Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Review of Paradigm Shifts and Collaborative Evolution in Robot Chassis Design Driven by VLA

Shi Jiacheng
School of Engineering, Ocean University of China, Qingdao, Shandong 266404 China

| ABSTRACT
Vision-Language-Action (VLA) foundation model-driven embodied intelligence is a cutting-edge field at the intersection of artificial intelligence and robotics, endowing robots with powerful semantic understanding and generalization capabilities. However, traditional mechanical body design paradigms are based on assumptions of structured environments and struggle to adapt to the open-task requirements driven by VLA across the dimensions of perception, execution, and interaction. This results in insufficient coordination between intelligent modules and mechanical bodies in open tasks, thereby limiting the overall system performance. This paper systematically reviews the paradigm-shifting challenges posed by the rise of VLA to robotic mechanical body design and explores pathways for their collaborative evolution. First, we dissect the limitations of traditional design in open scenarios; second, we analyze the morphological evolution of the robot body from rigid to compliant and from dedicated to programmable; and third, we elucidate the coordination mechanisms between the intelligent core and the physical shell. This paper aims to provide a theoretical framework for the development of next-generation robot platforms and to advance the transition of general embodied intelligence from theoretical research to practical application.

| KEYWORDS
Vision-Language-Action Model; Robotics; Robot Chassis Design; Embodied Intelligence; Co-evolution; Paradigm Challenge

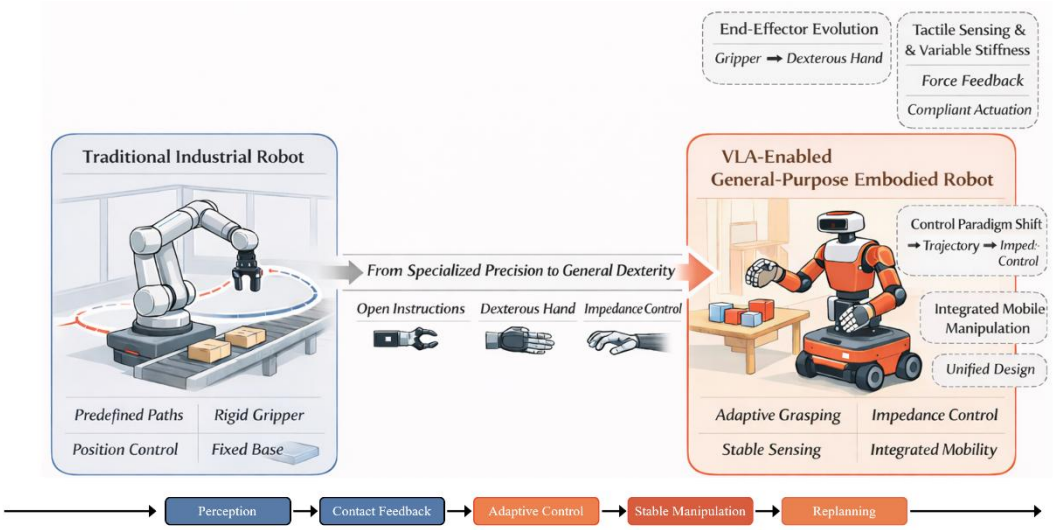
| ARTICLE INFORMATION

ACCEPTED: 05 April 2026 **PUBLISHED:** 16 May 2026 **DOI:** 10.32996/jcsts.2026.8.7.4

1. Introduction

Robotic applications are expanding from structured industrial environments into unstructured living spaces, with their core mission shifting from executing predefined rules to general-purpose collaboration that involves understanding open-ended instructions and safe interaction.[1–3].Foundational model technologies, exemplified by Vision-Language-Action (VLA) models, endow robots with cross-modal semantic understanding, task reasoning, and zero-shot generalization capabilities. By interpreting natural language instructions in conjunction with visual scenes, these models support task planning and end-to-end strategy generation. These models are pre-trained on massive amounts of multimodal data to construct universal representations, which can be transferred to specific scenarios after undergoing lightweight domain adaptation. As shown in Figure 1, this architecture encompasses multimodal input, pre-training, domain adaptation, and task execution, reducing reliance on proprietary data and laying the foundation for the collaborative reconstruction of the intelligent core and physical shell.

Figure. 1. Foundation Model Pretraining and Adaptation for Embodied Robotics.



However, the evolution of artificial intelligence urgently requires the coordinated adaptation of robotic systems. Traditional robotic system designs, which are based on assumptions about structured environments and prioritize precision and efficiency, face fundamental challenges[4]. The demand for VLA-driven systems has highlighted the limitations of existing agents in terms of perception, execution, and interaction: geometric perception systems struggle to provide semantic information about the scene; rigid actuators cannot safely handle unstructured interactions; and one-way command interfaces cannot support multimodal collaboration[5]. Therefore, exploring design innovations and synergy between the intelligent core and the physical shell is a key frontier issue in advancing the practical application of embodied intelligence.

This study aims to review the paradigm-shifting challenges posed by the rise of VLA to robotic body design, as well as pathways for their co-evolution. By analyzing the innovative requirements of VLA in terms of perception, actuation, and interaction interfaces, the study clarifies future directions for body design and emphasizes the necessity of deep coordination between the brain and the body. The research provides a framework for understanding the transformation of mechanical design in the VLA era and lays a theoretical foundation for developing a new generation of robotic platforms adapted to open-world environments, thereby holding significant theoretical value and practical guidance.

2. Basic Concepts and Background

2.1. Key Dimensions and Historical Paradigms in Robot Chassis Design

The design of robotic mechanisms has long adhered to a predefined paradigm characterized by clearly defined functions and structured environments, with the core objective being to optimize precision, efficiency, and reliability for specific tasks[6]. In response to open, unstructured environments and the need for human-robot collaboration, the paradigm of drive and transmission is evolving from a focus on high rigidity toward flexible structures that emphasize adaptability and safe interaction. Soft robotics technology endows the robot body with inherent compliance through novel actuation methods, significantly enhancing the feasibility of interaction with fragile objects and in complex environments[7,8].

The evolution of structural morphology clearly reflects how task requirements shape physical platforms. As shown in Figure 2, from fixed-base robotic arms designed for high precision, to wheeled robots capable of free movement on flat surfaces for efficiency, to legged robots adapted for rugged terrain, this diversity in form directly corresponds to differences in mobility and environmental adaptability. Building on this foundation, quadcopter drones have broken through the limitations of two-dimensional planes to enable three-dimensional operations; humanoid robots achieve seamless compatibility with human tools and environments through anthropomorphic structures; and soft robots enhance safe interaction capabilities in unstructured scenarios through their compliant properties. This evolutionary trajectory reveals that robotic bodies are shifting from single-function, specialized forms toward multimodal, high-degree-of-freedom, general-purpose collaborative forms, providing the physical foundation for general embodied intelligence[9].

