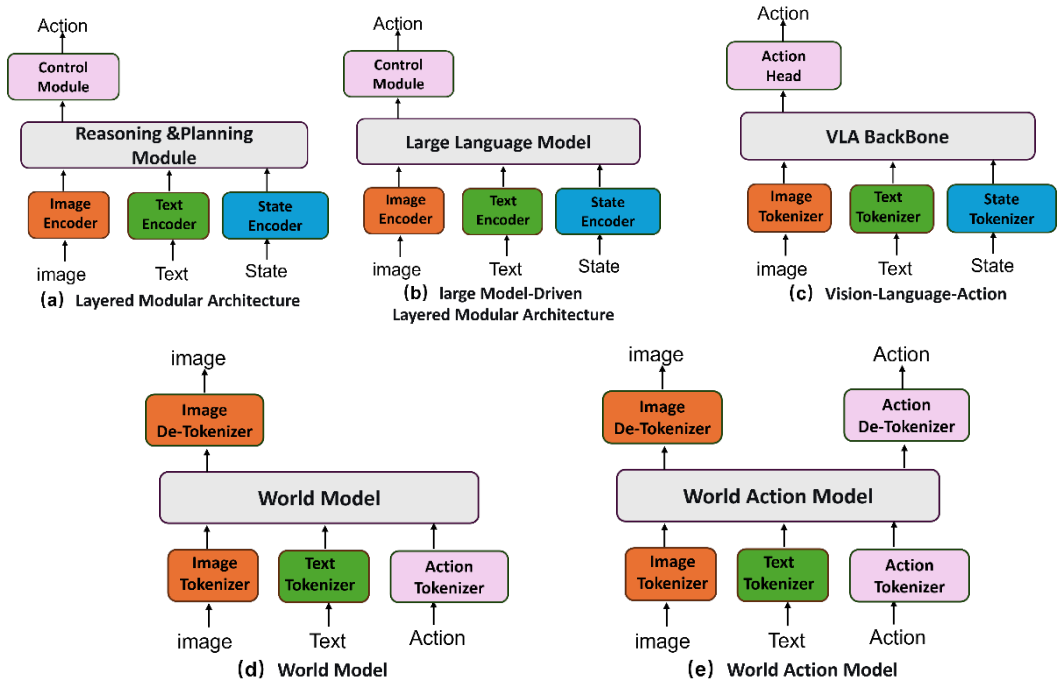


Figure 3. Typical brain architectures for the embodied intelligence.

3.3.1. Traditional hierarchical modular architecture

The traditional hierarchical modular architecture is the classical paradigm for engineering embodied intelligence systems. Its basic idea is to divide perception, localization, planning, and control into relatively independent modules and transfer intermediate representations through predefined interfaces.

This approach provides high reliability and interpretability in structured scenarios such as autonomous driving, industrial robots, and mobile robots, and facilitates local debugging, constraint verification, and safety redundancy design. However, it also has obvious limitations. The interfaces between modules compress information, local optimality cannot guarantee global optimality, and errors may accumulate progressively along the perception–planning–control pipeline. Therefore, the hierarchical modular architecture is more suitable as an engineering baseline for highly reliable systems than as the ultimate architecture for general embodied intelligence in open environments.

3.3.2. Large-model-driven hierarchical modular architecture

To compensate for the limitations of the traditional hierarchical architecture in high-level reasoning and task generalization, researchers have begun to employ large-scale pretrained models as the core of decision-making and planning, forming the large-model-driven hierarchical modular architecture.

This paradigm retains the reliability of low-level motion planning and feedback control while utilizing large language models or vision-language models for open-ended instruction understanding, task decomposition, and dynamic replanning. Representative works include SayCan [47], which balances semantic feasibility and physical feasibility; Code as Policies [48] , which composes low-level skills through code generation; and Inner Monologue [49] , which injects environmental feedback into language models to support closed-loop execution.

The advantage of this architecture lies in its strong interpretability and engineering reusability. However, a representation gap still exists between high-level semantic understanding and low-level motion control.

3.3.3. VLA architecture

Although the large-model-driven hierarchical architecture improves the generalization capability of high-level cognition, it still retains predefined module boundaries and interfaces, resulting in information compression and accumulated errors during transmission. The VLA model abandons explicit hierarchical decomposition and instead jointly models perception, reasoning, and action in an end-to-end manner.

The VLA model takes visual observations and natural language task instructions as inputs and directly outputs executable robot control actions, realizing an end-to-end mapping from high-dimensional perceptual semantics to low-dimensional physical actions. The essence of VLA is to transfer the scene understanding, commonsense reasoning, and generalization capabilities learned by large-scale vision-language models from Internet data to physical robot manipulation tasks through the introduction of the action modality.

The core principle of VLA is to discretize continuous actions into a finite number of action tokens, which, together with visual tokens and language tokens, form a unified multimodal sequence for autoregressive prediction by a large language model. Through joint training on Internet vision-language data and robot demonstration data, the model learns semantic understanding from vision-language priors and maps it into the physical action space. During inference, the model takes the current image and task instruction as input, generates future action sequences token by token, and then decodes them into joint commands or end-effector control commands to drive the robot. Recent analysis of VLA models for generalist robots further shows that model architecture, action representation, observation history, and the diversity of embodied training data are important factors affecting their generalization and performance [50].

The VLA model can be formally expressed as learning an action policy distribution conditioned on visual observations and language instructions:

$$P_\theta(A_{t:t+H-1} \mid h_t, L) = \int q_\theta(c_t|h_t, L) P_\theta^{act}(A_{t:t+H-1}|c_t)\, d_{c_t} \tag{1}$$

In the above equation, the left-hand side represents the conditional policy distribution for generating the future H-step action sequence $A_{t:t+H-1}$ given the interaction history h_t and the language instruction L, which corresponds to the input-output relationship of the VLA model. The right-hand side represents its implementation, where the conditional representation model $q_\theta(c_t|h_t, L)$ fuses the visual or multimodal observation history with the language instruction into a task-conditioned representation c_t, and the action generation model $P_\theta^{act}(A_{t:t+H-1}|c_t)$ generates the action sequence based on this representation by integrating over all possible conditional representations. Here, $h_t = (O_{1:t}, A_{1:t})$ denotes the interaction history up to time step t; $O_{1:t}$ denotes the visual or multimodal observation history; L denotes the language task instruction; $A_{t:t+H-1}$ denotes the future action chunk; H denotes the action prediction horizon; c_t denotes the task-conditioned representation that integrates the environmental state, task semantics, and historical information; and θ denotes all learnable parameters of the VLA model. P_θ^{act} is the unified action probability model, which can be implemented using autoregressive models, diffusion models, or flow matching. The output of the VLA model is the action sequence and does not require explicitly generating the future world. The representative VLA Models are summarized in Table 5.

Table 5. Typical VLA Models for embodied intelligence.

Series	Model	Year	Open Source	Main Features
RT	RT-1 [51]	2022	Y	The first VLA model based on Robotics Transformer, using sequence modeling for robot actions.
	RT-2 [44]	2023	Y	Integrates Internet-scale vision-language knowledge to enable semantic reasoning and action generation.
	RT-2-X [52]	2023	Y	Strong cross-robot generalization capability, trained on the Open X-Embodiment dataset.
	RT-X [52]	2023	Y	Trained on multiple robots and multiple tasks, enabling cross-platform transfer
π	π0 [53]	2024	Y	Uses Flow Matching for action generation, providing strong continuous action control capability.
	π0-FAST [54]	2025	Y	Introduces the FAST Action Tokenizer to achieve low-latency action generation.
	π0.5 [55]	2025	Y	Designed for open-world scenarios and supports generalization to unseen environments.
	π0.6 [56]	2025	Y	Introduces hierarchical planning and an improved VLM backbone, improving manipulation success rate.
	π*0.6 [56]	2025	Y	Incorporates reinforcement learning based on π0.6, enabling policy optimization from real execution experience.
	π0.7 [57]	2026	Y	Supports compositional skills for completing unseen tasks and achieves the strongest generalization in open environments.
GEN	GEN-0	2025	N	The first model to validate robot scaling laws, specifically designed for physical tasks with native support for perception and action.
	GEN-1	2026	N	Achieves a success rate of over 99%, 2–3× faster execution, 10× higher data efficiency, and trains most parameters from scratch.
GROOT	GR00T N1 [58]	2025	Y	The world's first open-source foundation model for humanoid robots, featuring a dual-system architecture with a slow reasoning system and a fast diffusion-based action system.
	GR00T N1.5	2025	Y	Expands the diffusion Transformer action head to 16 layers compared with GR00T N1.
	GR00T N1.6	2026	Y	Uses a 32-layer diffusion Transformer action head with only four diffusion steps, achieving 22.8 Hz inference on an RTX 4090, relative-state behavior prediction, and strong generalization capability.

3.3.4. World Model

Although VLA models establish an end-to-end mapping from perception to action, they still directly generate actions based on the current observation and lack an explicit predictive mechanism for future environmental evolution. To address this limitation, the World Model introduces an internal model of environmental dynamics, enabling embodied intelligence systems to predict future states before taking actions [59]. The core idea of the World Model is to learn the state transition patterns of the external environment from large-scale interaction data and construct an internal predictive model that can simulate future environmental evolution. Instead of directly outputting actions, the World Model predicts future observations or latent states conditioned on the current observation and historical information, thereby providing environmental predictions for planning and decision-making. This mechanism enables the agent to anticipate future states before acting, thereby significantly improving long-horizon planning and decision-making in complex tasks.

The World Model can be formulated as:

$$P_\phi(O_{t+1:t+H} \mid h_t, L, A_{t:t+H-1}) = \int q_\phi(s_t|h_t, L)\, P_\phi^{world}(O_{t+1:t+H}|s_t, A_{t:t+H-1}) d_{s_t} \quad (2)$$

In the above equation, the left-hand side represents the probability distribution for predicting the future observations $O_{t+1:t+H}$ given the interaction history h_t, the language task instruction L, and the candidate action sequence $A_{t:t+H-1}$, which corresponds to the input-output relationship of the language-conditioned World Model. The right-hand side represents its implementation, where the conditional world representation model $q_\phi(s_t|h_t, L)$ encodes the current world state and task semantics into a task-conditioned world

representation s_t and the future world generation model $P_\phi^{world}(O_{t+1:t+H}|s_t, A_{t:t+H-1})$ predicts the future observations corresponding to the candidate action sequence by integrating over all possible conditional world representations. Here, s_t can be regarded as the combination of the world state representation z_t and the task-conditioned representation c_t , *i.e.*, $s_t = (z_t, c_t)$; ϕ denotes the parameters of the World Model; and H denotes the prediction horizon. The output of the World Model is the predicted future world corresponding to the candidate action sequence. The representative World Models are summarized in Table 6.

Table 6. Typical World Models for embodied intelligence.

