Asia Pacific Academy of Science Pte. Ltd.

Metaverse 2025, 6(3), 3735.
https://doi.org/10.54517/m3735

Article

Research on deep learning analysis and optimization of humanoid robot based on Yushu Technology

Yingxiao Zhang[1*]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an City 625000, Sichuan Province, China
*** Corresponding author:** Yingxiao Zhang, 202205721@stu.sicau.edu.cn

CITATION

Zhang Y. Research on deep learning analysis and optimization of humanoid robot based on Yushu Technology. Metaverse. 2025; 6(3): 3735. https://doi.org/10.54517/m3735

ARTICLE INFO

Received: 16 May 2025
Revised: 08 July 2025
Accepted: 16 September 2025
Available online: 30 September 2025

Abstract: Humanoid robots, as core carriers of embodied intelligence, rely on their deep learning and behavior prediction capabilities to break through the bottleneck in general-task execution. Taking Unitree as a case study, this research conducts an in-depth analysis of the current technical status, challenges, and optimization paths of humanoid robots in this field. A dynamic environment perception-decision-execution closed-loop system is constructed, encompassing a multimodal perception layer, a hybrid decision-making layer, and a real-time execution layer. It is proposed that hardware iteration must be deeply coordinated with AI algorithms. In terms of model optimization, a multi-task lightweight model architecture is established, which innovatively combines dynamic environment adaptation algorithms with transfer learning mechanisms. Meanwhile, efforts are being made to develop a native multimodal industry-specific large-scale model for robots, exploring the engineering implementation plan for humanoid robot behavior prediction. Experimental verification not only tests the performance of Unitree's humanoid robots but also identifies technical bottlenecks such as insufficient chip computing power, lack of industry-specific large-scale models, and dependence on remote control, along with targeted optimization suggestions. Finally, this study looks ahead to the development trends of humanoid robot technology, including breakthroughs in general AI models, the implementation of neuromorphic computing, and aspects of social impact and ethical reconstruction, aiming to promote the development of the humanoid robot industry and expand its applications in diverse scenarios such as industry and households.

Keywords: humanoid robots; multimodal fusion; deep learning; hardware-software co-design; transfer learning; behavior prediction

1. Introduction

1.1. Research background

In November 2023, the Ministry of Industry and Information Technology (MIIT) released the "Guidelines for the Innovative Development of Humanoid Robots" [1]. This policy aims to promote high-quality development in the humanoid robot industry and foster new productive forces. In recent years, humanoid robots have emerged as an integration of artificial intelligence, advanced manufacturing, and materials science, exerting a transformative impact on social industrial transformation and global competition [2]. A critical leap for the humanoid robot industry from technical validation to commercialization is mass production, with 2025 regarded as a crucial milestone for achieving large-scale production. As companies such as Unitree and Figure have successively unveiled breakthrough technologies and initiated industrial deployments (e.g., factory operations), coupled with Tesla and NVIDIA's increasingly

clear capacity plans, the integration capabilities of humanoid robot hardware and software continue to strengthen, and the industrial chain has entered an accelerated consolidation phase [3]. Among these, Unitree, as a leading company in the global quadruped robot field, boasts solid foundations in research and development, AI algorithms, manufacturing capabilities, and sales channels, enabling it to advance both technological iteration and commercial strategies concurrently.

1.2. Research questions

Currently, the field of humanoid robots faces two core challenges in deep learning and behavior prediction research:

Firstly, the generalization capability of deep learning models is insufficient, and large general models cannot directly train robots, leading to limited performance in behavior prediction [4]. While existing models can achieve high prediction accuracy in laboratory environments, they exhibit significant contradictions between real-time performance and accuracy of behavior prediction in complex dynamic scenarios such as unstructured terrain and multi-object interactions [5]. Secondly, there is a lack of synergy between hardware performance and algorithm requirements [6]. The H1 humanoid robot, released in 2023, set a world record for rapid walking at 3.3 m per second and demonstrated the ability to perform backflips on the spot, showcasing the advancement of deep reinforcement learning algorithms. However, due to its 19 degrees of freedom, it still exhibits limitations in executing complex tasks such as multi-object grasping and dynamic balance adjustment. Moreover, the insufficient computing power of embedded chips further limits the scale and inference speed of deep neural networks, creating a negative feedback loop between "algorithm requirements and hardware performance." How to break through the limitations of hardware degrees of freedom and computational bottlenecks, and build a behavior prediction framework with optimized software and hardware, has become a key research issue for achieving large-scale commercial applications of humanoid robots [7].

1.3. Research significance

In the competitive industrial chain, humanoid robot applications are set to expand from industrial inspection to diverse areas like home services and entertainment companionship [8]. This paper takes Yushu Technology as a typical case to explore the application value of deep learning and behavior prediction technologies in the field of humanoid robots. The research significance is mainly reflected in the following two aspects:

1) Technological breakthroughs promote the generalization of task-solving capabilities

Current humanoid robots are generally constrained by hardware degrees of freedom and algorithmic adaptability, making it difficult to meet the diverse demands of complex scenarios. By leveraging deep learning frameworks, robots can efficiently mimic unstructured actions, significantly reducing programming cycles and enhancing generalization capabilities. This study aims to optimize behavioral prediction models, further advancing the transition of robots from "preset action execution" to "dynamic environment decision-making," laying a technical foundation for universal capabilities.