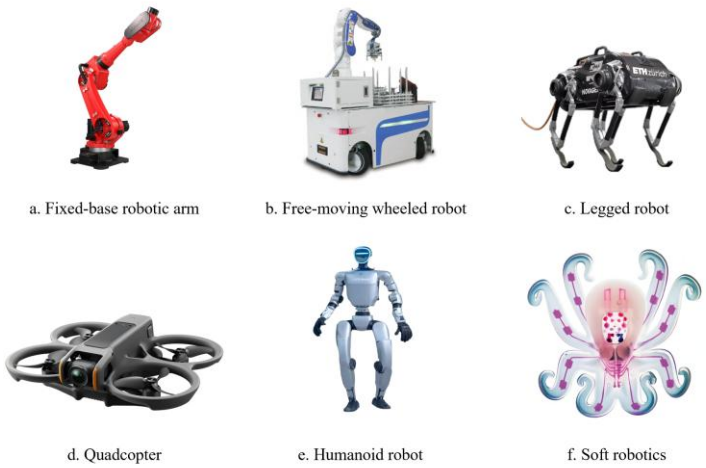
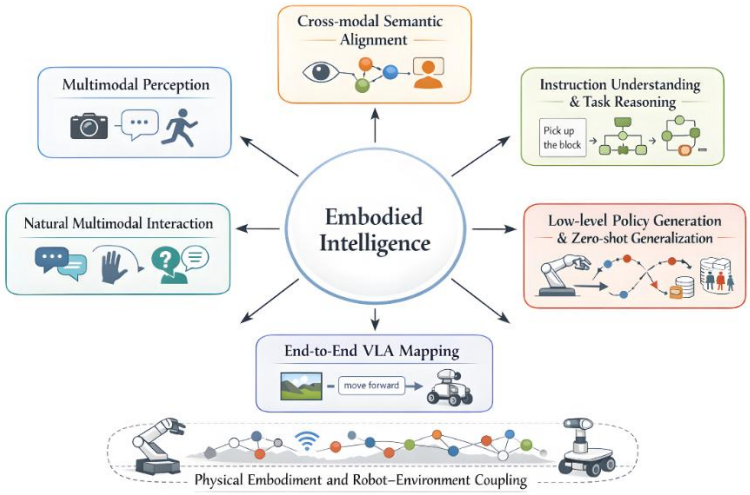


Figure 2. A visualization of the evolution of structural forms

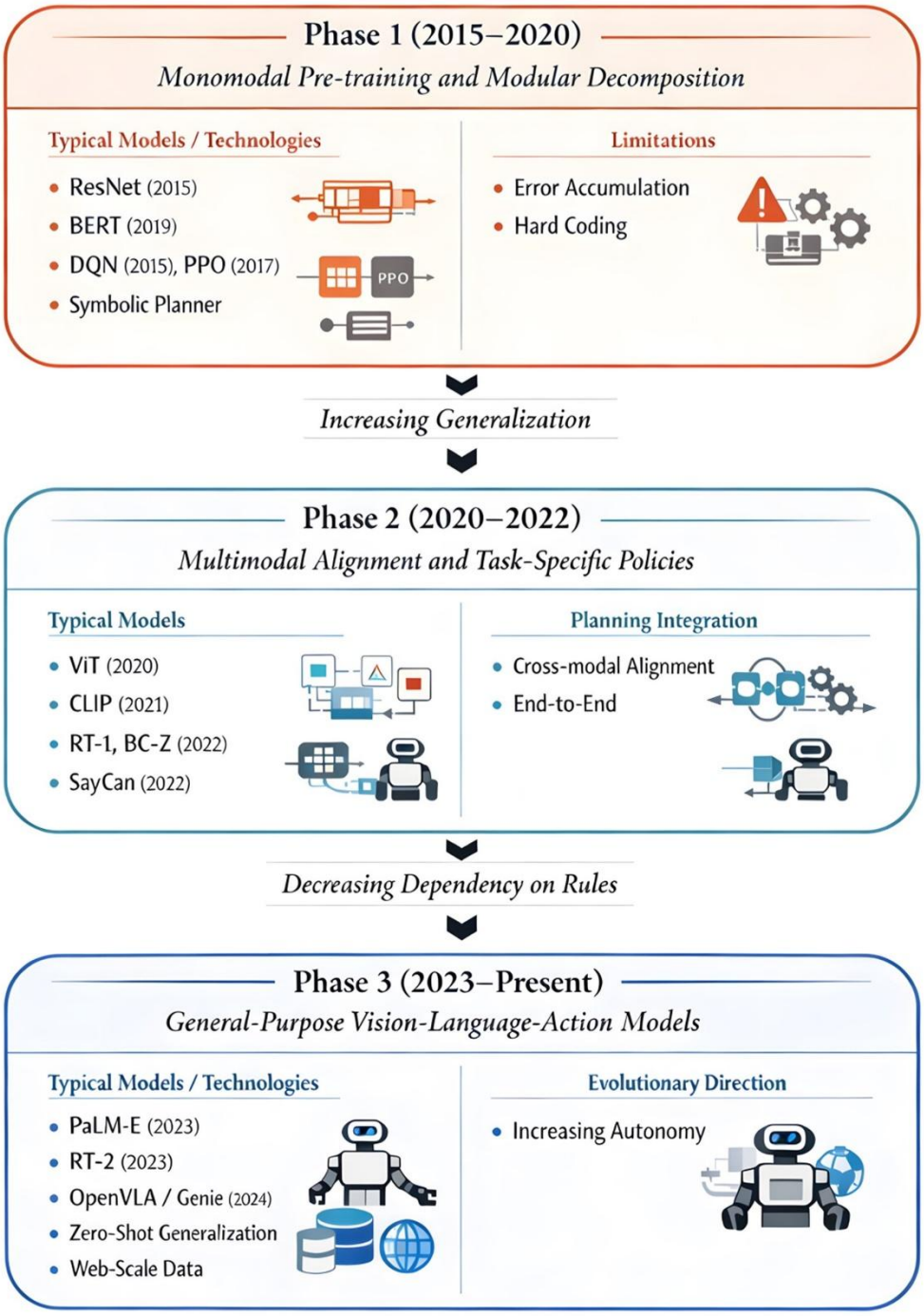
In recent years, composite mobile operation platforms and modular robots have emerged as hot topics in research. The modular concept enables the rapid construction of optimal configurations tailored to different tasks through the reconfiguration of standardized components. This marks a paradigm shift from "dedicated configurations" to "programmable configurations," making the configuration itself an optimizable variable[10,11]. In terms of perception system integration, the evolution has progressed from a single-feedback system to a complex, multi-layered multi-sensor fusion system. The configuration logic has evolved from simple data superposition to a spatiotemporally synchronized system featuring deep integration of hardware and software architectures, with the aim of establishing a unified and robust environmental representation[12,13]. Perception has moved beyond the realm of control inputs to become the foundation for robots to understand and interpret the world.