Model	Release Time	Open Source	Main Features	Key Performance Metrics	Training Data
Dreamer 4 [60]	2025	Y	Based on a latent-space world model, enabling long-horizon planning and extremely low real-world interaction training	Requires only approximately 1/100 of the real-world data used by VPT for the Minecraft Diamond task	Minecraft interaction data
VerseCrafter [61]	2025	Y	Text-driven generation of editable and physically consistent large-scale 3D worlds	Supports open-world 3D environment generation	Indoor/industrial 3D scene datasets
Ctrl-World [62]	2025	Y	Supports future world generation driven by control signals such as keypoints and trajectories	Improves policy performance by 44.7% with zero real-robot data	Zero real-robot data
NeoVerse [63]	2026	Y	Outputs diverse 3D environments with geometric consistency, semantic rationality, and direct interaction capability	State-of-the-art geometric consistency and strong simulation-to-real alignment	Industrial/home digital twin data
Cosmos WFM [64]	2026	Y	Unified modeling of videos, physical states, and action tokens	Supports joint generation of video and physical tokens; 8× spatiotemporal compression	10 million hours of video + 9000T tokens
Genie-Envisioner [65]	2025	Y	Infers and generates an operable virtual world from a single image, achieving a transition from static observation to dynamic prediction	State-of-the-art cross-platform generalization and long-horizon prediction	3000 hours of real-robot data
DexWorldModel [66]	2026	Y	Models dexterous hand contact dynamics and enables fine manipulation prediction	94% success rate on RoboTwin; 50% zero-shot Sim2Real improvement	Simulation data

3.3.5. World Action Model

Classical World Models mainly focus on predicting the system's own states or observations, while their modeling of the causal effects caused by action interventions is insufficient. Based on this, the World Action Model explicitly introduces action conditions and establishes a closed-loop reasoning mechanism of action–state–outcome, providing a more complete simulation environment for model-based behavior decision-making [67]. Different from the VLA architecture, which focuses on the direct mapping from perception to action, and also different from the World Model, which uses environmental prediction as an auxiliary tool for policy optimization, the core characteristic of the World Action Model is that it does not pursue the shortest path from perception to execution. Instead, it emphasizes constructing a predictable physical engine inside the model, allowing the agent to first simulate the possible physical consequences of actions before execution, thereby forming a cognitive closed loop of understanding–reasoning–correction.

Within this closed loop, prediction itself is an intrinsic component of decision-making rather than an external auxiliary module. VLA models determine appropriate actions based on scene observations, whereas World Models predict future outcomes resulting from actions. World Action Models unify these capabilities within a single computational framework by integrating predictive reasoning with action-based decision-making.

Formally, the World Action Model unifies action generation and world reasoning into a joint optimization problem. Specifically, the model takes the current environmental state and task instruction as conditions and simultaneously outputs the predicted state at the next time step and the executable action at the current time step. Different from VLA, which only outputs actions, and the World Model, which mainly outputs state predictions, the key characteristic of the World Action Model is that state reasoning and action generation share the same computational graph, where prediction and decision-making mutually constrain and jointly optimize each other in the latent space. By expanding this joint generation process, the model can be further extended to multi-step reasoning, thereby supporting forward-looking decision-making in long-horizon tasks.

The World Action Model can be formally expressed as integrating action generation and world reasoning into the same conditional joint distribution:

$$P_\varphi\left(A_{t:t+H-1}, O_{t+1:t+H} \mid h_t, L\right) = \int q_\varphi(s_t|h_t, L)\, P_\varphi^{joint}(A_{t:t+H-1}, O_{t+1:t+H}|s_t) d_{s_t} \quad (3)$$

In the above equation, the left-hand side represents the probability distribution for jointly generating the future actions $A_{t:t+H-1}$ and future observations $O_{t+1:t+H}$ conditioned on the interaction historyh_t and the language task instruction L, which corresponds to the input-output relationship of the World Action Model. The right-hand side represents its implementation logic, where the conditional world representation model $q_\varphi(s_t|h_t, L)$ encodes the current world state and task semantics into the task-conditioned world representation s_t, and the joint action-world generation model $P_\varphi^{joint}(A_{t:t+H-1}, O_{t+1:t+H}|s_t)$ simultaneously generates executable actions and their corresponding future world states, while integrating over all possible conditional world representations. Here, $s_t = (z_t, c_t)$, where z_t represents the current world state and c_t represents the task condition; η denotes the parameters of the WAM; and H denotes the generation horizon. The joint generation model P_φ^{joint} can be implemented using video-action joint diffusion, autoregressive generation, imagination-before-execution, or future-feature-conditioned policies.

The output of WAM simultaneously encompasses action generation and future-state prediction. Based on Equations (1)–(3), the technological evolution from VLA models to World Models and finally to World Action Models can be characterized by a progressive shift from direct action generation, to consequence prediction conditioned on actions, and ultimately to the joint generation of actions and their predicted consequences. The typical World Action Models are shown in Table 7.

Table 7. Typical World Action Models for embodied intelligence.

Model	Release Time	Open Source	Main Features	Training Data
VLA-JEPA [68]	2025	Y	Learns a shared latent space, enabling mutual prediction among language instructions, visual observations, and action sequences	Multimodal robot manipulation data
RynnVLA-002 [69]	2025	Y	Implements a real-time closed-loop vision-language-action flow. It not only generates actions but also continuously updates its belief about the world in real time	Three-tokenizer unified encoding
GigaBrain-0 [70]	2025	Y	A world-model-enhanced vision-language-action foundation model that unifies environmental understanding, future state prediction, and robot action generation within a single framework	Large-scale robot interaction data
Dream Zero [71]	2026	Y	A world action model that jointly predicts future video frames and continuous robot actions, enabling zero-shot task generalization, cross-embodiment skill transfer, and real-time closed-loop robotic control	Large-scale multi-robot interaction data, human videos, and robot demonstration data

3.3.6. Comparison of five brain model architectures

The above five types of embodied brain architectures can be understood as a continuous spectrum from engineering controllability to frontier exploration. The hierarchical modular architecture emphasizes reliable interfaces and verifiable execution; the large-model-driven hierarchical architecture enhances open semantic understanding; VLA pursues end-to-end mapping from perception to action; the World Model provides future state prediction; and the World Action Model further unifies prediction and action generation. The main differences among them are shown in Table 8.

Table 8. Comparison of the five brain model architectures.

Architecture Paradigm	Core Idea	Main Advantages	Main Disadvantages	Typical Application Scenarios	Maturity
Hierarchical Modular Architecture	Divides perception, localization, planning, and control into independent modules	Clear interfaces, easy debugging, and convenient safety verification	Module errors are prone to cascading, and cross-task generalization is limited	Industrial robots, autonomous driving, structured mobile robots	High
Large-Model-Driven Hierarchical Modular Architecture	Uses LLM/VLM for task parsing, skill invocation, and semantic planning	Strong open instruction understanding, good interpretability and engineering reusability	A gap exists between high-level semantic objectives and low-level continuous control	Service robots, tool use, language-conditioned task planning	Medium–High
VLA Model	Unifies vision, language, and action into an end-to-end policy model	Fast response and strong potential for cross-task semantic generalization	Insufficient explicit dynamics modeling, counterfactual reasoning, and safety constraints	General manipulation, mobile manipulation, cross-scenario instruction execution	Medium, rapidly maturing
World Model	Learns latent state transitions and predicts future states internally	High sample efficiency, suitable for long-horizon planning and imagination-based trial and error	Prediction and action generation are separated, with a simulation-to-reality gap	Long-horizon planning, simulation training, risk anticipation	Medium
World Action Model	Jointly models future prediction and action generation within the same framework	Potential for forward-looking decision-making and safety evaluation	Still in the early stage, with immature data, evaluation, and controllability	Long-term autonomous interaction, safety-critical tasks, open-domain exploration	Low–Medium, frontier exploration

These architectures should not be regarded as mutually exclusive alternatives at the same abstraction level. Hierarchical architectures primarily describe how perception, planning, control, and other functions are organized into different modules or levels, whereas VLA models provide a unified mapping from visual observations and language instructions to robot actions. World Models introduce an explicit representation or prediction of environmental states and dynamics, enabling an agent to reason about possible future outcomes before acting. World Action Models further couple world modeling with action generation, aiming to jointly predict how actions affect the environment and determine appropriate actions. Consequently, these paradigms can coexist within a single embodied system: a hierarchical system may incorporate a VLA policy, a VLA agent may use a World Model for prediction, and a World Action Model can serve as an integrated mechanism for prediction and action generation.

3.4. Data collection

Data collection determines what kind of behavior distribution embodied intelligence models can learn, and also determines the upper bound of their generalization capability. Unlike Internet image-text data, robot interaction data has characteristics including high collection costs, large differences in hardware configurations, scarcity of failure samples, and insufficient coverage of long-tail scenarios. Therefore, data collection is not merely a preparation stage before training, but a core infrastructure connecting the real world, simulation environments, and policy learning.

The demonstration data collection paradigm directly affects the generalization capability and training efficiency of models. According to the data acquisition mechanism, current mainstream methods can be divided into three categories: teleoperation-based data collection, motion capture-based data collection, and human demonstration-based data collection [72]. Teleoperation-based data collection records data through the chain of operator–control interface–robot. The operator controls the master device, and the control commands are transmitted to the slave robot in real time after forward kinematics and interpolation calculation. During this process, the system synchronously records joint states, end-effector poses, visual observations, and control signals, forming structured state-action pairs. The ALOHA system proposed by Zhao *et al.* [73] is a typical example of this paradigm. It adopts a master-slave dual-arm configuration and low-cost joint encoders, enabling the collection of high-precision bimanual coordination demonstrations in fine manipulation tasks and providing reliable supervision for behavior cloning and Vision-Language-Action model training.

Motion capture-based data collection directly generates demonstrations from human movements. This type of method utilizes visual tracking or inertial sensing systems to obtain the three-dimensional motion trajectories of human keypoints in real time, and maps human motions to the target robot configuration through motion retargeting and inverse kinematics solving. Fu *et al.* [74] adopted a monocular RGB camera in the HumanPlus framework to estimate full-body human poses in real time. Through whole-body inverse kinematics optimization, executable joint trajectories for humanoid robots are generated, achieving automated conversion from natural human movements to robot demonstrations. Compared with teleoperation, this approach does not require operators to directly manipulate robots, provides higher naturalness of motion, and significantly improves data collection efficiency. However, it relies on accurate motion calibration and mapping algorithms to overcome the kinematic differences between humans and robots. Human demonstration-based data collection further reduces equipment constraints by directly recording multimodal behavioral information during human task execution. Operators typically hold portable devices integrated with visual sensors and operation interfaces to perform tasks in natural environments, while the data acquisition equipment simultaneously records first-person-view images, hand motion sequences, interaction forces, and other signals. The Universal Manipulation Interface (UMI) designed by Chi *et al.* [75] adopts a handheld gripper and wrist-mounted camera to capture real-world demonstration data without relying on a fixed robotic platform. Subsequently, human operations are mapped into robot policies through behavior cloning methods such as diffusion policies. The essence of this paradigm is to abstract human operating experience into generalizable policy representations through the process of observation–recording–learning. Its data collection process has low dependence on facilities and is more suitable for covering long-tail scenarios; however, it imposes higher requirements on noise processing of demonstration data and the robustness of motion mapping. The data collection processes of the three methods are shown in Figure 4.