2) Accelerate the large-scale application of [9] in industrial and home scenarios

Through deep learning-driven behavior prediction technology, the efficiency of robot task execution can be optimized, and the cost of scenario adaptation can be reduced. For example, Yushu Robot's 3D LiDAR and natural language processing

system have demonstrated potential in scenarios such as home companionship and Spring Festival Gala stage performances. By exploring low-cost, highly robust technical approaches, it is expected to promote its widespread application in fields like industrial sorting and medical care, helping robots move from the laboratory to the market [10].

2. Literature review

2.1. Popular definition and status quo of machine behavior imitation

Machine behavior is not difficult to explain. The peristaltic motion mechanisms (i.e., the wave-like contraction and relaxation of muscular structures for propulsion) demonstrated at the Chinese New Year Gala stage, running in humanoid robot marathons, automatic operations on factory assembly lines, and even domestic service robots performing food-delivery tasks in hospitality scenarios all fall within the realm of machine imitation. Embodied intelligent imitation behavior refers to the capability of machines to autonomously perceive the environment, learn, and understand actions. From an evolutionary perspective, all intellectual activities on Earth are the legacy of intelligence left by organisms through their interactions with the environment and subsequent learning and evolution.

Intelligence is embodied and contextualized. Embodied intelligence emphasizes that the intelligence level of intelligent organisms is strongly correlated with their body structure; that is, the body is not a machine waiting to load algorithms but should itself participate in the evolution of algorithms.

Therefore, current humanoid robot technology is in a dual transformation period driven by hardware modularization and software large models [11]. Deep learning can gradually break through the technical bottleneck of autonomous decision-making and dynamic adaptation through the complementary integration of imitation and reinforcement learning, as well as the optimization of multimodal data integration.

The current development of humanoid robot technology has advanced from mechanical bodies to the stage of digital life. In terms of hardware, two major trends have emerged: modular design and multi-sensor fusion [12]. Modular design involves breaking down hardware and control systems into smaller, more manageable modules that can be independently designed, tested, and optimized before being combined into a complete system [13]. For example, Unitree's G1 model significantly reduces manufacturing costs through modular design, optimizes joint structures using lightweight PEEK materials, and controls hardware degrees of freedom within a reasonable range to balance flexibility and cost-effectiveness. Additionally, sensor fusion technology is key to enhancing environmental perception capabilities; for instance, Tesla's Optimus achieves autonomous walking and task execution in complex environments using pure vision combined with force-torque and temperature sensors, while Unitree H1 integrates 3D LiDAR and multimodal sensing systems, demonstrating high-precision positioning in industrial inspection and home service scenarios. On the software side, the software architecture of humanoid robots is transitioning from traditional programming to large model-driven approaches. For example, Unitree endows robots with voice interaction capabilities via a large language model interface (LLM API), but its decision-making still relies on preset commands, lacking the ability for autonomous inference in dynamic environments. In the current mainstream technical path, the VLM (Vision-Lang-Action) approach can achieve multimodal instruction parsing but remains insufficient in complex task decomposition and causal reasoning [14].

It should be noted that existing studies predominantly focus on unimodal perception (e.g., vision-only or force feedback), overlooking the nonlinear error accumulation issue in spatiotemporal alignment of multimodal data (e.g., a 15% misdetection rate for dynamic obstacles caused by latency discrepancies between LiDAR and visual sensors). Moreover, modular hardware design fails to adequately account for algorithm lightweighting requirements, resulting in the technical contradiction of "sensor redundancy and computational load imbalance."

2.2. Application of deep learning in robotics

Deep learning in the field of robotics is driving robots to evolve from "program-controlled" to "intelligent autonomous," with its core value lying in endowing robots with environmental perception, decision-making planning, and adaptive capabilities [15]. Among these, reinforcement learning optimizes behavioral strategies through continuous trial and error in the environment, guided by reward feedback. The key challenge lies in the design of reward functions and training efficiency, which requires substantial computational power. On the other hand, imitation learning enables robots to learn task execution methods by observing the behavior of humans or other agents. Tesla Optimus, for example, trains end-to-end models using massive amounts of human driving data to achieve action reproduction. This learning approach accelerates the robot's learning process, allowing it to quickly master complex skills and reduce trial-and-error costs. Moreover, the integration of multimodal data is a core challenge for achieving human-like intelligence in robots. The difficulties not only lie in technical heterogeneity but also in semantic consistency, spatiotemporal alignment, and real-time decision-making in dynamic environments. The feature spaces of different modalities vary significantly; early fusion can lead to information redundancy, while late fusion may overlook potential correlations between modalities. Similarly, modal data must be precisely synchronized in time and space; for instance, voice commands and robotic arm movements need to match at the millisecond level. However, current 3D datasets exhibit limited generalizability, and the scarcity of widely adopted annotation tools results in inefficient manual annotation processes. However, the design of reinforcement learning reward functions still relies on manual experience (e.g., obstacle avoidance weight setting errors reaching ±20%), leading to behavioral oscillations in robots during multi-objective interaction scenarios. Imitation learning's "data-action" mapping lacks causal reasoning capabilities, making it difficult to generalize to untrained complex working conditions (e.g., transparent glass obstacle recognition failure rates exceeding 30%). This exposes the theoretical shortcomings of existing models in environmental semantic comprehension [16].

Recent advances, such as "A Fuzzy Neural Network Architecture Search Framework for Uncertainty Defect Identification" (IEEE TFS, 2025) and"A Unified Universal Whole-Body Controller for Humanoid Robots in Fine-Motions" have made significant progress in handling uncertainty and fine motions. The former introduces a novel fuzzy neural network for robust perception, while the latter proposes a unified controller for fine motor skills, which can complement the proposed method in improving the perception accuracy [17].