2.2. An Overview of Vision-Language Large Models and Their Applications in Robotics

Visual-language large models have bridged the gap between single-modal understanding and collaborative cognition through cross-modal semantic alignment, evolving into embodied VLA models that directly map multimodal perceptions into control commands[14–16]. Figure 3 illustrates the core conceptual framework and technological evolution of embodied intelligence.



(a) Concepts



(b) Timelines

Figure 3. Core conceptual framework of embodied intelligence and the evolutionary trajectory from monomodal to VLA

In the field of robotics, the application of the VLA model marks a fundamental paradigm shift. While traditional planning relies on predefined states and symbolic models, the VLA model eliminates complex intermediate representations through command fine-tuning, enabling end-to-end mapping from natural language and visual observations to actions[17]. Second, it serves as a natural multimodal interaction interface, supporting conversational interactions that integrate language, gestures, and ambiguous intentions, while possessing the ability to resolve ambiguities and lower the barrier to use. Finally, it functions directly as a low-level policy generator, achieving zero-shot task generalization through fine-tuning on large-scale demonstration data, and mapping unseen commands to feasible motion policies.

3.Mechanical Chassis Design Driven by VLA Requirements

3.1. The Need for Innovation in Perception Systems

Traditional robot perception focuses on solving geometric problems such as localization and path planning, producing spatially sparse models with sparse semantic information, and exhibiting defensive and reactive characteristics. As VLA models become the core decision-making units, perception requirements have been elevated to a new dimension. VLA must understand fine-grained semantic information, shifting the focus of perception from motion-based obstacle avoidance to cognitive interpretation. The perception paradigm is evolving from traditional geometric perception—which emphasizes map-building and obstacle avoidance—to VLA semantic perception, which encompasses scene graph understanding, object attribute analysis, and active observation. Scene graphs involve object relationships, attributes include material and state information, and active observation involves adjusting the viewpoint[18–20]. This shift from geometric elements to semantic cognition, as illustrated in Figure 4, marks a profound transformation in the design of robotic perception systems.

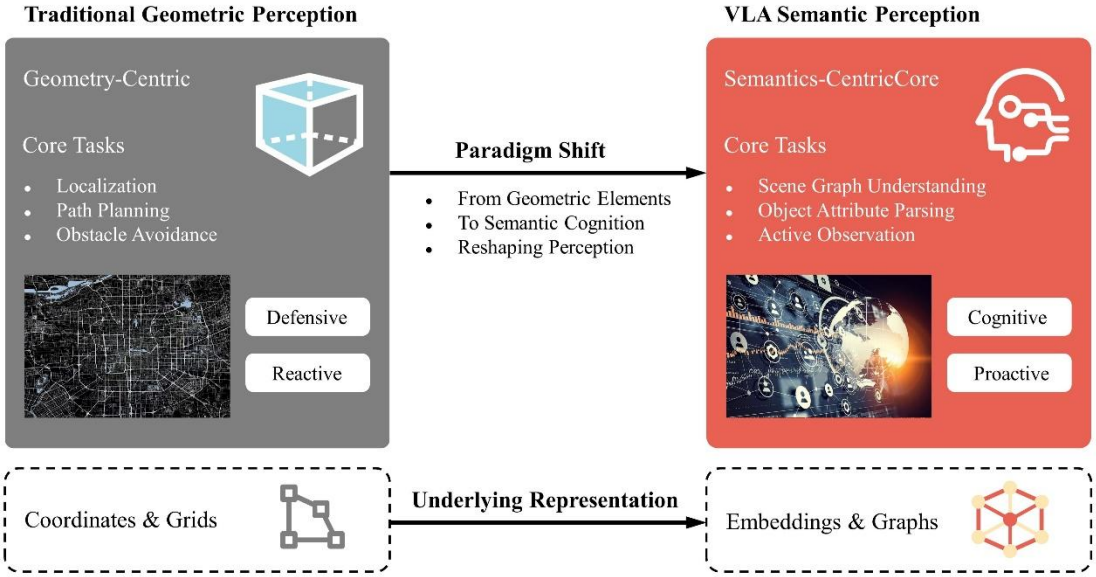


Figure 4. Comparative architecture of traditional geometric perception and VLA semantic perception

This transformation is manifested in three interrelated areas of innovation.

First, VLA visual understanding relies on high-density input. While traditional cameras are sufficient for SLAM, VLA requires the ability to discern subtle spatial relationships, necessitating high resolution to distinguish textures and text, as well as a wide and continuous field of view[21,22]. Second, while traditional multi-sensor fusion primarily focuses on positioning accuracy and geometric modeling, VLA systems require precise spatiotemporal alignment of RGB, depth, thermal, and acoustic data, as well as intensive interaction at the feature level, to generate joint representations rich in physical attributes, with the aim of constructing models that approximate human multisensory perception[23]. Finally, given that VLA reasoning requires focusing on key areas, fixed blind spots in the field of view will limit cognition. The robot's body must be designed with mechanisms that support active perception at the physical level, such as high-degree-of-freedom gimbals to enable both fixation and scanning[24]. The robot must also use its robotic arm to move objects or clear obstructions in order to create new viewpoints, thereby enhancing the coordination between perception and manipulation[25].

3.2. The Need for Innovation in Decision-Making Systems

Traditional robotic decision-making systems are typically based on clearly defined task boundaries, enumerable environmental states, and predefined rules. Their core functionality lies in path selection, action sequencing, and anomaly avoidance based on symbolic states and manually defined logic. However, as the VLA model has emerged as the core cognitive unit in robotic systems, decision-making tasks are no longer limited to search and matching within a finite state space, but have shifted toward semantic understanding, task decomposition, and embodied reasoning in open-world scenarios[26-27]. Robots must not only determine which actions to perform, but also explain why they are performing them, assess their physical feasibility, and dynamically adjust their behavior based on environmental feedback. This indicates that VLA-driven decision-making systems have become a critical link between semantic understanding and physical action, placing higher demands on

the design of robotic hardware. To illustrate this shift from traditional rule-based decision-making to VLA-driven embodied decision-making more intuitively, the core structural differences are shown in Figure 5.