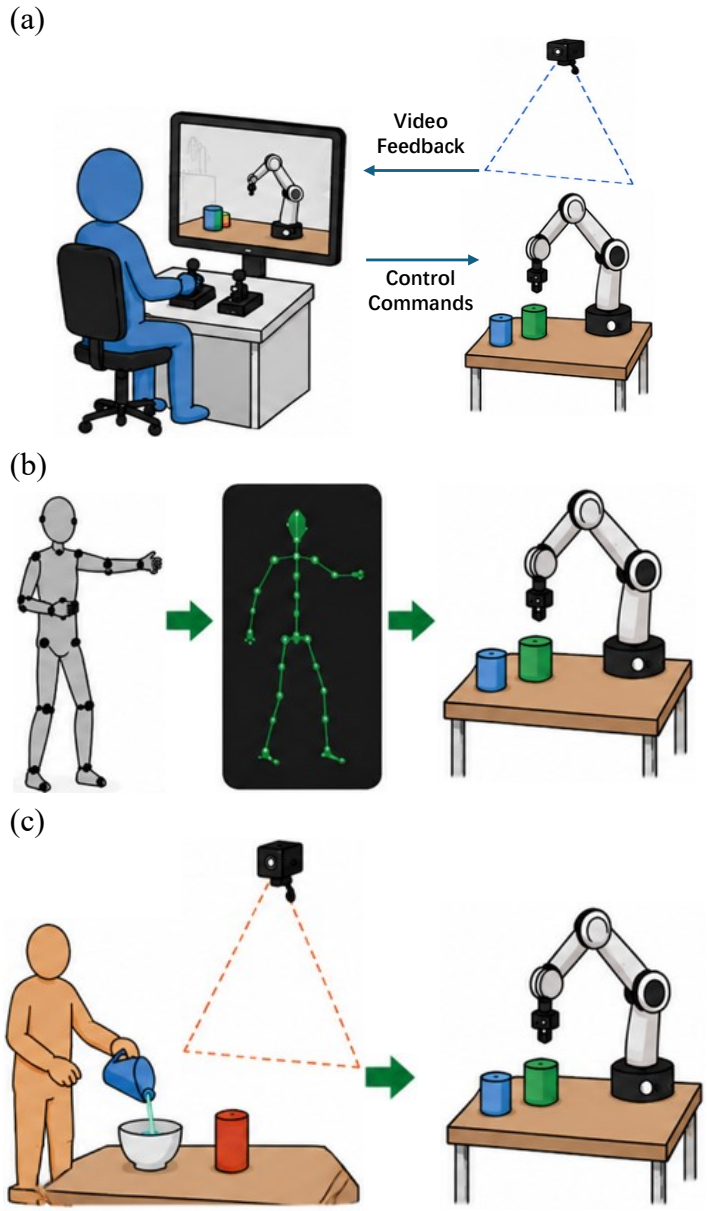


Figure 4. Comparison of three types of data collection processes. The conceptual figure was created with the assistance of ChatGPT and subsequently reviewed, edited, and refined by the authors. **(a)** Teleoperation-based data collection process; **(b)** Motion capture-based data collection process; **(c)** Human Demonstration-based data collection process.

3.5. Model training

3.5.1. Model training methods

After obtaining multi-task and multi-scenario embodied interaction data, training methods determine how agents form transferable perception–motion policies from data. Different from traditional supervised learning, embodied intelligence training not only requires models to fit expert actions, but also needs to handle distribution shifts, environmental feedback, physical constraints, and safety risks. Existing training methods can be summarized into three categories: imitation learning, which extracts policies from expert demonstrations; offline reinforcement learning, which optimizes policies from previously collected interaction data without additional environment interaction; and reinforcement learning from

human feedback, which converts human preferences into learnable rewards. Recent work further emphasizes the bidirectional relationship between embodiment and learning, highlighting that physical embodiment can shape the learning process while learning simultaneously improves an agent's ability to exploit its embodied capabilities[76].

Imitation learning formulates policy learning as supervised regression and minimizes the negative log-likelihood [77]:

$$\mathcal{L}_{\mathrm{BC}} = -E_{(s,a)\sim\mathbb{D}}[\log \pi_\theta\,(a|s)] \tag{4}$$

Where $\mathcal{L}_{\mathrm{BC}}$ is the behavior cloning loss function; (s, a) is the state-action pair;s denotes the state;a denotes the action;$\mathbb{D}$ is the expert demonstration dataset; π_θ is the policy network parameterized by θ; and$E_{(s,a)\sim\mathbb{D}}$ represents the expectation over dataset $\mathbb{D}$.This paradigm is direct and efficient, but suffers from the covariate shift problem, where the mismatch between training and deployment state distributions leads to error accumulation. Inverse reinforcement learning indirectly alleviates this problem by inferring the reward functionR_ϕ, enabling the policy to align with intentions rather than action appearances.

Offline reinforcement learning learns conservatively from a fixed dataset $\mathbb{D}$ without environmental interaction. The interaction data contained in this dataset may have been collected previously through expert demonstrations, human teleoperation, existing policies, or trial-and-error exploration; however, such data collection is distinct from the subsequent offline policy training process. Conservative Q-Learning (CQL) suppresses this problem by penalizing extrapolated values [78]：

$$\min_Q \alpha\left(E_{s\sim\mathbb{D},a\sim\mu}[Q(s,a)] - E_{(s,a)\sim\mathbb{D}}[Q(s,a)]\right) + L_{\mathrm{Bellman}} \tag{5}$$

where $Q(s,a)$ is the action-value function;μ is an arbitrary distribution used to generate candidate actions; a is the penalty coefficient; $\mathbb{D}$ is the offline dataset; and L_{Bellman} is the standard Bellman mean squared error.

It is also commonly combined with behavior regularization constraints to limit the deviation between the policy and the data-generating policy:

$$E_{s\sim\mathbb{D}}\left[D_{\mathrm{KL}}\left(\pi_\theta(\cdot\,|s)\,|\,\pi_\beta(\cdot\,|s)\right)\right] \le \epsilon \tag{6}$$

where *DKL* is the Kullback–Leibler divergence; $\pi_\theta$denotes the learned policy, $\pi_\beta$tdenotes the behavior policy that generated the offline dataset, D denotes the offline dataset, and ϵ denotes the prescribed upper bound on the expected policy deviation from the behavior policy [79].

Reinforcement learning from human feedback replaces manually designed rewards with qualitative comparison signals [80]. Given two trajectory segments (σ^1, σ^2) from the same state, the human preference distribution follows the Bradley–Terry model:

$$P(\sigma^1 \succ \sigma^2) = \frac{\exp \widehat{R_\psi}\,(\sigma^1)}{\exp \widehat{R_\psi}\,(\sigma^1) + \exp \widehat{R_\psi}\,(\sigma^2)} \tag{7}$$

where $\sigma^1 \succ \sigma^2$ represents the preference event that segment 1 is better than segment 2; and $\widehat{R_\psi}$ is the reward model parameterized by ψ.

The reward model is trained by minimizing the cross-entropy loss of preference prediction:

$$L_R wr = -E_{\left((\sigma_1,\sigma_2,y)\sim D_p ref\right)}\left[y log P_{(\sigma_1\succ\sigma_2)} + (1-y) log\left(1 - P_{(\sigma_1\succ\sigma_2)}\right)\right] \tag{8}$$

where $L_R wr$ denotes the reward model training loss,σ_1 and σ_2 denote two trajectory segments being compared, y∈{0,1} denotes the human preference label, $P_{(\sigma_1 \succ \sigma_2)}$denotes the probability that trajectory segment σ_1 is preferred over σ_2, and $D_p ref$ denotes the human preference dataset.

The processes of the three training methods are shown in Figure 5.

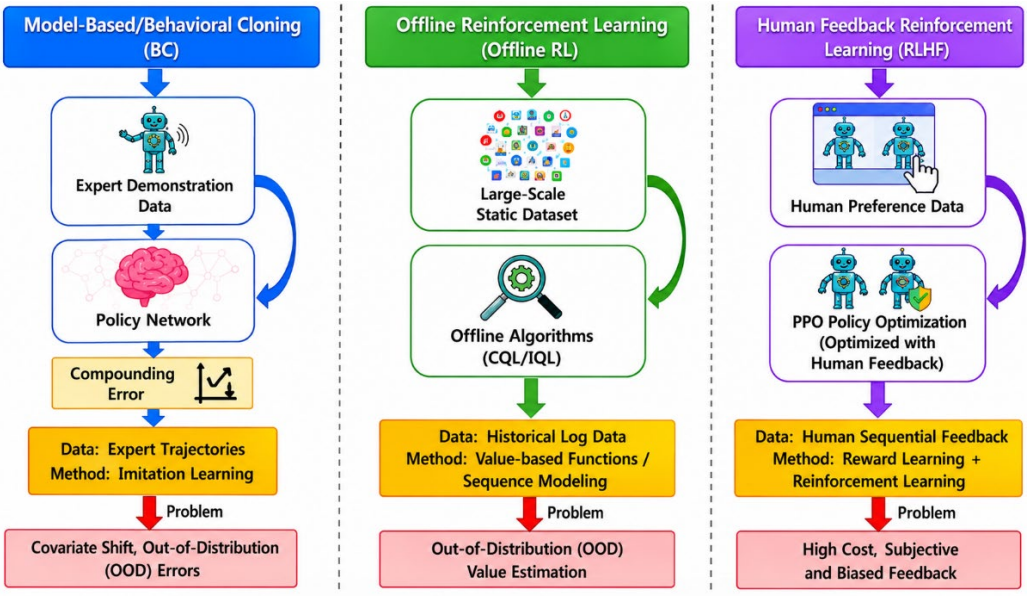


Figure 5. Flowchart of typical training methods. The conceptual figure was created with the assistance of ChatGPT and subsequently reviewed, edited, and refined by the authors.

3.5.2. Model training framework

The engineering deployment of embodied intelligence models not only relies on the performance of the model architecture itself, but also highly depends on the support of a full-process training framework and toolchain covering data collection, model training, simulation validation, and edge deployment. This review systematically summarizes four major categories of current mainstream training frameworks: engineering full-stack frameworks, VLA specific training frameworks, data synthesis and augmentation frameworks, and simulation training platforms.