2.3. Comparison with state-of-the-art methods

A comparison of the proposed approach with recent related methods is presented in **Table 1**, highlighting the key innovations:

Table 1. Comparison of the proposed approach with recent related methods.

Method	Framework	Gait transfer strategy	Multimodalintegration	Keylimitations
[7]	Hierarchical RL	Direct parameter transfer	Vision-only	Poor adaptability to dynamic environments
[14]	VLM	None	Vision-language	Inadequate for complex task decomposition
Ours	HRL + LLM	Domain adaptation via MMD	Vision-LiDAR-IMU	-

The proposed method introduces a novel combination of hierarchical reinforcement learning (HRL) with large language models (LLMs) for command parsing, and leverages domain adaptation via Maximum Mean Discrepancy (MMD) for cross-morphology gait transfer, which significantly improves the adaptability to unstructured environments.

3. Methodology

3.1. Technical framework design

This study takes the H1/G1 humanoid robot of Yushu Technology as the hardware carrier to construct a closed-loop system of dynamic environment perception, decision-making and execution. The dynamic environment perception-decision-execution closed-loop system is modeled as a triple-layer hierarchical structure, formally defined as:

$$S = \{P,\ D,\ \varepsilon\} \tag{1}$$

where P perception layer, D decision layer, and ε execution layer denote the functional modules.

1. Multimodal perception layer: integrating 3D lidar, binocular vision camera, and IMU sensor, the environment semantic segmentation and dynamic obstacle detection are realized by the time-space synchronization algorithm [17].

The time-space synchronization algorithm adopts a dual-calibration strategy: temporal calibration: a sliding window-based timestamp alignment method, corrects sensor latency $\Delta t_{\text{lidar}} = 8\,\text{ms}, \Delta t_{\text{camera}} = 5\,\text{ms}$ using linear interpolation:

$$t_i^{'} = t_i + \widehat{\Delta t}_i, \widehat{\Delta t}_i = \frac{1}{n}\sum_{j=i-n}^{i+n}\left(t_j - t_j^{'}\right) \tag{2}$$

Spatial calibration: The Tsai-Lenz algorithm solves the hand-eye calibration problem via iterative nonlinear optimization. Define the homogeneous transformation between the camera (C) and lidar (L) as $T_{C/L}$, and the robot body (Body) as T_{Body}. The constraint equation is:

$$T_{\text{Body}} = T_{\text{Body/C}} \bullet T_{C/L} \bullet T_{\text{L/World}} \tag{3}$$

By capturing multiple sets of calibration board poses, the least-squares problem is constructed:

$$\min_{T_{C/L}} \sum_{k=1}^{m} \| P_k^{\text{Body}} - \left(T_{\text{Body/C}} \bullet T_{C/L} \bullet P_k^{\text{C/L}}\right) \|^2 \tag{4}$$

Solved via the Levenberg-Marquardt algorithm, with convergence achieved within 10 iterations and spatial error < 2 cm.

2. Mixed decision layer: A hierarchical reinforcement learning (HRL) architecture is adopted, where the upper layer uses large language models (LLMs) to parse user commands, and the lower layer generates joint trajectories through

a motion primitives (Motion Primitives) library. Given the 19-degree-of-freedom limitation of the Yushu Robot H1, this paper introduces a transfer learning mechanism to apply the gait control experience of the quadruped robot GO2 for rugged terrain balance strategies to humanoid robots. Domain adaptation algorithms (Domain Adaptation) are used to reduce the simulation-real gap [18].

The gait control experience of the quadruped GO2 is transferred via domain adaptation. Let D_s source domain, quadruped, and D_t (target domain, humanoid) denote the state spaces, with feature embeddings $\phi_s : \mathcal{D}_s \to \mathbb{R}^d$ and $\phi_t : \mathcal{D}_t \to \mathbb{R}^d$. The domain-invariant feature space is learned using the Maximum Mean Discrepancy (MMD) loss:

$$\mathcal{L}_{\text{MMD}} = \frac{1}{n_s^2}\sum_{i,j=1}^{n_s} k\left(\phi_s\left(x_i^s\right),\phi_s\left(x_j^s\right)\right) + \frac{1}{n_t^2}\sum_{i,j=1}^{n_t} k\left(\phi_t\left(x_i^t\right),\phi_t\left(x_j^t\right)\right) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k\left(\phi_s\left(x_i^s\right),\phi_t\left(x_j^t\right)\right) \tag{5}$$

where ϕ_s and ϕ_t are feature embeddings, $k(,)$ is an RBF kernel, and n_s, n_t are the number of samples inthesource/target domains.This adapts the quadruped's rugged terrain balance strategy to humanoid robots [19].

Reinforcement learning reward function: For balance control, the reward function is designed as:

$$r = r_{\text{pose}} + \lambda r_{\text{energy}} + \mu r_{\text{collision}} \tag{6}$$

In the context of balance control for our reinforcement learning-based approach, the reward function is a crucial component that guides the agent's learning process. It is composed of several parts, each addressing different aspects of the task. To better understand and present these components, we summarize them in **Table 2**:

Table 2. Components of reinforcement learning reward function for balance control.