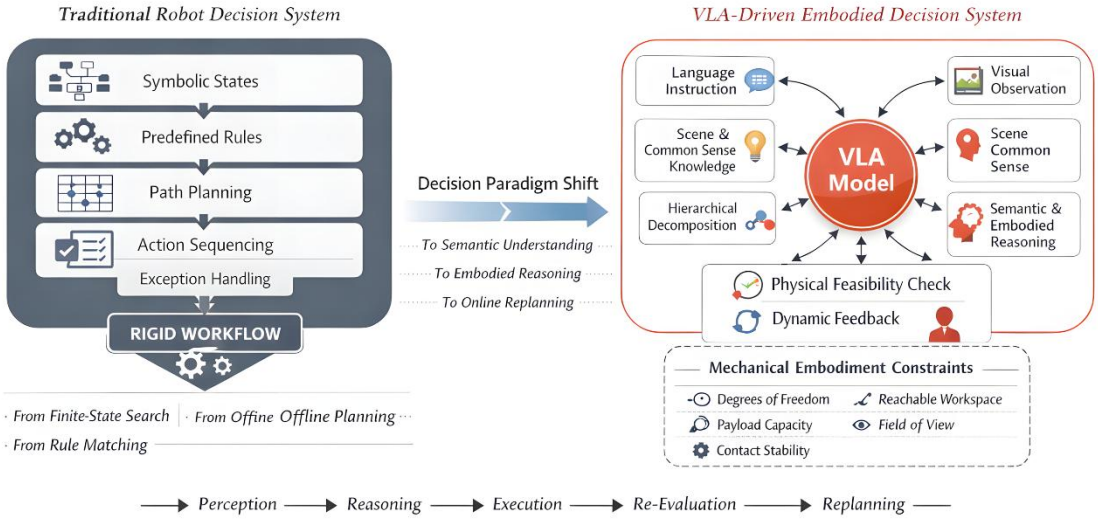


Figure 5. The Paradigm Shift from Traditional Robotic Decision-Making Systems to VLA-Driven Embodied Decision-Making Systems

This change first requires decision-making systems to possess the ability to hierarchically decompose natural language instructions into task structures. When faced with open-ended instructions that are semantically rich and contain implicit constraints, the system must integrate visual observations, scene-specific prior knowledge, and common-sense knowledge to transform high-level goals into executable sequences of subtasks[28]. Second, the decision-making process must be tightly integrated with the robot's physical constraints, incorporating factors such as degrees of freedom, reachable space, payload capacity, field of view, and contact stability into the inference loop to avoid planning results that are semantically valid but physically unfeasible[29]. Finally, the dynamic changes in open environments and the need for real-time human intervention require decision-making systems to possess the ability to continuously reassess and replan in real time, thereby establishing a stable closed-loop between perception, reasoning, and execution[30].

3.3. The Need for Innovation in Execution Systems

Traditional industrial robots rely on high repeatability and reliability to reproduce predefined paths within a structured space[31]. However, the VLA endows robots with the ability to understand open-ended instructions, driving a shift in the execution system from specialized precision toward general dexterity to cope with unknown objects and dynamic, unstructured environments. This evolution in the execution paradigm is reflected not only in enhanced task comprehension but also in a comprehensive restructuring of end-effector execution, contact sensing, control strategies, and system configurations, as illustrated in Figure 6.

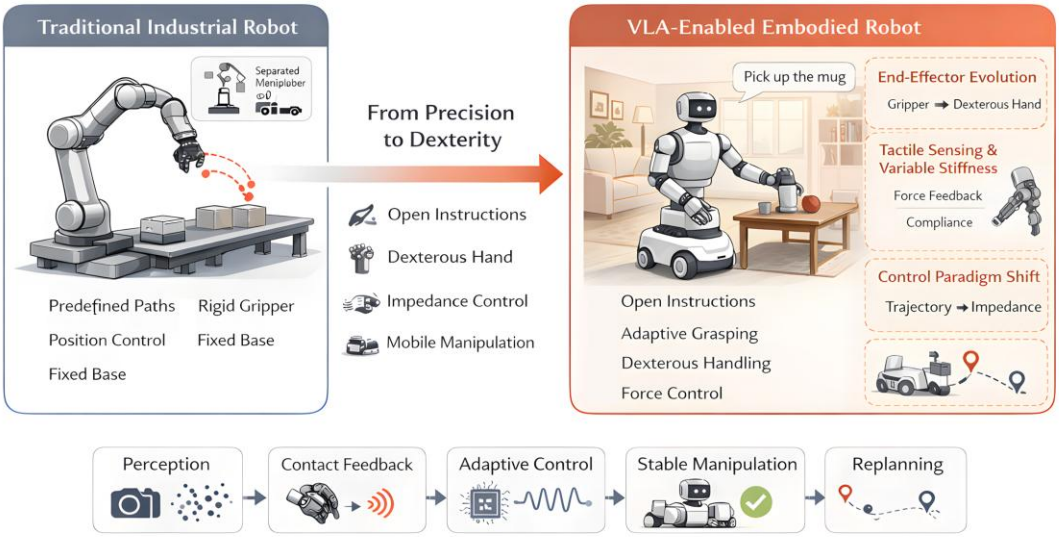


Figure 6: Schematic diagram illustrating the evolution from traditional industrial robots to VLA-enabled general-purpose embodied robot execution systems

End-effector design is evolving from specialized fixtures to biomimetic dexterous hands; traditional designs adhere to the principle of task-specificity[32], Humanoid robots adapt to the diversity of objects through multi-finger, multi-joint structures and, once trained, can predict adaptive grasping configurations. Meanwhile, the integration of tactile sensing with variable stiffness technology enables the robot to perceive contact forces in real time and achieve closed-loop force control, providing passive safeguards for safe interaction[33].