Engineering full-stack frameworks integrate hardware drivers and upper-level applications, providing an end-to-end solution from prototype development to practical applications, thereby reducing development complexity and cost. VLA-specific training frameworks focus on the training of vision-language-action models, unify multimodal training processes, and improve policy generalization and learning efficiency through methods such as reinforcement learning. Data synthesis and augmentation frameworks systematically address the scarcity of high-quality training data by automatically generating large-scale and diverse synthetic data through simulation environments and performing data augmentation. Simulation training platforms provide high-fidelity virtual environments, allowing robots to perform trial-and-error learning, algorithm validation, and policy training, thereby enabling transfer from simulation to reality. The comparison of representative architectures of the four training frameworks is shown in Table 9.

Table 9. Comparison of representative training frameworks.

Framework Category	Framework Name	Core Positioning	Data Collection	Model Training	Model Inference	Main Characteristics
Engineering Full-Stack Training Framework	FluxVLA Engine	Standardized engineering full-stack foundation	✓	✓	✓	Unifies the entire VLA development workflow and addresses industry fragmentation through standardization
	LeRobot	Unified development framework for robot learning	✓	✓	✓	Provides unified APIs covering the complete closed loop from hardware drivers to edge deployment
VLA Training Framework	VLA Foundry [81]	Full lifecycle pre-training and fine-tuning of VLA	×	✓	✓	Deeply integrates with Hugging Face and controls the entire process from LLM to VLA
	Dexbotic [82]	Multi-model VLA experimental comparison platform	✓	✓	✓	Provides unified interfaces for multiple VLA policies and enables rapid experimental comparison through script-based configuration
	IB-Robot [83]	Industrial-grade ROS 2 bridging framework	✓	✓	✓	Deeply integrates LeRobot with ROS 2 and is suitable for product-level deployment
Data Synthesis and Augmentation Framework	RoboTwin [84]	Dual-arm manipulation data synthesis	✓	✓	✓	Automatically synthesizes high-quality dual-arm collaborative data through generative digital twins
	MimicGen [85]	Automatic demonstration data augmentation	✓	×	×	Automatically expands a small number of demonstrations and increases dataset scale at low cost
	RoboGen [86]	Fully automated data generation exploration	✓	✓	×	AI automatically generates tasks and data, exploring fully automated production processes
	GenAug [87]	Semantic-level visual data augmentation	✓	×	×	Enhances visual data at the semantic level to improve policy generalization capability
Simulation Training Framework	Isaac Lab [88]	GPU-accelerated high-fidelity simulation training	✓	✓	✓	NVIDIA GPU-accelerated robot simulation platform supporting large-scale parallel training and Sim-to-Real transfer
	ManiSkill [89]	Fine manipulation skill simulation training	✓	✓	✓	A fine manipulation policy training and evaluation framework based on SAPIEN

4. Application scenarios

Key technologies can only be transformed into effective capabilities when integrated with real-world scenarios and tasks. This paper summarizes the typical applications of embodied intelligence into three categories: precision manipulation in structured environments, adaptive interaction in unstructured environments, and exploration and navigation in high-risk open domains, as shown in Figure 6. These three types of scenarios correspond to different technical requirements: structured scenarios emphasize precision, stability, and engineering integration; unstructured scenarios emphasize environmental adaptability and compliant interaction; high-risk open domains emphasize uncertainty reasoning, safety redundancy, and human–robot collaboration.

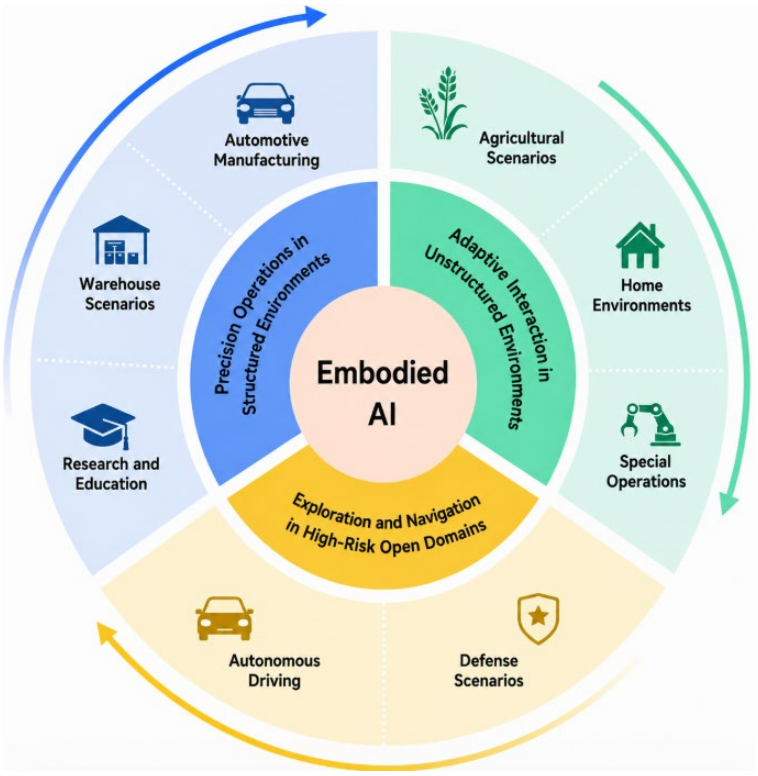


Figure 6. Typical applications of embodied intelligence. The conceptual figure was created with the assistance of ChatGPT and subsequently reviewed, edited, and refined by the authors.

4.1. Precision manipulation in structured environments

Structured environments are characterized by fixed spatial layouts, clear task objectives, and relatively known operating objects. They mainly include scenarios such as intelligent manufacturing, warehouse logistics, and scientific education. The core requirements of these scenarios are high precision, high stability, and repeatable operation. Therefore, they place greater emphasis on engineering reliability, process integration, and long-term data closed loops.

In the automotive manufacturing field, robot systems based on VLA models (such as RT-1 [50] and RT-2 [44]) have recently begun to achieve end-to-end control from vision-language inputs to robot action outputs. Tesla Optimus achieves 3D perception capability comparable to automotive systems by reusing the FSD autonomous driving chip (Tesla AI Day Technical Report, 2021). Figure 02 [90] further validates engineering reliability in high-precision assembly tasks. Similarly, a recent embodied learning approach for peg-in-hole assembly achieves over 98% success rate by combining PPO, compliant force control, and force-vision-position fusion [91].

In warehouse scenarios, Locus Robotics [92] uses a cloud-native intelligent platform as the core, achieving human–robot collaborative operations through multi-sensor fusion localization and dynamic task allocation. Autonomous delivery robots realize automated last-mile logistics delivery through multi-sensor fusion perception and autonomous navigation technologies.

Scientific research and education scenarios demonstrate the potential of embodied intelligence in automated scientific discovery. The robotic chemist platform [93] developed by the University of Science and Technology of China shortens the development cycle of high-entropy catalysts through robotic experiments and intelligent optimization. The Chemifarm automated synthesis facility developed

by the University of Glasgow [94] demonstrates an end-to-end process from code to compounds. The significance of these systems lies not only in improving experimental efficiency, but also in embedding embodied intelligence into the scientific discovery loop, enabling the gradual autonomous coordination of hypothesis generation, experimental execution, and result feedback.

4.2. Adaptive interaction in unstructured environments

Unstructured environments are characterized by changing object states and task objectives that require real-time adjustment based on environmental feedback. Since pre-programmed rules cannot cover all possible situations, embodied intelligence systems must possess real-time perception, semantic understanding, dynamic decision-making, and compliant interaction capabilities to operate reliably in scenarios such as agriculture, household services, and specialized operations. For agricultural operation scenarios, multi-source fusion localization, visual perception, and autonomous planning technologies have been applied to unmanned agricultural machinery, harvesting robots, and plant protection robots [95]. In household environments, healthcare robots developed by teams from Stanford University and other institutions employ imitation learning and full-body teleoperation collaborative training, jointly optimizing mobile manipulation data with existing datasets. Their core functions include basic household tasks, dietary assistance, and emergency assistance. Specialized operation scenarios require embodied intelligence systems to operate reliably in environments with high loads, high risks, or long-term human inaccessibility. The heavy-duty robotic arm system developed by the Institute of Plasma Physics, Chinese Academy of Sciences, achieves a high level of performance in fusion maintenance scenarios in terms of payload capability and dual-arm end-effector repeatability accuracy. These applications demonstrate that the value of embodied intelligence in extreme industrial and scientific environments lies not only in replacing human operations, but also in reducing systemic risks through remote perception, autonomous planning, and highly reliable execution.

4.3. Exploration and navigation in high-risk open domains

High-risk open domains are characterized by unknown or partially observable environments, extremely limited safety margins, and the need for real-time decision-making under incomplete information. Such scenarios not only challenge perception and planning capabilities but also require systems to possess redundant safety mechanisms, anomaly detection, task termination, and human–robot collaborative takeover mechanisms.

Autonomous driving is an important validation scenario for embodied intelligence in high-risk open domains. Waymo [96] achieves autonomous driving capabilities in complex road environments through multi-sensor fusion, high-definition maps, and predictive planning. Tesla FSD [97] adopts an end-to-end neural network architecture based on visual inputs and continuously optimizes model performance through large-scale real-world driving data. In the inspection field, DJI M300 RTK drones are equipped with full-frame cameras and AI recognition algorithms. In logistics delivery, JD Logistics drones adopt multi-source fusion obstacle avoidance using vision and LiDAR.

From application validation perspectives, structured precision manipulation emphasizes reliable reproduction and engineering integration, unstructured adaptive interaction emphasizes generalization and compliant decision-making, and high-risk open domains emphasize uncertainty reasoning under safety

constraints. These three types of scenarios jointly constitute a validation ladder for embodied intelligence from controllable deployment to open autonomy, and reveal an important trend, *i.e.*, the maturity of embodied intelligence should not be evaluated only by single-task success rates, but should also consider long-term operation, anomaly recovery, cross-scenario transfer, and human–robot collaboration capabilities.

5. Evaluation benchmarks

After reviewing the typical application scenarios of embodied intelligence, how to systematically measure and analyze the performance of different agents in complex environments has become the core question addressed by evaluation benchmarks. This paper categorizes embodied intelligence evaluation benchmarks into three types: offline data-based evaluation, simulation-based evaluation and real-world hardware deployment evaluation. Data-based evaluation provides low-cost and reproducible preliminary screening; simulation-based evaluation examines interaction capabilities and robustness in long-tail scenarios; real-world based evaluation ultimately verifies safety, stability, and deployability under real physical constraints.

5.1. Data-based evaluation benchmarks

Data-based evaluation, also known as offline evaluation, refers to evaluating embodied intelligence models on fixed datasets without requiring real-time interaction between the model and the environment. This type of evaluation typically relies on pre-collected multimodal data, including visual observations, language instructions, robot states, action trajectories, and task labels. Model capabilities are measured by comparing model outputs with corresponding ground-truth annotations or predefined evaluation criteria. In this setting, datasets serve as the basis for evaluation, while the trained models or agents constitute the objects being evaluated.