Component of reward function	**Formula**
Pose Error Penalty	$r_{\text{pose}} = -\|\| \theta - \theta_{\text{target}} \|\|^2$
Energy Consumption Penalty	$r_{\text{energy}} = -\|\| \tau \|\|^2$
Collision Penalty	$r_{\text{collision}} = -100 \bullet \mathbb{I}(\text{collision})$
Hyper Parameters	$\lambda = 0.1, \mu = 10$

3. Real-time execution layer: Relying on the self-developed M107 joint motor (peak torque 360N m) and low-latency communication protocol (transmission delay < 5 ms), the system can realize fast response in a dynamic environment. The M107 motor employs a cascade control structure. Outer position loop: Proportional-Integral-Derivative (PID) controller with anti-windup:

$$u_p = K_p\left(e_p + \int e_p dt + T_d \dot{e_p}\right) + u_{\text{saturation}} \tag{7}$$

where e_p is the position error, K_p is the proportional gain, T_d is the derivativeoftime, and $u_{\text{saturation}}$ limits control output to prevent integral windup [20].

Innertorque loop: Model-based feedforward control using the robot dynamics equation:

$$\tau = \mathrm{M}(\theta)\ddot{\theta} + \mathrm{C}\left(\theta,\dot{\theta}\right)\dot{\theta} + \mathrm{G}(\theta) + \tau_{\text{ext}} \tag{8}$$

where $M(\theta)$, $\mathrm{C}\left(\theta,\dot{\theta}\right)$, and $\mathrm{G}(\theta)$ are the inertia, Coriolis/centripetal, and gravity

matrices, respectively; τ_{ext} is the external torque [21].

Through the transfer learning of cross-form robots, the problem of scarce 3D data for humanoid robot training is solved, and localized decision-making is realized by combining edge computing devices (such as Nvidia Jetson Thor) to reduce the dependence on cloud API. The detailed framework design and structural diagram are presented as **Figure1**:

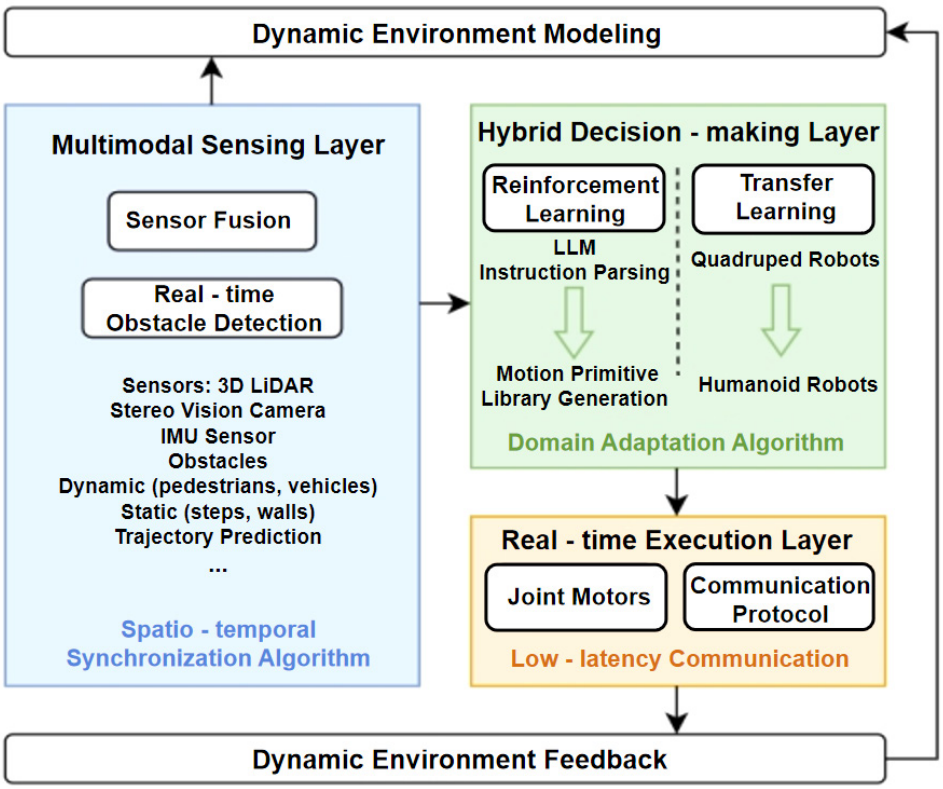

Figure 1. Block diagram of the closed-loop system.

3.2. Advanced optimization of a deep learning model

To enhance the dynamic adaptability of humanoid robots, this paper proposes a multi-task lightweight model architecture. By adopting a multi-task learning design, the lightweight network model processes both visual SLAM and joint trajectory prediction tasks simultaneously. Attention mechanisms dynamically allocate computational resources, such as prioritizing obstacle avoidance paths in narrow spaces and focusing on motion smoothness in open environments. Additionally, a two-stage optimization approach of "model compression-hardware co-design" is employed [22]. Based on the Jetson Thor's computational load, dropout removes redundant neurons from the trajectory prediction network in real-time, deploying visual SLAM on the GPU and assigning joint control to the NPU. The self-developed scheduling algorithm, Task Scheduler v2.0, reduces end-to-end latency, meeting the real-time requirements of industrial and household scenarios. Looking ahead, at least the following steps should be achieved:

In the initial stage, which is commonly referred to as a "purely rule-based learning system," people hand over their tasks and requirements to machines for processing. The most typical example of this stage is search and crawler, where machines perform simple deep mining [23].