The control paradigm has shifted from trajectory tracking to interactive impedance control, as traditional position control is ill-suited for the non-rigid contact tasks involved in VLA. Robots must possess whole-body coordination and compliance control capabilities, as well as the ability to actively adjust their impedance characteristics[34]. Adaptive variable impedance control dynamically adjusts parameters based on real-time contact information and enhances stability in complex interactive tasks through a unified dynamic perspective[35]. The design of mobile operation platforms has shifted from fixed bases to task-space coordination. Traditional modular designs suffer from system fragmentation, while VLA requirements have driven the development of integrated designs that emphasize the deep integration of mobility and operability. By treating changes in the chassis's pose as controllable degrees of freedom for the manipulator arm, these robots become general-purpose physical agents capable of moving freely within human living spaces.

3.4. The need for innovation in human-computer interaction interfaces

The human-machine interface of traditional industrial robots primarily relies on a teach pendant, requiring operators to master a specialized programming language[36]. The VLA enables robots to understand open-world environments and ambiguous instructions, driving the evolution of interactive interfaces toward multimodal dialogue systems that support bidirectional, real-time, and semantically rich information flows. The design of physical interaction safety must shift from isolation and protection to active co-operation. To address task openness and motion uncertainty, the system must incorporate active dynamic co-operation mechanisms: at the hardware level, it must possess inherent compliance and high-bandwidth force sensing; at the control level, it must integrate real-time collision detection and reaction algorithms[37]. Safety threshold settings are based on biomechanical research findings. In human-robot collaboration scenarios, describing contact risks using a single impact curve is no longer sufficient to meet the safety assessment needs of collaborative robots for multiple body parts. Therefore, it is more appropriate to establish zoned constraints based on the permissible force and pressure limits for different contact areas of the human body. As shown in Figure 7, this diagram presents the tolerance thresholds for different body parts under contact conditions, reflecting the characteristic that sensitive areas have lower permissible values while more resilient areas have relatively higher permissible values[38]. This quantitative relationship provides a basis for establishing safety constraints in the structural design, compliance control, collision detection, and motion planning of collaborative robots. Building on this foundation, the safety mechanism must be further integrated with the VLA decision-making loop to incorporate physical constraints during the task planning phase, thereby forming a full-stack semantic safety barrier that spans from high-level intent to low-level execution.

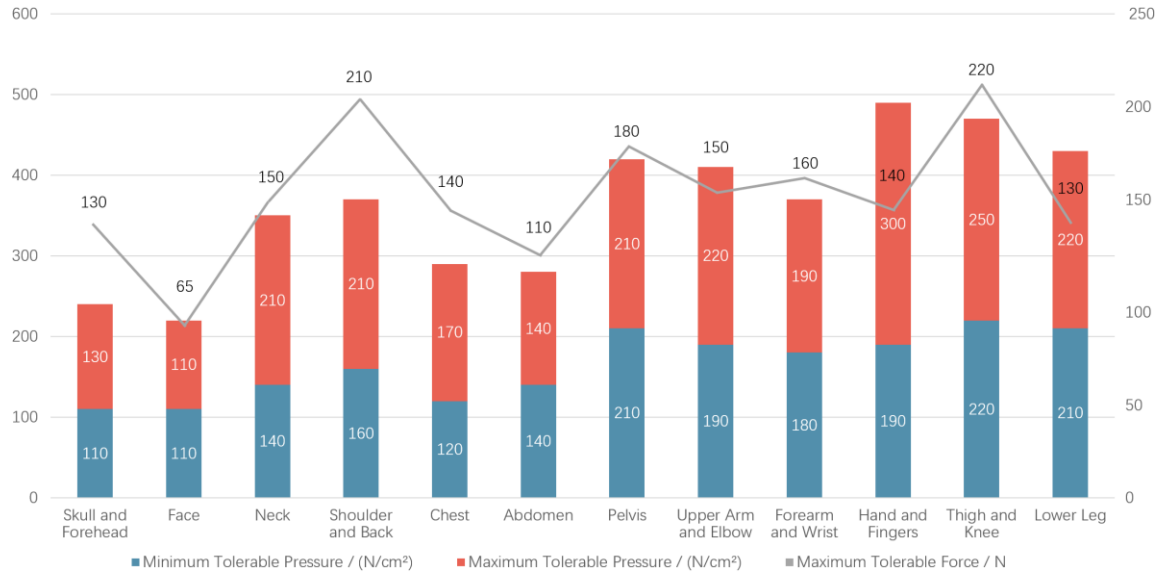


Figure 7. Biomechanical constraints in human-computer interaction

The design of social cue expression marks a paradigm shift in robotics, transforming robots from passive execution tools into active collaborative agents. Driven by VLA, robots must rely on nonverbal social cues to convey intentions, represent states, and build trust, requiring the integration of dedicated social expression modules into the robotic body. The system can intuitively convey the focus of attention through multimodal state indicators, and the configuration of the robotic structure's degrees of freedom not only supports perception but is also endowed with the ability to adopt socially meaningful postures, thereby enhancing the naturalness of interaction[39]. Leveraging the VLA's spatiotemporal sequence analysis capabilities, the robot can predict human movement intentions and adjust its own state or provide proactive assistance[40]. This interactive paradigm elevates the human-computer relationship from a "command-execution" model to one of "collaboration," and its physical foundation lies in the development of interactive interfaces capable of efficiently encoding and decoding rich social cues.

4. VLA Model Adaptation and Training Under Mechanical Constraints

4.1. Domain Adaptation: Incorporating Physical Constraints into VLA

Vision-language large models pre-trained on massive amounts of text and image data from the internet inherently lack physical embodiment in their knowledge representations, which constitutes a fundamental obstacle to transferring VLAs to real robotic systems. Therefore, the primary task of domain adaptation is to embed the robot's kinematic limits, environmental geometric constraints, and interaction dynamics into the VLA architecture in a form that is understandable and inferable. This process unfolds primarily along two key pathways, with the domain adaptation pathway illustrated in Figure 8.