In current research, data-based evaluation methods can mainly be divided into four categories: behavior imitation-based evaluation, offline reinforcement learning evaluation, perception and association evaluation, and multimodal embodied reasoning evaluation. These evaluation methods generally rely on fixed datasets or pre-collected multimodal data as the basis for assessment, while the specific evaluation targets and metrics vary across categories. For behavior imitation-based evaluation, historical multimodal data and expert action trajectories can be used as references to assess the consistency between model-generated actions and demonstrated behaviors. For other evaluation settings, datasets may instead provide scene observations, language instructions, task descriptions, state-action-reward sequences, or other annotations for assessing perception, reasoning, planning, or policy performance.

In offline reinforcement learning evaluation, models are evaluated on previously collected state-action-reward sequences without online environment interaction. The prior collection of these data may involve environment interaction, but it is distinct from subsequent offline training and evaluation on the fixed dataset. The D4RL work proposed by Fu *et al.* [98] provides standardized offline reinforcement learning datasets and tasks that support systematic evaluation without online environment interaction. Perception and association evaluation assesses how embodied agent s interpret task instructions and associate linguistic and visual information with relevant objects, states, and task requirements, emphasizing multimodal perception, semantic grounding, and task understanding as the basis for subsequent action planning.

BabyAI [99] proposes a hierarchical task evaluation framework under controllable language instructions to systematically analyze instruction understanding and planning capabilities. ALFRED [100] combines language understanding with long-horizon manipulation and has become a representative benchmark for embodied task evaluation. With the introduction of large language models, SayCan [47] combines language model reasoning with executability evaluation, while PaLM-E [101] unifies vision, language, and robot control through multimodal large models, promoting the transition of perception and association evaluation from single-module testing toward comprehensive cognitive capability assessment.

Multimodal embodied reasoning evaluation does not require models to execute actual actions. Instead, it evaluates spatial reasoning, task planning, and commonsense reasoning capabilities by providing scene descriptions, visual inputs, and task instructions. Huang *et al.* [102] first explored behavior generation evaluation using language models as zero-shot planners. The Voyager system [103] demonstrated a long-term autonomous exploration evaluation framework based on continual learning and skill accumulation in open-world environments. Meanwhile, comprehensive evaluation platforms such as EmbodiedBench [104] and SafeAgentBench [105] have begun to focus on unified multi-task evaluation and safety dimensions, reflecting the transition of embodied intelligence evaluation from single-metric performance testing toward multidimensional comprehensive evaluation systems. Representative datasets and benchmarks used for offline evaluation of embodied intelligence are summarized in Table 10.

Table 10. Representative datasets and benchmarks for offline evaluation of embodied intelligence.

Benchmark Dataset	Core Tasks and Characteristics	Data Type	Data Scale	Task Scenarios
Open X-Embodiment [106]	Contains 22 robot platforms and more than 1 million trajectories; currently one of the largest multimodal general-purpose datasets	Real robot trajectories: RGB images, depth, language instructions, joint states, end-effector actions	1M+ real robot trajectories	Cross-robot multi-task manipulation (pick-and-place, door opening, drawer operation, kitchen tasks, *etc.*)
BridgeData V2 [107]	Focuses on multi-task manipulation in kitchen environments, with high-quality data suitable for validating model generalization	Real robot demonstrations: RGB images, natural language instructions, 7-DoF actions	60,000+ demonstration trajectories	Multi-skill manipulation in kitchen scenarios (opening doors, placing bowls, wiping surfaces, folding towels, *etc.*)
RT-1 [50]	Industrial-scale real robot dataset containing 130,000 vision-language-action samples	Real robot data: RGB images, natural language instructions, end-effector actions	130,000+ vision-language-action samples	Office and kitchen manipulation (pick-and-place, drawer opening/closing, object moving, *etc.*)
D4RL [98]	Industrial standard for offline RL, containing diverse trajectories including maze tasks and robotic arm control	Low-dimensional observations/states, actions, rewards, with some environments providing visual observations	Millions of time steps (e.g., around 1M steps for MuJoCo tasks and around 5M steps for Adroit tasks)	2D maze navigation, robotic arm control, Franka kitchen manipulation, dexterous hand operation
RL Unplugged [108]	Focuses on evaluating model capabilities in complex control tasks	Pixel-level RGB images/low-dimensional states, actions, rewards	Billions of interaction frames (around 200M frames per Atari game, DMC tasks, *etc.*)	Atari 2600 games, DeepMind Control Suite continuous control tasks
ALFRED [100]	Requires models to complete multi-step interactions in household environments according to long natural language instructions	Egocentric RGB videos, natural language instructions, discrete action sequences	25,743 instructions, 8055 task instances	Multi-step daily tasks in household environments (AI2-THOR), such as "heating an apple" and "cleaning objects"
BabyAI [99]	Focuses on logical reasoning and tests compositional generalization through millions of automatically generated instructions	Grid-world state representations, programmatically generated synthetic language instructions, discrete actions	Millions of automatically generated instruction-trajectory pairs	Finding, picking, and moving objects of specific colors/shapes according to compositional instructions in grid environments
ScanRefer [109]	Emphasizes 3D spatial understanding and evaluates the accuracy of associating language instructions with 3D physical entities	3D indoor scene scans (RGB-D point clouds), natural language descriptions, 3D object bounding boxes	51,583 language descriptions covering 800+ real indoor scenes	Localizing unique target objects in indoor 3D scenes based on long natural language descriptions

Table 10. *Cont.*

Benchmark Dataset	Core Tasks and Characteristics	Data Type	Data Scale	Task Scenarios
EmbodiedBench[104]	Comprehensive evaluation covering spatial navigation, object manipulation, and long-horizon logical planning	Visual observations from multiple simulators (RGB/RGB-D), language instructions, discrete/continuous action spaces	Integrated multiple environments with approximately 1,000+ evaluation tasks	Indoor navigation, multi-step object manipulation, long-horizon commonsense reasoning and planning
SafeAgentBench [105]	Focuses on safety evaluation and assesses agent capabilities under constraints	Natural language task descriptions, safety rules, environmental states and feedback	500+ safety-critical tasks	Embodied manipulation or Web interaction requiring avoiding damage and following safety rules
MineDojo [110]/Voyager [103]	Based on Minecraft, evaluating skill accumulation and continual learning in open-world environments	Minecraft gameplay videos, large-scale encyclopedia text, procedurally generated tasks	330,000 hours of videos, 730K+ Wiki pages, thousands of programmable tasks	Open-world exploration, tool crafting, technology tree unlocking, and survival challenges in Minecraft

However, the above benchmark datasets still exhibit significant limitations in evaluating embodied intelligence. Most task scenarios are highly controlled and involve limited closed-loop interactions, while evaluation metrics primarily rely on task success rates. As a result, they fail to comprehensively reflect the robustness, safety, and long-horizon continual learning capabilities of models in real-world open environments. For perception robustness, adversarial testing provides a complementary evaluation dimension by deliberately probing model vulnerability under challenging visual conditions; DSAdv, for example, demonstrates a diffusion-based framework for generating high-fidelity adversarial examples to evaluate the robustness of autonomous driving perception algorithms [111]. Furthermore, a generalization gap from simulation to reality remains prevalent.

5.2. Simulation-based evaluation benchmarks

Current mainstream simulation platforms for embodied intelligence evaluation can be categorized into three types according to their core capabilities. The first category is represented by NVIDIA Isaac Lab [112], a large-scale parallel training and evaluation platform that supports large-scale robot learning and evaluation based on GPU parallel computing. By utilizing vectorized environments, it enables thousands of parallel simulations, achieving high-throughput evaluation of robotic policies. It also supports parameterized scenario generation and domain randomization techniques to automatically construct test cases covering different lighting conditions, object properties, and environmental disturbances. The second category consists of task-specific manipulation evaluation platforms, represented by ManiSkill [89] and ManipulaTHOR [113]. ManiSkill provides a hierarchical evaluation framework covering task types such as grasping, placing, and assembly. ManipulaTHOR, in contrast, focuses on vision-guided manipulation tasks by providing simulation environments specifically designed for vision-driven robotic manipulation. The third category comprises navigation and interaction platforms, represented by Habitat [114] and iGibson [115]. Habitat emphasizes large-scale scene coverage and statistical significance, supporting systematic evaluation across thousands of different 3D scanned environments. iGibson reconstructs realistic household and office environments, providing a validation platform for the long-term autonomous operation of service robots in complex real-world settings. Representative simulation platforms are summarized in Table 11.

Table 11. Typical simulation platforms for embodied intelligence.

Simulation Platform Category	Platform	Rendering Engine	Ray Tracing	Physically Based Rendering (PBR)	Parallel Rendering	Open Source
Large-Scale Parallel Training and Evaluation Platform	Isaac Lab [112]	Vulkan	✓	✓	✓	Y
Task-Specific Manipulation Evaluation Platform	ManiSkill [89]	SapienRenderer (Vulkan-based)	✓	✓	✓	Y
	ManipulaTHOR [113]	OpenGL-based	✓	✓	✓	Y
Navigation and Interaction Platform	Habitat [114]	Magnum/OpenGL	×	Partial	×	Y
	iGibson [115]	PyBullet OpenGL / TinyRender	×	×	×	Y

Although simulation-based evaluation provides an efficient and safe approach for validating embodied intelligence systems, its effectiveness depends on whether it can faithfully reflect the key risks encountered in deployment environments. Current simulation-based evaluation faces at least three major challenges. First, the simulation-to-reality (Sim-to-Real) gap still causes policies to fail in real-world environments. Second, there exists a trade-off between evaluation coverage and interpretability: overly simplified scenarios cannot expose the boundary capabilities of models, whereas excessively complex scenarios make it difficult to identify the causes of failure. Third, although generative simulation based on multimodal foundation models [116] has improved scenario diversity, its physical fidelity and task validity still require further calibration. Therefore, simulation should not be regarded as a substitute for the real world, but rather as an intermediate layer connecting offline validation and real-world deployment.

5.3. Physical deployment-based evaluation benchmarks

Evaluation in real physical environments has become a core criterion for assessing the practicality of embodied intelligence systems. A series of standardized physical deployment benchmarks have been proposed to promote reproducible capability evaluation from different perspectives. For long-horizon and complex manipulation, BEHAVIOR Robot Suite [117] transfers BEHAVIOR simulation tasks to real robot execution, while FurnitureBench [118] focuses on real-world furniture assembly and evaluates planning robustness in high-rigidity tasks through under-constrained part assembly and long-horizon interactions. In terms of multi-task generalization, RH20T [119] establishes a unified evaluation protocol covering more than 50 real-world manipulation tasks, with an emphasis on cross-task skill transfer. For long-horizon language instruction following and functional manipulation, FMB [120] specifically evaluates the ability to use functional tools such as scissors and lighters that require task-specific manipulation to accomplish their intended functions. For language-guided manipulation in open environments, these benchmarks collectively cover a broad evaluation spectrum ranging from dexterous assembly and multi-task manipulation to language instruction following, functional interaction, and mobile navigation, reflecting the transition of physical evaluation from demonstrating isolated capabilities toward systematic and multidimensional reliability assessment. Representative physical deployment benchmarks are summarized in Table 12.