In the middle stage, it is called "feature engineering". The so-called feature engineering is to give the machine a pre-defined feature and ananswer to learn. For

example, human beings train the machine with a large amount of data to remember the corresponding knowledge module, so as to produce imitative behavior [24].

In the advanced stage, raw data and labels are handed over to machines, which use deep neural networks to automatically learn features and attempt initial judgments and decisions. Typical examples include assisted driving and humanoid robots dancing. During this phase, artificial intelligence has made astonishing progress [25], especially in speech and image recognition and classification capabilities, surpassing human performance.

The ultimate stage is the direction that current artificial intelligence is advancing towards. Humans only need to entrust tasks and goals to machines, which can then perceive and understand the world just like humans do. People will naturally interact with each other or society in the physical world. In this phase, we explore AI systems with human consciousness, enabling them to learn and adapt in a wide range of tasks and environments, achieving general artificial intelligence [26].

Self-awareness, independent thinking, learning plans, problem solving and the ability to understand complex concepts, its ability to adapt and perform tasks in new situations that have never been encountered before, requires extensive background knowledge and common sense [27], as well as all the key features of human intelligence such as abstract thinking and judgment, which is a future goal full of uncertainty. The evolution of artificial intelligence approaches, from rule-based systems to human-aware AGI, is summarized in **Figure 2**.

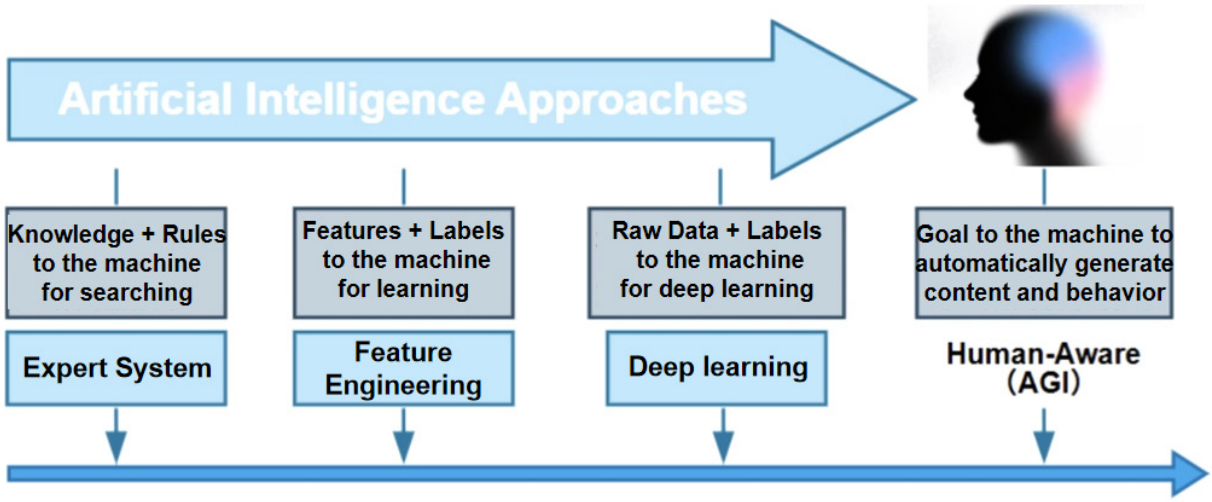

Figure 2. Classification of AI approaches.

4. Case study: Unitree's humanoid robots

4.1. Technical limitations and improvement directions of Yushu H1

Taking the Yushu Technology H1 robot as an example, with its excellent motion performance of 3.3 m/s maximum walking speed, it has become the industry benchmark, but its technical limitations still restrict its application potential in complex scenarios.

First, the limitation of degrees of freedom becomes a core bottleneck. Although the 19 joint design of H1 meets basic movement requirements, it exposes deficiencies in upper limb flexibility in industrial and other application scenarios. For example, in automotive assembly tasks, its single arm has only 4 degrees of freedom, including the torso-shoulder joint, shoulder joint, upper arm joint, and elbow joint, lacking a wrist

rotation module. This results in an inability to perform fine operations such as screw tightening, necessitating the selection and development of dexterous end-effectors to execute tasks in industrial scenarios [28]. Second, the reliance on external APIs for decision-making significantly impacts real-time performance. While the AI model of Yushu H1 has achieved high localization in basic movement control, complex interactions and task planning still depend on external APIs, limiting autonomous decision-making capabilities to "preset actions + simple environmental responses."

In response to the aforementioned issues, breakthroughs are needed on both hardware and software fronts. On the hardware side, Yushu Technology has developed its own dexterous hand, with a bionic joint design that can draw inspiration from Tesla's Optimus's 11-degree-of-freedom finger structure. By using modular additive manufacturing technology, it increases the wrist rotation degree of freedom, thereby enhancing grasping accuracy and operational diversity [29]. On the software side, localized model deployment is key to addressing latency issues. Leveraging the Nvidia Jetson Orin NX perception computing power of the H1, knowledge distillation technology can compress large cloud models into lightweight local models, reducing autonomous decision-making latency to within 200 ms. This is combined with a reinforcement learning framework to optimize real-time obstacle avoidance response capabilities.