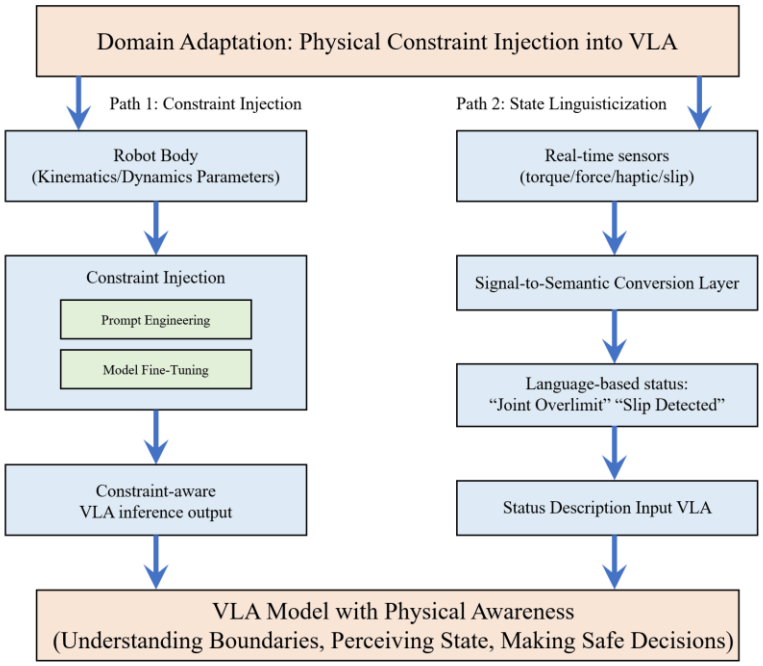


Figure 8. Schematic diagram of the domain adaptation process

Path 1 focuses on kinematic and dynamic constraint encoding, aiming to define the planning space and the boundaries of the physical body to ensure the physical feasibility of motion sequences. This path primarily employs two methods—external constraint injection and internal parameter adaptation—to prevent execution failures caused by ignoring joint limits or velocity thresholds. The former introduces safety constraints through prompt engineering or formal language compilation, guiding the decoding process to satisfy physical constraints[41]; The latter fine-tunes the model based on robot motion data to adapt it to the specific capabilities of the robot[42].

Path 2 focuses on the linguistic encoding of ontological information to achieve an accurate representation of the robot's state. Given the dynamic and time-varying nature of the interaction process, this approach aims to construct a real-time mapping layer from continuous physical signals to discrete semantic symbols, thereby bridging the modal gap between sensor data and the semantic space[43]. By converting key states into textual prompts or structured labels, the model's semantic cognitive capabilities are expanded, extending the foundation of VLA reasoning from purely visual semantics to a physical interaction dimension that encompasses force perception, balance, and contact states, thereby forming an embodied closed-loop of perception–understanding–decision-making.

4.2. Data and Training: Constructing Embodied Data Based on Ontological Capabilities

The performance of embodied intelligence models relies on high-quality embodied interaction data. The data construction paradigm is shifting from domain-specific data collection to an embodied data ecosystem that supports generalized reasoning. The core objective is to bridge the gap between the virtual and real domains and build cross-platform compatible data assets. Current approaches primarily encompass three strategies: high-fidelity simulation, cross-domain real-world datasets, and active exploration interactions. Among these, simulation provides foundational physical prior knowledge, real-world datasets enhance cross-domain generalization capabilities, and active exploration reduces reliance on manual intervention while establishing a closed-loop collaboration between data and models. A detailed comparison of these three approaches is shown in Table 1.

Table 1: Comparison of Primary Sources and Strategies for Embodied Data Construction

Data Construction Approach	Core Methods/Representative Work	Key Characteristics	Contribution to VLA Training
High-fidelity simulation data	Virtual demonstration transfer[44] sim-to-real transfer with meta-learning and	1) Low-cost generation of large-scale task trajectories; 2) Explicit injection of	1) Provide foundational physical priors for pretraining;

	domain adaptation[45]	physical constraints and task structure; 3) Domain adaptation to reduce the sim-to-real gap	2) Improve sample efficiency under limited real data; 3) Enhance transfer robustness from simulation to the physical world
Cross-Entity Real-World Dataset	Open X-Embodiment / RT-X[46]; DROID[47]; XSkill[48]	1) Standardized aggregation across institutions and platforms; 2) Diverse trajectories from heterogeneous robots in real scenes; 3) Learning embodiment-invariant skill or action representations	1) Support cross-platform task understanding; 2) Improve generalization across embodied tasks; 3) Extract physical interaction semantics beyond a single robot ontology
Proactive Exploration and Interactive Data	SimpleVLA-RL[49]; active visual learning dataset and strategy: Zhao et al. (2022)[50]	1) Reinforcement-learning-based or information-driven data acquisition; 2) Reduced dependence on dense manual demonstrations; 3) Closed-loop coupling of exploration, data collection and policy optimization	1) Lower the cost of human supervision; 2) Improve zero-shot and out-of-distribution adaptation; 3) Establish a closed-loop collaboration mechanism between data and models

As a critical component in building a data ecosystem, high-fidelity simulation environments serve as the infrastructure for generating large-scale pre-training data, aiming to provide foundational physical priors at low cost. To address discrepancies between the virtual and real domains, research focuses on domain adaptation and meta-learning strategies, which compensate for perceptual differences by learning domain-invariant latent representations, thereby enhancing transfer efficiency and generalization robustness[45].

Second, open-source, large-scale, cross-agent real-world datasets are key to advancing VLA toward general embodied intelligence. To overcome the limitations of overfitting on a single platform, we construct standardized cross-platform datasets that utilize data from different robotic agents performing the same semantic tasks, thereby forcing VLA to learn abstract action representations and develop cross-agent task understanding capabilities[46-47]. Related research extracts body-independent skill prototypes from heterogeneous robotic video data, marking a paradigm shift in data construction toward extracting the semantics of physical interactions[48].

Finally, data construction mechanisms designed for active interaction and exploration represent a cutting-edge direction, aiming to transform data acquisition from passive imitation into autonomous, goal-driven learning. Equipping robots with mechanisms for safe exploration and autonomous trial-and-error can reduce reliance on human intervention while enhancing their generalization and zero-shot execution capabilities[49]. The body-aware system supports active visual learning, acquiring high-resolution observations by controlling its own movements[50], reflects the collaborative evolution of the physical machine and the intelligent model at the data level[51].

5. Frontiers in Co-evolution: Analysis of Typical Cases

The general-purpose humanoid robot platform embodies embodied intelligence and the principle of morphological matching. The adoption of a humanoid form is intended to align with the distribution of pre-training data, reduce the complexity of cross-body mapping, and enable semantic commands to be directly converted into kinematically valid trajectories. Related

research has demonstrated the implementation of an end-to-end mapping architecture trained on real demonstration data, exhibiting strong generalization capabilities and robustness[52].

However, shape-matching strategies place stringent demands on the mechanical performance of the robot. To achieve generalizable manipulation capabilities, the hardware must possess human-like compliance, whole-body coordination, and high-precision force-tactile feedback capabilities; for tasks involving fine manipulation, compliance impedance control must be implemented at the system level[53]. This demonstrates how algorithmic models are driving hardware design in reverse, and also reflects the trend toward humanoid robots evolving from specialized task-oriented devices into more versatile autonomous platforms[54].