Table 12. Typical physical benchmarks for embodied intelligence.

Benchmark	Robot Configuration	Task Type	Task Environment	Task Duration/ Complexity	Core Evaluation Metrics	Data Collection and Training Paradigm	Unique Contribution and Positioning
BEHAVIOR Robot Suite [117]	Fixed-base single-arm robot	Long-horizon household tasks	Standardized real kitchens and storage rooms	Long-horizon, requiring multi-step object interactions	Task success rate, generalization, reproducibility score	Simulation-to-real policy transfer, partially supported by teleoperation	Physical realization of the BEHAVIOR simulation benchmark
FurnitureBench [118]	Fixed-base robotic arm	Real furniture assembly	Laboratory-based real assembly workstation	Tens of minutes per assembly task with under-constrained assembly	Assembly completion, force-control robustness, error recovery	Teleoperation and autonomous policy learning	Physical benchmark for long-horizon high-rigidity assembly
RH20T [119]	Fixed-base or mobile single-arm robot	Large-scale real-world robot manipulation dataset and benchmark	Standardized multi-task tabletop and real-world scenarios	Multi-step long-horizon tasks including navigation and manipulation	Multi-task success rate, cross-task skill transfer	Teleoperation demonstrations with unified multimodal settings	Physical benchmark for multi-task generalization
FMB [120]	Fixed-base or mobile single-arm robot	Functional object manipulation	Standardized tabletop tool scenarios	Continuous action sequences enabling reproducible dexterous behaviors	Functional success rate, tool generalization	Teleoperation and autonomous policies	Benchmark for tool use and functional interaction

However, while pursuing standardization and reproducibility, current physical deployment benchmarks still face several critical limitations. First, the coverage of scenarios and tasks remains insufficient. Most benchmarks rely on structured tabletop settings or fixed indoor layouts and therefore cannot adequately represent open-world factors commonly encountered in real applications, such as visual distractions, dynamic objects, and changing human intentions. Second, the evaluation dimensions remain relatively limited, with task success rate serving as the primary metric, while comprehensive aspects such as execution efficiency, operational safety, energy consumption, error recovery capability, and the naturalness of human–robot interaction are rarely evaluated systematically. Third, benchmark construction and maintenance are expensive, as scene resetting, equipment calibration, and manual demonstrations require substantial human effort, limiting large-scale expansion and rapid iteration. Fourth, current evaluation protocols have yet to fully incorporate assessments of long-term continuous operation, lifelong learning, out-of-distribution generalization, and continual adaptation in human–robot coexisting environments.

To address these challenges, recent studies are reshaping the paradigm of physical deployment evaluation from multiple directions. In terms of automation and scalability, large vision-language models and generative methods are being employed to automatically construct task variants and perform fine-grained semantic evaluation, significantly reducing the cost of manual resetting and annotation while promoting the emergence of scalable open benchmarks. Regarding the expansion of evaluation dimensions, comprehensive benchmarks incorporating mobile manipulation, whole-body dexterous coordination, and shared human–robot manipulation are emerging, integrating safety, natural interaction, and task execution efficiency into a unified evaluation framework. For long-term and adaptive evaluation, lifelong learning benchmarks and incremental evaluation protocols in continuously evolving environments have been proposed, enabling robots to continually receive new tasks and undergo periodic re-evaluation in real environments, thereby characterizing their long-term reliability and capability development. Meanwhile, a trend toward unified cross-embodiment evaluation is beginning to emerge. By abstracting task representations and skill primitives, researchers are attempting to establish comparable capability metrics for different robot embodiments, such as single-arm manipulators, mobile dual-arm robots, and humanoid robots. These advances are driving physical deployment evaluation beyond static snapshots of capability

toward dynamic, open, and sustainable lifecycle validation, providing a more solid evaluation foundation for the future deployment of general-purpose embodied intelligence systems in the real world.

6. Key challenges

Although significant progress has been achieved in the key technologies of embodied intelligence, substantial limitations remain in deploying these systems in real-world complex environments, such as precision manipulation tasks in manufacturing [43] and long-horizon operation tasks [121]. From the perspective of the current state of the art, the major challenges of embodied intelligence are illustrated in Figure 7.



Figure 7. Key challenges of embodied intelligence. The conceptual figure was created with the assistance of ChatGPT and subsequently reviewed, edited, and refined by the authors.

(1) Scarcity of High-Quality Data. High-quality robot interaction data remain the primary bottleneck limiting the scalability of embodied intelligence. The lack of unified standards across different robot embodiments, sensor configurations, and task protocols makes data difficult to reuse across platforms. In addition, key modalities such as force, tactile, and contact-state information are expensive to collect and inherently noisy, making it challenging for models to learn fine-grained physical interactions. Future research should establish cross-embodiment, cross-task, and cross-scenario data standards and sharing mechanisms while improving long-tail coverage through simulation-based data generation, active sampling, and failure-case collection.

(2) Long-Horizon Memory and Planning. Embodied tasks often involve multi-step operations, delayed feedback, and irreversible consequences, making myopic reactive policies insufficient for ensuring overall task success. World Models and World Action Models provide a promising foundation for anticipatory planning, yet long-horizon reasoning still suffers from error accumulation, state compression distortion, and uncertainty propagation. Future research should integrate long-term memory, hierarchical planning, and uncertainty estimation to enable embodied agents to perform revisable multi-step decision-making in open environments.

(3) Multimodal Data Fusion. Vision, language, force, tactile sensing, and proprioception differ significantly in temporal scales, spatial resolutions, and noise characteristics. Current models tend to

rely excessively on visual semantics while underutilizing low-level physical signals such as contact, force control, and material properties. Future multimodal fusion should evolve from simple feature concatenation toward task-oriented representation learning, with particular emphasis on cross-modal alignment, robustness to missing modalities, and the interpretability of physical signals.

(4) Edge Deployment and Inference Acceleration. Embodied intelligence systems must perform real-time decision-making under constraints of limited computational resources, power consumption, and communication bandwidth. However, the inference latency of large foundation models is inherently incompatible with millisecond-level motion control. Techniques such as model compression, knowledge distillation, sparse activation, and edge-cloud collaboration can alleviate this issue, but they must simultaneously preserve policy safety and cross-scenario generalization. Future edge deployment should focus not only on making models smaller but also on establishing hierarchical decision-making mechanisms that coordinate fast and slow systems.

(5) Model Safety and Trustworthiness. Since embodied intelligence directly interacts with the physical world, failures may result in equipment damage, human injury, or task failure. The black-box nature of end-to-end models makes accident tracing and responsibility attribution more difficult, while the stability and safety guarantees developed in classical control theory are not readily applicable to neural network-based policies. Future research should integrate interpretable decision-making, formal constraints, safety monitoring, and human intervention mechanisms into a unified framework, transforming safety from a post hoc remedy into an intrinsic design principle.

(6) Continual Learning of Skills. Real-world environments evolve continuously, requiring embodied agents to update their capabilities after deployment. However, learning new tasks may degrade previously acquired knowledge, creating a trade-off between stability and plasticity. Future research should balance parameter-efficient adaptation, memory replay, task boundary identification, and online evaluation, enabling models to accumulate long-term experience while minimizing catastrophic forgetting.

(7) Unified Modeling of Navigation and Manipulation. Navigation and manipulation focus on large-scale spatial mobility and fine-grained local interaction, respectively, and differ substantially in coordinate systems, action primitives, and feedback frequencies. Simply cascading navigation and manipulation modules cannot effectively address challenges such as manipulation during locomotion, navigation error propagation, and dynamic target changes. Future research should develop unified spatial-action representations that enable embodied agents to perform continuous closed-loop operations encompassing reaching, alignment, manipulation, and replanning within a single framework.

(8) Multi-Agent Collaboration. A single embodied agent is insufficient for large-scale, long-duration, and highly complex tasks. Multi-agent collaboration is therefore expected to become an important direction for applications such as agriculture, inspection, warehousing, and emergency response. The key challenges include distributed perception consistency, task allocation across heterogeneous platforms, collaborative decision-making under communication constraints, and safety guarantees for collective behaviors. Future research should extend from individual policy learning toward group intention modeling and collaborative robustness evaluation.

7. Conclusion

Embodied intelligence is shifting the focus of artificial intelligence research from information processing in digital space toward autonomous interaction in the physical world. This paper first analyzes the

fundamental differences between embodied intelligence, teleoperation, and conventional automation, and systematically reviews the evolutionary pathway from disembodied intelligence to spatial intelligence, physical intelligence, and embodied intelligence. It then presents a comprehensive overview of the key technological framework of embodied intelligence from five aspects: robotic embodiment, the cerebellum, the brain, data collection, and training methods. Particular attention is given to several representative embodied brain paradigms, including the traditional hierarchical modular architecture, large model-driven hierarchical architecture, VLA models, World Models, and World Action Models. Subsequently, the current application status of embodied intelligence is reviewed across three representative scenarios: precision manipulation in structured environments, adaptive interaction in unstructured environments, and exploration in high-risk open domains. Furthermore, three major evaluation paradigms, namely offline data-based evaluation, simulation-based evaluation, and real-world physical evaluation, are analyzed in terms of their respective advantages and practical limitations. Finally, considering deployment in real-world complex environments, eight major technical challenges are identified, including data infrastructure, long-horizon planning, multimodal fusion, edge deployment, safety and trustworthiness, continual learning, unified navigation and manipulation, and multi-agent collaboration. By systematically reviewing the fundamental concepts, key technologies, representative application scenarios, evaluation methodologies, and future challenges of embodied intelligence, this paper provides a clear conceptual framework, a structured technical roadmap, and practical guidance for advancing embodied intelligence from laboratory demonstrations toward real-world deployment.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used ChatGPT to improve English language readability and to assist in the creation of conceptual figures. After using this tool, the authors reviewed, edited, and refined the generated content and figures as needed. The authors take full responsibility for the accuracy, integrity, originality, and final content of the manuscript and figures.

Acknowledgments

This research was supported by the Intelligent Driving Innovation Consortium under Grand Union of Innovation, the Youth Innovation Promotion Association of the CAS under Y2021115, and the Anhui Provincial Key Science and Technology R&D Program under Grant No. 202423i08050022.