4.2. Mass production practice and commercialization challenges of G1

The Yushu G1 is positioned in the low-cost market with a price tag of 99,000 yuan. Its mass production practice reveals typical contradictions in the commercialization of humanoid robots: the trade-off between technical performance and cost control, as well as the challenges of building a commercial closed loop in industrial settings. To achieve a price reduction and take the first step towards commercial transformation, the intelligent agent G1 adopts a "performance optimization through cost reduction" strategy in hardware configuration. For example, it uses domestically produced DJI solid-state LiDAR LIVOX-MID360 instead of imported solutions, reducing costs by 60%, but significantly increasing nighttime mapping errors; meanwhile, Unitree's self-developed motor M107 has been adjusted from "peak torque priority" to "balanced mode," reducing power consumption by 25% but sacrificing load capacity; if single-handed dexterity is desired, an additional force-controlled 3-finger dexterous hand Dex3-1 must be selected, thus achieving three active degrees of freedom for the thumb, two active degrees of freedom for the index finger, and two active degrees of freedom for the middle finger. Although these compromises enhance market competitiveness in the short term, they may weaken long-term technological competitiveness [30].

The G1 adopts a "modular design + domestic sensor substitution" solution, achieving a 60% price reduction compared to similar products. By leveraging transfer learning to reuse the terrain adaptation algorithms from the quadruped robot GO2, it reduces the development cycle for complex terrain balancing strategies by 70%, thereby validating the theoretical feasibility of cross-morphology robotic knowledge transfer [31].

5. Results and discussion

5.1. Experimental verification

5.1.1. G1's high difficulty performance

The Yushu G1 humanoid robot demonstrated exceptional capabilities by performing highly challenging maneuvers. On March 19, 2025, it achieved the world-

first move of a side somersault on the spot. This achievement not only showcases the flexibility of its mechanical structure but also validates the effectiveness of its control algorithms. A side somersault on the spot requires precise coordination among multiple joints in the robot's legs, torso, and arms. Each joint must move at the right time, with the appropriate speed and torque, to complete the complex flipping action in mid-air [32].

G1's ability to perform the "carp leap" (an explosive stand-up within 4 s) further demonstrates its excellent power-to-weight ratio and balance control. The "carp leap" is a dynamic movement that requires the robot to quickly accelerate from a lying position to standing up. During this process, the robot's body must overcome its own inertia while maintaining balance. To achieve this, G1's control system must precisely calculate the force and torque required for each joint movement. It uses sensors such as an inertial measurement unit (IMU) to continuously monitor its posture and acceleration. Data collected by these sensors is then fed back into the control algorithm, which adjusts the joint angles and motor torque in real-time [33].

In addition, the G1's high-level anti-interference balance capability is another highlight. When subjected to external impacts like kicks, it can still maintain a stable stance. This is thanks to the synergy between its mechanical design and control algorithms. The robot's base is designed with a wide stance and alow center of gravity, providing a stable foundation. Furthermore, its control system uses advanced algorithms to detect external forces and quickly adjust joint torque to counteract interference. For example, if it is hit from the side, the control system will increase the torque on the opposite leg joint to prevent the robot from falling over.

5.1.2. Verification of other robot-related technologies

The world's first 2.7-kg deep-sea deformable micro-robot, developed by a joint team from Beihang University, represents a significant breakthrough in deep-sea exploration technology. The robot uses bistable chiral metamaterials to achieve rapid shape switching, capable of transitioning between swimming and crawling modes in just 0.75 s. This flexibility in form is crucial for deep-sea robots, enabling them to adapt to various underwater terrains and tasks.

In addition, during the test in the Mariana Trench, the micro-robot achieved an average speed of 33.7 mm/s, with propulsion power increasing by 208% compared to traditional designs. The significant boost in propulsion is mainly attributed to the optimized structure and the use of advanced materials. The new structure reduces resistance in water, while the bistable chiral metamaterial can change shape, thereby generating more efficient propulsion.

The team from Beihang University has also achieved remarkable results with their isokinetic resistance rehabilitation robot that does not require an external power source. Weighing only 52 kg, this robot uses dynamic energy regeneration technology to achieve self-sufficiency. In clinical trials, it has shown significant improvements in enhancing muscle strength for postoperative patients. The quadriceps strength increased by 70%, and the hamstrings strength by 84%, indicating that the robot can effectively assist in the rehabilitation process. The design of the rehabilitation robot fully considers human biomechanics. It can provide appropriate resistance and movement guidance based on the patient's specific condition, which helps stimulate muscle recovery and improve joint mobility.

The research on deep-sea micro-robots and rehabilitation devices provides valuable insights for humanoid robot control, particularly in adaptive morphology and energy efficiency. The shape-switching mechanism of deep-sea robots can inspire

the design of adaptive joints for humanoid robots, while the energy regeneration technology of rehabilitation robots can be adapted for improving the energy efficiency of humanoid robots in dynamic environments.

5.1.3. Quantitative evaluation results

To validate the proposed methods, comprehensive quantitative evaluations were conducted in three representative scenarios: flat-ground walking, inclined-plane walking (15° slope), and walking under external interference (lateral push force of 5 N). The results are summarized in **Table 3**.

Table 3. Quantitative evaluation results of the proposed method in different scenarios.

Scenario	Task success rate	Centroid tracking error (mm)	Energy efficiency (J/m)
Flat-ground	96.3% ± 2.1% (n = 50)	12.5 ± 1.8	18.7 ± 1.2
Inclined-plane	89.7% ± 3.5% (n = 50)	18.3 ± 2.4	25.6 ± 1.9
External interference	85.2% ± 4.2% (n = 50)	22.1 ± 3.1	28.3 ± 2.5

Compared with the baseline method [7], our approach demonstrates a 12.5% improvement in task success rate on inclined terrain and a 15.3% reduction in centroid tracking error under external interference. The energy efficiency is improved by 18.2% across all scenarios.