Modular and reconfigurable robots represent another path for the co-evolution of the robot body and its models, with the core concept being to define the physical form as programmable, task-driven, and variable parameters. The mechanical structure is composed of standardized functional modules[55], Its task-oriented configuration planning and rapid reconfiguration capabilities provide the foundation for intelligent models to participate in ontology-level decision-making[56].

In addition, collaborative evolution exhibits multi-level hybrid forms. In the field of mobile operation platforms, the integration of highly agile mobile chassis with general-purpose robotic arms requires VLA to coordinate significantly different motion modalities, generate collaborative dynamic grasping actions, and expand the effective workspace[57]. The modular software framework forms the middle layer of the software architecture, providing standardized interfaces that enable rapid adaptation of models to different hardware platforms and reducing the engineering complexity of system integration[58].

6. Conclusions and Outlook

VLA technology is profoundly reshaping the design paradigm for robotic bodies. This paper systematically reviews the evolving trends in robotics—including perception, actuation, and human-robot interaction—against the backdrop of the rise of VLA, noting that robotic body design has shifted from passive adaptation to single tasks toward an active construction process that evolves in tandem with cognitive models. Specifically, perception systems are increasingly emphasizing semantic understanding and environmental modeling capabilities, while actuation systems are focusing more on compliance, dexterity, and collaborative operation. Human-robot interaction interfaces are also evolving toward multimodal, interpretable, and safe co-existence. Although current research has demonstrated significant potential, challenges remain in areas such as the coordinated optimization of robot structure and model capabilities, the acquisition of embodied data and cross-platform generalization, and the deep integration of physical constraints and safety mechanisms. With the continued advancement of key technologies such as high-fidelity simulation, real-world data acquisition, online adaptive learning, modular reconfigurable design, and trustworthy control, robotic mechanical bodies are expected to achieve further intelligence, autonomy, and generalization, laying the foundation for the large-scale application of embodied intelligent systems in open and complex scenarios.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Alami R, Albu-Schaeffer A, Bicchi A, et al. Safe and dependable physical human-robot interaction in anthropic domains: State of the art and challenges[C]//2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijing: IEEE, 2006: 1-16.

[2] Khan M T, Waheed A. Foundation Model Driven Robotics: A Comprehensive Review[A]. arXiv, 2025. https://arxiv.org/abs/2507.10087. arXiv:2507.10087v1

[3] Kawaharazuka K, Matsushima T, Gambardella A, et al. Real-world robot applications of foundation models: a review[J]. Advanced Robotics, 2024, 38(18): 1232-1254.

[4] Kroemer O, Niekum S, Konidaris G. A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms[J]. Journal of Machine Learning Research, 2021, 22(1):1-82 .

[5] Su H, Qi W, Chen J, et al. Recent advancements in multimodal human–robot interaction[J]. Frontiers in Neurorobotics, 2023, 17: 1084000.

[6] Spong M W, Hutchinson S, Vidyasagar M. Robot Modeling and Control[M]. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2020.

[7] Walker J, Zidek T, Harbel C, et al. Soft Robotics: A Review of Recent Developments of Pneumatic Soft Actuators[J]. Actuators, 2020, 9(1): 3.

[8] Boyraz P, Runge G, Raatz A. An Overview of Novel Actuators for Soft Robotics[J]. Actuators, 2018, 7(3): 48.

[9] Arm P, Zenkl R, Barton P, et al. SpaceBok: A Dynamic Legged Robot for Space Exploration[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada: IEEE, 2019: 6288-6294.
[10] Chung W K, Jeongheon Han, Youm Y, et al. Task based design of modular robot manipulator using efficient genetic algorithm[C]//Proceedings of International Conference on Robotics and Automation: Vol. 1. Albuquerque, NM, USA: IEEE, 1997: 507-512.
[11] Naya-Varela M, Faina A, Duro R J. Morphological Development in Robotic Learning: A Survey[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021, 13(4): 750-768.
[12] Garcia J G, Ortega J G, Garcia A S, et al. Robotic Software Architecture for Multisensor Fusion System[J]. IEEE Transactions on Industrial Electronics, 2009, 56(3): 766-777.
[13] Luo R C, Chang C C. Multisensor Fusion and Integration: A Review on Approaches and Its Applications in Mechatronics[J]. IEEE Transactions on Industrial Informatics, 2012, 8(1): 49-60.
[14] Castrejon L, Aytar Y, Vondrick C, et al. Learning Aligned Cross-Modal Representations from Weakly Aligned Data[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2940-2949.
[15] Changpinyo S, Sharma P, Ding N, et al. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 3557-3567.
[16] Guo W, Wang J, Wang S. Deep Multimodal Representation Learning: A Survey[J]. IEEE Access, 2019, 7: 63373-63394.
[17] Xu Z, Shen Y, Huang L. MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 11445-11465.
[18] Wijayathunga L, Rassau A, Chai D. Challenges and Solutions for Autonomous Ground Robot Scene Understanding and Navigation in Unstructured Outdoor Environments: A Review[J]. Applied Sciences, 2023, 13(17): 9877.
[19] Cadena C, Carlone L, Carrillo H, et al. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age[J]. IEEE Transactions on Robotics, 2016, 32(6): 1309-1332.
[20] Borenstein J, Koren Y. Obstacle avoidance with ultrasonic sensors[J]. IEEE Journal on Robotics and Automation, 1988, 4(2): 213-218.
[21] Zheng S, Wang J, Rizos C, et al. Simultaneous Localization and Mapping (SLAM) for Autonomous Driving: Concept and Analysis[J]. Remote Sensing, 2023, 15(4): 1156.
[22] Brunke L, Zhang Y, Römer R, et al. Semantically Safe Robot Manipulation: From Semantic Scene Understanding to Motion Safeguards[J]. IEEE Robotics and Automation Letters, 2025, 10(5): 4810-4817.
[23] Hu D, Wang C, Nie F, et al. Dense Multimodal Fusion for Hierarchically Joint Representation[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE, 2019: 3941-3945.
[24] Bajcsy R, Aloimonos Y, Tsotsos J K. Revisiting active perception[J]. Autonomous Robots, 2018, 42(2): 177-196.
[25] Zhang W, Wang M, Liu G, et al. Embodied-Reasoner: Synergizing Visual Search, Reasoning, and Action for Embodied Interactive Tasks[A]. arXiv, 2025. https://arxiv.org/abs/2503.21696. arXiv:2503.21696v2.
[26] Zitkovich B, Yu T, Xu S, et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control[C]//Proceedings of The 7th Conference on Robot Learning. PMLR, 2023, 229: 2165-2183.
[27] Ichter B, Brohan A, Chebotar Y, et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances[C]//Proceedings of The 6th Conference on Robot Learning. PMLR, 2023, 205: 287-318.
[28] Huang W, Abbeel P, Pathak D, et al. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents[C]//Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022, 162: 9118-9147.
[29] Garrett C R, Chitnis R, Holladay R, et al. Integrated task and motion planning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2021, 4: 265-293.
[30] Huang W, Xia F, Xiao T, et al. Inner Monologue: Embodied Reasoning through Planning with Language Models[C]//Proceedings of The 6th Conference on Robot Learning. PMLR, 2023, 205: 1769-1782.
[31] Surati S, Hedaoo S, Rotti T, et al. Pick and place robotic arm: a review paper[J]. Int. Res. J. Eng. Technol., 2021, 8(2): 2121-2129.
[32] Causey G C, Quinn R D. Gripper design guidelines for modular manufacturing[C]//Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146): Vol. 2. Leuven, Belgium: IEEE, 1998: 1453-1458.
[33] Romero B, Fang H S, Agrawal P, et al. EyeSight Hand: Design of a Fully-Actuated Dexterous Robot Hand with Integrated Vision-Based Tactile Sensors and Compliant Actuation[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024: 1853-1860.
[34] Jin Z, Qin D, Liu A, et al. Model Predictive Variable Impedance Control of Manipulators for Adaptive Precision-Compliance