Authors' contribution

Conceptualization, Y.L. and B.Y.; methodology, Y.L.; software, Y.C.; validation, Y.L., X.H. and L.L.; formal analysis, B.L., Y.L.; investigation, Y.L., L.S., and H.M.; resources, B.Y. and C.C.; data curation, Y.L.; writing—original draft preparation, Y.L. writing—review and editing, Y.C., X.H., L.L., L.S., H.M., S.Z., S.Z., B.Y., and C.C.; visualization, Y.L.; supervision, B.Y.; project administration, B.Y.; funding acquisition, B.Y. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Newell A, Simon HA. Computer science as empirical inquiry: symbols and search. *Commun. ACM* 1976, 19(3):113–126.

[2] McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence: August 31, 1955. *AI Mag*. 2006, 27(4):12–14.

[3] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986, 323(6088):533–536.

[4] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems 25 (NIPS 2012)*, Nevada, USA, December 3–8, 2012, pp. 1097–1105.

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, California, USA, December 4–9, 2017, pp. 5998–6008.

[6] Moravec HP. Sensor fusion in certainty grids for mobile robots. *AI Mag.* 1988, 9(2):61.

[7] Zhang Y, Tian J, Xiong Q. A review of embodied intelligence systems: A three-layer framework integrating multimodal perception, world modeling, and structured strategies. *Front. Robot. AI* 2025, 12:1668910.

[8] Li X, He X, Zhang L, Wu M, Li X, *et al.* A comprehensive survey on world models for embodied AI. *arXiv* 2025, arXiv:2510.16732.

[9] Sun F, Chen R, Ji T, Luo Y, Zhou H, *et al.* A comprehensive survey on embodied intelligence: advancements, challenges, and future perspectives. *CAAI Artif. Intell. Res.* 2024, 3:9150042.

[10] Guan W, Hu Q, Li A, Cheng J. Efficient vision-language-action models for embodied manipulation: a systematic survey. *arXiv* 2025, arXiv:2510.17111.

[11] Li G, Wang R, Xu P, Ye Q, Chen J. The developments and challenges towards dexterous and embodied robotic manipulation: a survey. *arXiv* 2025, arXiv:2507.11840.

[12] Varela FJ, Thompson E, Rosch E. *The embodied mind, revised edition: Cognitive science and human experience.* Cambridge: MIT Press, 2017.

[13] Brooks RA. Intelligence without representation. *Artif. Intell.* 1991, 47(1–3):139–159.

[14] Wilson M. Six views of embodied cognition. *Psychon. Bull. Rev.* 2002, 9(4):625–636.

[15] Sheridan TB. *Telerobotics, Automation, and Human Supervisory Control.* Cambridge: MIT Press, 1992.

[16] Vogel-Heuser B, Fischer J, Feldmann S, Ulewicz S, Rösch S. Modularity and architecture of PLC-based software for automated production systems: an analysis in industrial companies. *J. Syst. Softw.* 2017, 131:35–62.

[17] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, *et al.* On the opportunities and risks of foundation models. *arXiv* 2021, arXiv:2108.07258.

[18] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* 2023, 42(4):1–14.

[19] Rus D. From chips to thoughts: building physical intelligence into robotic systems. In *Proceedings of the 2025 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, USA, February 16-20, 2025, pp.16–22.

[20] Liu Y, Chen W, Bai Y, Liang X, Li G, *et al.* Aligning cyber space with physical world: a comprehensive survey on embodied AI. *IEEE/ASME Trans. Mechatron.* 2025, 30(6):7253–7274.
[21] OpenAI. GPT-4o System Card. 2024. Available: https://openai.com/index/gpt-4o-system-card/ (accessed on 17 July 2026).
[22] Ren Z, Wei Y, Guo X, Zhao Y, Kang B, *et al.* Videoworld: exploring knowledge learning from unlabeled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, June 10–17, 2025, pp. 29029–29039.
[23] Boston Dynamics. Atlas. 2026. Available: https://www.bostondynamics.com/atlas (accessed on 1 June 2026).
[24] Figure AI, Inc. Figure humanoid robot platform. 2026. Available: https://www.figure.ai/ (accessed on 1 June 2026).
[25] 1X Technologies. NEO humanoid robot. 2026. Available: https://www.1x.tech/neo (accessed on 1 June 2026).
[26] Unitree. Unitree H2. 2026. Available: https://www.unitree.com/h2 (accessed on 1 June 2026).
[27] Fourier Intelligence. GR-3 humanoid robot. 2026. Available: https://www.fourierintelligence.com (accessed on 1 June 2026).
[28] SuperDroid Robots. Honey Badger 4.0. 2026. Available: https://www.superdroidrobots.com (accessed on 1 June 2026).
[29] ANYbotics. ANYmal X. 2026. Available: https://www.anybotics.com (accessed on 1 June 2026).
[30] Hello Robot. Stretch 3. 2026. Available: https://hello-robot.com (accessed on 1 June 2026).
[31] 1X Technologies. EVE. 2026. Available: https://www.1x.tech (accessed on 1 June 2026).
[32] AgiBot. AGIBOT G2 humanoid robot. 2026. Available: https://www.agibot.com (accessed on 1 June 2026).
[33] Robot Era. R1 Pro. 2026. Available: https://www.robotera.com (accessed on 1 June 2026).
[34] Galbot. Galbot G1. 2026. Available: https://www.galbot.com (accessed on 1 June 2026).
[35] RIVR Systems. RIVR autonomous delivery robot. 2026. Available: https://www.rivr.ai (accessed on 1 June 2026).
[36] AgileX Robotics. TRON 2. 2026. Available: https://www.agilex.ai (accessed on 1 June 2026).
[37] Unitree. Unitree B2-W. 2026. Available: https://www.unitree.com (accessed on 1 June 2026).
[38] Grandia R, Jenelten F, Yang S, Farshidian F, Hutter M. Perceptive locomotion through nonlinear model-predictive control. *IEEE Trans. Robot.* 2023, 39(5):3402–3421.
[39] Minniti MV, Farshidian F, Grandia R, Hutter M. Whole-body MPC for a dynamically stable mobile manipulator. *IEEE Robot. Autom. Lett.* 2019, 4(4):3687–3694.
[40] Zhang T, Yue L, Zhang H, Zhang L, Zeng X, *et al.* Whole-body control framework for humanoid robots with heavy limbs: a model-based approach. *arXiv* 2025, arXiv:2506.14278.
[41] Di Carlo J, Wensing P M, Katz B, Bledt G, Kim S. Dynamic locomotion in the MIT cheetah 3 through convex model-predictive control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, October 1–5, 2018, pp. 7440–7447.
[42] Lee J, Hwangbo J, Wellhausen L, Koltun V, Hutter M. Learning quadrupedal locomotion over challenging terrain. *Sci. Robot.* 2020, 5(47):eabc5986.
[43] Zhuang Z, Yao S, Zhao H. Humanoid parkour learning. *arXiv* 2024, arXiv:2406.10759.

[44] Brohan A, Brown N, Carbajal J, Chebotar Y, Chen X, *et al.* RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv* 2023, arXiv:2307.15818

[45] Ha D, Schmidhuber J. Recurrent world models facilitate policy evolution. *In Advances in Neural Information Processing Systems*, Montréal, Canada, December 3–8, 2018, pp. 2455–2463.

[46] Ye S, Ge Y, Zheng K, Gao S, Yu S, *et al.* World action models are zero-shot policies. *arXiv* 2026, arXiv:2602.15922.

[47] Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, *et al.* Do as I can, not as I say: Grounding language in robotic affordances. *arXiv* 2022, arXiv:2204.01691.

[48] Liang J, Huang W, Xia F, Xu P, Hausman K, *et al.* Code as policies: Language model programs for embodied control. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, May 29–June 2, 2023, pp. 9493–9500.

[49] Huang W, Xia F, Xiao T, Chan H, Liang J, *et al.* Inner monologue: Embodied reasoning through planning with language models. *arXiv* 2022, arXiv:2207.05608.

[50] Li X, Li P, Qian L, Liu M, Wang D, *et al.* What matters in building vision–language–action models for generalist robots. *Nat. Mach. Intell.* 2026, 8(2):158–172.

[51] Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, *et al.* RT-1: Robotics transformer for real-world control at scale. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 10–14, 2023.

[52] O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, *et al.* Open X-Embodiment: robotic learning datasets and RT-X models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 13–17, 2024, pp. 6892–6903.

[53] Black K, Brown N, Driess D, Esmail A, Equi M, *et al.* π_0: a vision-language-action flow model for general robot control. *arXiv* 2026, arXiv:2410.24164.

[54] Pertsch K, Stachowicz K, Ichter B, Driess D, Nair S, *et al.* FAST: efficient action tokenization for vision-language-action models. *arXiv* 2025, arXiv:2501.09747.

[55] Intelligence P, Black K, Brown N, Darpinian J, Dhabalia K, *et al.* $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv* 2025, arXiv:2504.16054.

[56] Intelligence P, Amin A, Aniceto R, Balakrishna A, Black K, *et al.* $\pi^*_{0.6}$: a VLA that learns from experience. *arXiv* 2025, arXiv:2511.14759.

[57] Intelligence P, Ai B, Amin A, Aniceto R, Balakrishna A, *et al.* $\pi_{0.7}$: a steerable generalist robotic foundation model with emergent capabilities. *arXiv* 2026, arXiv:2604.15483.

[58] Bjorck J, Castañeda F, Cherniadev N, Da X, Ding R, *et al.* Gr00t n1: an open foundation model for generalist humanoid robots. *arXiv* 2025, arXiv:2503.14734.

[59] Hafner D, Yan W, Lillicrap T. Training agents inside of scalable world models. *arXiv* 2025, arXiv:2509.24527.

[60] Hafner D, Pasukonis J, Ba J, Lillicrap T. Mastering diverse control tasks through world models. *Nature* 2025, 640(8059):647–653.

[61] Zheng S, Yin M, Hu W, Li X, Shan Y, *et al.* VerseCrafter: dynamic realistic video world model with 4D geometric control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Colorado, USA, June 5–7, 2026, pp. 40277–40290.

[62] Guo Y, Shi L, Chen J, Finn C. Ctrl-world: a controllable generative world model for robot manipulation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, April 23–27, 2026, pp. 6121–6138.

[63] Yang Y, Fan L, Shi Z, Peng J, Wang F, *et al.* NeoVerse: enhancing 4D World Model with in-the-wild monocular videos. *arXiv* 2026, arXiv:2601.00393.

[64] Agarwal N, Ali A, Bala M, Balaji Y, Barker E, *et al.* Cosmos world foundation model platform for physical AI. *arXiv* 2025, arXiv:2501.03575.

[65] Liao Y, Zhou P, Huang S, Yang D, Chen S, *et al.* Genie envisioner: a unified world foundation platform for robotic manipulation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, April 23–27, 2026, pp. 88446–88463.

[66] Deng Y, Liu G, Jia K. DexWorldModel: causal latent world modeling towards automated learning of embodied tasks. *arXiv* 2026, arXiv:2604.16484.