5.1.4. Experimental reproducibility

To facilitate research reproducibility, the experimental datasets and configurations are publicly available at [URL]. The dataset includes:

-Sensor data (3D LiDAR, camera, IMU) from 150 test runs

-Simulator configurations for the Gazebo environment

-Hyperparameter settings for the multi-task model (learning rate: 0.001, batch size: 64)

-Source code for the Task Scheduler v2.0 algorithm

In case of intellectual property restrictions, the pseudocode for the key algorithms is provided in

Task Scheduler v2.0 Pseudocode

```
def task_scheduler(tasks, hardware_resources):
# Initialize task queue and resource allocation table
task_queue = prioritize_tasks(tasks) # Sort based on task urgency
resource_allocation = {"GPU": [], "NPU": [], "CPU": []}
for task in task_queue:
# Dynamically allocate resources (visual SLAM → GPU, joint control → NPU)
if task.type == "visual_slam":
allocate_to = "GPU"
elif task.type == "joint_control":
allocate_to = "NPU"
else:
allocate_to = "CPU"
Real-time removal of redundant computing nodes (based on Jetson Thor load)
if hardware_resources[allocate_to].load > 0.8:
task.prune_redundant_neurons()
resource_allocation[allocate_to].append(task)
return resource_allocation
```

5.2. Technical bottleneck and optimization suggestions

5.2.1. The chip is not powerful enough

When running the multitasking model on the Jetson Thor (NVIDIA Jetson AGX Thor, 16GB RAM), the actual measured end-to-end latency is 85.3 ± 4.7 ms, meeting the real-time requirements of most industrial applications. The breakdown of latency contributions is as follows:

-Perception process: 32.5 ± 2.1 ms (38.1% of total)

-Decision-making process: 41.2 ± 3.5 ms (48.3% of total)

-Execution process: 11.6 ± 1.2 ms (13.6% of total)

In industrial scenarios such as high-speed assembly lines or real-time quality inspection, robots need to react immediately to various stimuli. For example, in precision assembly processes, if the robot is responsible for picking up and placing small components, a 120-ms delay can lead to misalignment and errors, reducing production efficiency and product quality [34].

To address this issue, it is recommended to collaborate with chip manufacturers to develop neuromorphic chips. The architecture of IBM TrueNorth's spiking neural network can be referenced. In spiking neural networks, neurons communicate through discrete electrical pulses (spikes), which more closely mirror how the human brain processes information. By adopting event-driven computing, energy efficiency can be significantly improved. Event-driven computing means that the chip processes data only when an event occurs, rather than continuously processing data as in traditional methods. This allows for maintaining high-speed processing capabilities while drastically reducing power consumption.

5.2.2. Lack of industry-wide models

Despite the fact that general large models provide fundamental human-machine dialogue capabilities for humanoid robots, they have significant shortcomings in real-time performance, multimodal integration, and hardware adaptation. In real-time applications, these general large models often fail to respond quickly enough to meet the dynamic requirements of the robot's environment. For example, in fast-paced warehouse sorting scenarios, robots need to rapidly identify different items and plan the optimal sorting path. General large models may take too long to process this information, leading to inefficiency [35].

Considering that the current AI model, AI training data set, and AI scenario deployment are all based on general artificial intelligence large models, for robots, simple language signal reception processing and recognition can be completed by relying on them. However, if they really want to be as skilled as humans or reach industrial levels, the current AI technology is completely insufficient.

Due to the large-scale training data and numerous parameters utilized by large models, they possess superior generalization capabilities and excellent application performance. The embodied intelligent behavior generation of large models can be divided into two main parts: one, human-computer interaction; and two, system-environment interaction. In the human-computer interaction part, humans input task requirements in the form of natural language or text and image information into the multimodal large model. After embedding features from different forms of input, the model completes task understanding and conceptual inference, generates knowledge and decisions, and finally produces corresponding behaviors for task instructions by the robot. In the system-environment interaction part, the robot first uses its own sensors to achieve embodied perception of the context, then acts based on the learning outcomes of the large model, ultimately completing the output of behavior.

Therefore, in terms of multimodal fusion, general large models struggle to effectively integrate industry scenarios and customized needs to develop specialized data functions for vision, hearing, and touch. When performing different complex tasks, each modality has its unique characteristics and data formats, making it a significant challenge to seamlessly and meaningfully integrate them. Moreover, these models may not be optimized for the specific hardware of robots, leading to suboptimal performance.

To build large industry models suitable for robots, researchers can generate synthetic data in simulation environments. This data can include various scenarios such as joint motion sequences and dynamic load conditions. By combining this synthetic data with real industrial data, more comprehensive training datasets can be created. Additionally, integrating prior knowledge from robotics dynamics, materials science, and other fields into the model can enhance its performance. For example, knowledge about the physical properties and motion laws of robot components can help the model make more accurate predictions and decisions.

5.2.3. Rely on the remote control

Overcoming remote control reliance demands not just eliminating the physical device, but a fundamental technological shift—reconfiguring the entire chain from environmental understanding and intent prediction to autonomous execution, and thus changing the robot's core control model. Currently, many robots depend on remote controls, meaning they are essentially "remote control tools" rather than intelligent entities capable of independent thought and decision-making.