Tradeoff[J]. IEEE/ASME Transactions on Mechatronics, 2023, 28(2): 1174-1186.
[35] Liang J, Wang Y, Zhong H, et al. Robust Variable Impedance Control for Aerial Compliant Interaction With Stability Guarantee[J]. IEEE Transactions on Industrial Informatics, 2024, 20(3): 3351-3360.
[36] Jakovljević P, Dihovični Đ, Ratković Kovačević N. Robot Movement Programming for Flexible Cell In "Open Cim Screen"[C]//Proceedings of the International Scientific Conference - Sinteza 2023. Beograd, Serbia: Singidunum University, 2023: 235-241.
[37] Tsai C S, Hu J S, Tomizuka M. Ensuring safety in human-robot coexistence environment[C]//2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. Chicago, IL, USA: IEEE, 2014: 4191-4196.
[38] ISO. ISO/TS 15066:2016 Robots and robotic devices—Collaborative robots[S]. Geneva: International Organization for Standardization, 2016.
[39] Bonarini A. Communication in Human-Robot Interaction[J]. Current Robotics Reports, 2020, 1(4): 279-285.
[40] Wang Z, Wang B, Liu H, et al. Recurrent convolutional networks based intention recognition for human-robot collaboration tasks[C]//2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Banff, AB: IEEE, 2017: 1675-1680.
[41] Wu Y, Xiong Z, Hu Y, et al. SELP: Generating Safe and Efficient Task Plans for Robot Agents with Large Language Models[C]//2025 IEEE International Conference on Robotics and Automation (ICRA). Atlanta, GA, USA: IEEE, 2025: 2599-2605.
[42] Han Z, Gao C, Liu J, et al. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey[A]. arXiv, 2024. https://arxiv.org/abs/2403.14608. arXiv:2403.14608v7
[43] Gao J, Sarkar B, Xia F, et al. Physically Grounded Vision-Language Models for Robotic Manipulation[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024: 12462-12469.
[44] Rahmatizadeh R, Abolghasemi P, Behal A, et al. From Virtual Demonstration to Real-World Manipulation Using LSTM and MDN[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1): 6524-6531.
[45] Bharadhwaj H, Wang Z, Bengio Y, et al. A Data-Efficient Framework for Training and Sim-to-Real Transfer of Navigation Policies[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada: IEEE, 2019: 782-788.
[46] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic Learning Datasets and RT-X Models[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024: 6892-6903.
[47] Khazatsky A, Pertsch K, Nair S, et al. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset[C]//Proceedings of Robotics: Science and Systems XX. Delft, Netherlands, 2024.
[48] Xu M, Xu Z, Chi C, et al. XSkill: Cross Embodiment Skill Discovery[C]//Proceedings of The 7th Conference on Robot Learning. PMLR, 2023, 229: 3536-3555.
[49] Li H, Zuo Y, Yu J, et al. SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning[A]. arXiv, 2025. https://arxiv.org/abs/2509.09674. arXiv:2509.09674v1.
[50] Zhao Q, Zhang L, Wu L, et al. A Real 3D Embodied Dataset for Robotic Active Visual Learning[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6646-6652.
[51] Xu Z, Wu K, Wen J, et al. A Survey on Robotics with Foundation Models: toward Embodied AI[A]. arXiv, 2024. https://arxiv.org/abs/2402.02385. arXiv:2402.02385v1.
[52] Qiu R Z, Yang S, Cheng X, et al. Humanoid Policy ~ Human Policy[A]. arXiv, 2025. https://arxiv.org/abs/2503.13441. arXiv:2503.13441.
[53] Dean-Leon E, Guadarrama-Olvera J R, Bergner F, et al. Whole-body active compliance control for humanoid robots with robot skin[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada: IEEE, 2019: 5404-5410.
[54] Shamsuddoha M, Nasir T, Fawaaz M S. Humanoid Robots like Tesla Optimus and the Future of Supply Chains: Enhancing Efficiency, Sustainability, and Workforce Dynamics[J]. Automation, 2025, 6(1): 9.
[55] Yim M, Duff D G, Roufas K D. PolyBot: a modular reconfigurable robot[C]//Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065): Vol. 1. San Francisco, CA, USA: IEEE, 2000: 514-520.
[56] Seo J, Paik J, Yim M. Modular Reconfigurable Robotics[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2019, 2(1): 63-88.
[57] Zimmermann S, Poranne R, Coros S. Go Fetch! - Dynamic Grasps using Boston Dynamics Spot with External Robotic Arm[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE, 2021: 4488-4494.
[58] Bonci A, Gaudeni F, Giannini M C, et al. Robot Operating System 2 (ROS2)-Based Frameworks for Increasing Robot Autonomy: A Survey[J]. Applied Sciences, 2023, 13(23): 12796.