[67] Wang S, Shi J, Fu Z, He X, Liu F, *et al.* World action models: the next frontier in embodied AI. *arXiv* 2026, arXiv:2605.12090.

[68] Sun J, Zhang W, Qi Z, Ren S, Liu Z, *et al.* VLA-JEPA: enhancing vision-language-action model with latent world model. *arXiv* 2026, arXiv:2602.10098.

[69] Cen J, Huang S, Yuan Y, Li K, Yuan H, *et al.* RynnVLA-002: a unified vision-language-action and world model. *arXiv* 2026, arXiv:2511.17502.

[70] Team G, Ye A, Wang B, Ni C, Huang G, *et al.* GigaBrain-0: a world model-powered vision-language-action model. *arXiv* 2025, arXiv:2510.19430.

[71] Ye S, Ge Y, Zheng K, Gao S, Yu S, et al. World Action Models are Zero-shot Policies. *arXiv* 2026, arXiv:2602.15922.

[72] Mandlekar A, Xu D, Wong J, Nasiriany S, Wang C, *et al.* What matters in learning from offline human demonstrations for robot manipulation. *arXiv* 2021, arXiv:2108.03298.

[73] Zhao T Z, Kumar A, Levine S, Finn C. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv* 2023, arXiv:2304.13705.

[74] Fu Z, Zhao Q, Wu Q, Wetzstein G, Finn C. HumanPlus: humanoid shadowing and imitation from humans. *arXiv* 2024, arXiv:2406.10454.

[75] Chi C, Xu Z, Pan C, Cousineau E, Burchfiel B, *et al.* Universal manipulation interface: in-the-wild robot teaching without in-the-wild robots. *arXiv* 2024, arXiv:2402.10329.

[76] Liu H, Guo D, Huang K. Learning for embodiment and embodiment for learning. *Nat. Rev. Electr. Eng.* 2025, 2(10):651–653.

[77] Osa T, Pajarinen J, Neumann G, Bagnell J, Abbeel P, *et al.* An algorithmic perspective on imitation learning. *Found. Trends Robot*. 2018, 7(1–2):1–179.

[78] Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, online, December 6–12, 2020, pp. 1179–1191.

[79] Chebotar Y, Hausman K, Lu Y, Xiao T, Kalashnikov D, *et al.* Actionable Models: unsupervised offline reinforcement learning of robotic skills. *arXiv* 2021, arXiv:2104.07749.

[80] Christiano PF, Leike J, Brown T, Martic M, Legg S, *et al.* Deep reinforcement learning from human preferences. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, USA, December 4–9, 2017, pp. 4299–4307.

[81] Mercat J, Keh S, Arora K, Huang I, Shah P, *et al.* VLA Foundry: a unified framework for training vision-language-action models. *arXiv* 2026, arXiv:2604.19728.

[82] Dexbotic Contributors. Dexbotic: open-source vision-language-action Toolbox. *arXiv* 2025, arXiv:2510.23511.

[83] openEuler Embedded. IB-Robot: intelligence boom robot—an intelligent robot development framework integrating LeRobot and ROS 2 (In Chinese). 2026. Available: https://pages.openeuler.openatom.cn/embedded/docs/build/html/openEuler-26.09/features/embodied_ai/ib_robot.html (accessed on 1 June 2026)

[84] Mu Y, Chen T, Chen Z, Peng S, Lan Z, *et al.* RoboTwin: dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Nashville, USA, June 10–17, 2025, pp. 27649–27660.

[85] Mandlekar A, Nasiriany S, Wen B, Akinola I, Narang Y, *et al.* MimicGen: a data generation system for scalable robot learning using human demonstrations. *arXiv* 2023, arXiv:2310.17596.

[86] Wang Y, Xian Z, Chen F, Wang TH, Wang Y, *et al.* RoboGen: towards unleashing infinite data for automated robot learning via generative simulation. *arXiv* 2024, arXiv:2311.01455.

[87] Chen Z, Kiami S, Gupta A, Kumar V. GenAug: retargeting behaviors to unseen situations via generative augmentation. *arXiv* 2023, arXiv:2302.06671.

[88] Mittal M, Roth P, Tigue J, Richard A, Zhang O, *et al.* Isaac Lab: a GPU-accelerated simulation framework for multi-modal robot learning. *arXiv* 2025, arXiv:2511.04831.

[89] Tao S, Xiang F, Shukla A, Qin Y, Hinrichsen B, *et al.* ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI. *arXiv* 2025, arXiv:2410.00425.

[90] Figure AI. Introducing Figure 02: the next generation humanoid robot. 2024. Available: https://www.figure.ai/company?utm (accessed on accessed on 1 June 2026)

[91] Li F, Qi H, Jin L, Yao X, Men Y, *et al.* Robot assembly strategy optimization based on embodied skill learning. *Robot Learn.* 2026(2):0017.

[92] Fragapane G, Ivanov D, Peron M, Sgarbossa F, Strandhagen JO. Increasing flexibility and productivity in industry 4.0 production networks with autonomous mobile robots and smart intralogistics. *Ann. Oper. Res.* 2022, 308(1):125–143.

[93] Yang R, Zhou D, Jia Z, Han Y, Tang L, *et al.* Toward practical design of high-entropy catalysts for chlorine evolution reaction via Pareto-guided multi-objective Bayesian optimization enabled by a robotic AI-chemist. *Angew. Chem. Int. Ed.* 2026, 65(27):e8794274.

[94] MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* 2020, 6(20):eaaz8867.

[95] Bechar A, Vigneault C. Agricultural robots for field operations: concepts and components. *Biosyst. Eng.* 2016, 149:94–111.

[96] Caesar H, Bankiti V, Lang AH, Vora S, Liong V, *et al.* nuScenes: a multimodal dataset for autonomous driving. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, June 13–19, 2020, pp. 11618–11628.

[97] Tesla AI Team. Full self-driving (supervised): end-to-end neural network architecture and technical overview. 2023. Available: https://www.tesla.com/SI (accessed on 13 July 2026).

[98] Fu J, Kumar A, Nachum O, Tucker G, Levine S. D4RL: datasets for deep data-driven reinforcement learning. *arXiv* 2021, arXiv:2004.07219.

[99] Chevalier-Boisvert M, Bahdanau D, Lahlou S, Willems L, Saharia C, *et al.* BabyAI: a platform to study the sample efficiency of grounded language learning. *arXiv* 2019, arXiv:1810.08272.

[100] Shridhar M, Thomason J, Gordon D, Bisk Y, Han W, *et al.* ALFRED: a benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, June 13–19, 2020, pp. 10737–10746.

[101] Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhury A, *et al.* PaLM-E: an embodied multimodal language model. *arXiv* 2023, arXiv:2303.03378.

[102] Huang W, Abbeel P, Pathak D, Mordatch I. Language models as zero-shot planners: extracting actionable knowledge for embodied agents. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, Baltimore, USA, July 17–23, 2022, pp. 9118–9147.

[103] Wang G, Xie Y, Jiang Y, Mandlekar A, Xiao C, *et al.* Voyager: an open-ended embodied agent with large language models. *arXiv* 2023, arXiv:2305.16291.

[104] Yang R, Chen H, Zhang J, Zhao M, Qian C, *et al.* EmbodiedBench: comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv* 2025, arXiv:2502.09560.

[105] Yin S, Pang X, Ding Y, Chen M, Bi Y, *et al.* SafeAgentBench: a benchmark for safe task planning of embodied LLM agents. *arXiv* 2025, arXiv:2412.13178.

[106] Vuong Q, Levine S, Walke HR, Pertsch K, Singh A, *et al.* Open X-embodiment: robotic learning datasets and RT-X Models. In *Proceedings of Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@CoRL* 2023, Atlanta, USA, November 6–9, 2023.

[107] Walke HR, Black K, Zhao TZ, Vuong Q, Zheng C, *et al.* BridgeData V2: a dataset for robot learning at Scale. In *Proceedings of the 7th Conference on Robot Learning (CoRL)*, Atlanta, USA, November 6–9, 2023, pp. 1723–1736.

[108] Gulcehre C, Wang Z, Novikov A, Paine T, Gómez S, *et al.* RL Unplugged: a suite of benchmarks for offline reinforcement learning. In *Proceedings of Advances in Neural Information Processing Systems*, Online, December 6–12, 2020, pp. 7248–7259.

[109] Chen DZ, Chang AX, Nießner M. ScanRefer: 3D object localization in RGB-D scans using natural language. *arXiv* 2019, arXiv:1912.08830.

[110] Fan L, Wang G, Jiang Y, Mandlekar A, Yang Y, *et al.* MineDojo: building open-ended embodied agents with internet-scale knowledge. In *Proceedings of Advances in Neural Information Processing Systems*, New Orleans, USA, November 28–December 9, 2022, pp. 18343–18362.

[111] Huang C, Shi G, Ji J, Zhang D, Qu C, *et al.* DSAdv: dual-stream diffusion adversarial attacks for autonomous driving perception algorithms. *Robot Learn.* 2026 3(1):0007.

[112] Isaac Lab Contributors. Isaac Lab Documentation. 2025. Available: https://isaac-sim.github.io/IsaacLab/main/index.html (accessed on 11 May 2026).

[113] Ehsani K, Han W, Herrasti A, VanderBilt E, Weihs L, *et al.* ManipulaTHOR: a framework for visual object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, June 20–25, 2021, pp. 4497–4506.

[114] Puig X, Undersander E, Szot A, Dallaire Cote M, Yang TY, *et al.* Habitat 3.0: a co-habitat for humans, avatars and robots. *arXiv* 2023, arXiv:2310.13724.

[115] Li C, Xia F, Martín-Martín R, Lingelbach M, Srivastava S, *et al.* iGibson 2.0: object-centric simulation for robot learning of everyday household tasks. *arXiv* 2021, arXiv:2108.03272.

[116] Yang M, Du Y, Ghasemipour K, Tompson J, Schuurmans D, *et al.* UniSim: learning interactive real-world simulators. *arXiv* 2023, arXiv:2310.06114

[117] Jiang Y, Zhang R, Wong J, Wang C, Ze Y, *et al.* Behavior robot suite: streamlining real-world whole-body manipulation for everyday household activities. *arXiv* 2025, arXiv:2503.05652.

[118] Heo M, Lee Y, Lee D, Lim JJ. FurnitureBench: reproducible real-world benchmark for long-horizon complex manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, July 10–14, 2023.

[119] Fang HS, Fang H, Tang Z, Liu J, Wang C, *et al.* RH20T: a comprehensive robotic dataset for learning diverse skills in one-shot. In *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, May 13–17, 2024, pp. 653–660.

[120] Luo J, Xu C, Liu F, Tan L, Lin Z, *et al.* FMB: a functional manipulation benchmark for generalizable robotic learning. *arXiv* 2024, arXiv:2401.08553.

[121] Mees O, Hermann L, Rosete-Beas E, Burgard W. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Rob. Autom. Lett.* 2022, 7(3):7327–7334.