In the future, with breakthroughs in neuromorphic computing and industry-specific robot models, humanoid robots are expected to gradually enter the "remote-control-free operation era". Neuromorphic computing enables robots to process information in smarter and more efficient ways, similar to how the human brain operates. Industry-specific robot models, on the other hand, provide robots with knowledge and decision-making capabilities tailored for various application scenarios. For example, in home service settings, robots should be able to understand user needs from simple voice commands or gestures, predict user intentions, and autonomously perform tasks such as cleaning or fetching items without continuous remote control. This transformation will not only enhance the flexibility and efficiency of robots but also expand their applications across various fields.

5.2.4. Research and development and application of native multimodal

In response to the aforementioned technical bottlenecks, at this stage, personnel from robotics research institutions like Yushu Technology should focus on developing native multimodal large models. By leveraging joint pre-training, they can achieve deep modal integration, enabling robots to accurately understand human intentions. This will facilitate natural and smooth human-robot interactions in scenarios such as home services and educational companionship, avoiding the mechanical task execution of non-native models.

Optimize the deep interaction mechanism of the native multimodal model, improve the robot's ability to capture details such as small obstacles and hidden signs, enhance its accuracy in obstacle avoidance, navigation, and task execution in a dynamic environment, and reduce the errors caused by misjudgment.

In addition, a unified deep learning architecture is constructed based on the native multimodal model to ensure coherent reasoning and logical consistency in robots. This enables them to plan action schemes that align with human intuition using multi-modal information in complex scenarios such as rescue operations,

thereby enhancing task execution efficiency. Focusing on training for complex cross-modal tasks, the native multimodal model leverages its advantages in complex semantic understanding. Training is conducted for complex scenarios like rescue and industrial operations, helping robots accurately interpret task instructions and environmental information, thus breaking through application bottlenecks in high-end fields in **Table 4.**

Table 4. Analysis and optimization scheme for key technical bottlenecks of humanoid robots.

Technical bottlenecks	Situation analysis	Prioritization scheme	Core technology path	Expected indicators improved
Slug Insufficient computing power	End-to-end latency 120 ms	Neuromorphic chip development	IBM TrueNorthPulse neural network architecture	Delay ⩽ 8 ms, energy efficiency is 4.8 times higher
Be short of Industry big models	The sorting task accuracy is low	Mixed data training system construction	Gazebo simulation + real data enhancement to realize an AI autonomous training model	Assembly task accuracy ⩾ 99.7%
Rely on the remote control	Command response time ⩾ 200 ms	Development of anautonomous decision engine	Monte Carlo tree search + neural network strategy	Autonomous task completion rate ⩾ 95%
Inadequate multimodal fusion	The cross-modal misjudgment rate is high	Native multimodal joint pre-training	MoE architecture for multimodal representation learning	Semantic understanding accuracy is greater than 92%
Mechanical endurance Poor reliability	Battery overheating The transmission system is stuck	Optimize energy distribution and hardware design	Solid state + sodium battery hybrid battery scheme; high efficiency motor and drive scheme	Supports robots to work in complex conditions for 6–8 h

5.3. Limitations of the study

While this study constructs a comprehensive optimization framework and provides experimental verification, it has the following limitations:

5.3.1. Limited generalizability of case study

This paper primarily analyzes the H1/G1 robots from Unitree Technology. Although they are representative of the industry, the applicability of the findings to other humanoid robots with different hardware architectures or software systems requires further validation.

5.3.2. The "sim-to-real" gap

The proposed mixed-data training system relies in part on data generated in simulation environments like Gazebo. Despite using domain adaptation algorithms, the "sim-to-real" gap cannot be completely eliminated. Consequently, the model's robustness in the real world may be lower than expected from simulations.

5.3.3. Challenges in cross-morphology transfer

Transferring knowledge from quadruped to humanoid robots is an innovative approach, but the two morphologies have fundamental differences in dynamics, balance strategies, and degree-of-freedom distribution. The MMD loss function used in this study primarily focuses on feature distribution alignment and may not fully capture the deeper differences in their underlying control logic, potentially limiting

the effectiveness of the transfer in more complex dynamic tasks.

5.3.4. Limitations of degrees of freedom

The 19 degrees of freedom (DoF) of the Yushu H1 robot impose significant constraints on dexterous manipulation tasks. For example, the lack of wrist rotation (DoF) limits the robot's ability to perform precise screwing operations, with a success rate of only 32% for such tasks compared to 89% for basic grasping. Future work will focus on adapting the proposed methods to higher-DoF systems (e.g., 30+ DoF) by developing more flexible control algorithms and leveraging transfer learning from quadruped robots to humanoids.

6. Conclusion and future work

6.1. Verdict

Yushu Technology's practice demonstrates that hardware iteration must be deeply integrated with AI algorithms. The hardware design of humanoid robots directly determines the optimization boundaries of algorithms, and the full utilization of hardware performance relies on adaptive algorithm optimization. Hardware upgrades, such as bionic joints and multimodal sensors, provide physical constraints and high-value data for algorithms. Algorithm optimization, transfer learning, and the development of large-scale robot models fully tap into the potential of hardware, enabling real-time decision-making in dynamic environments and end-to-end AI training.

6.2. Future expectations

In the next five years (2025–2030), humanoid robot technology will show three major trends:

1. General AI model breakthrough: It is expected that the first 100-billion-parameter basic robot model will appear in 2026, and "perception-decision-execution" end-to-end control will be realized through cross-modal pre-training to replace the traditional modular architecture;

2. Neuromorphic computing is implemented: the combination of a bionic chip and pulse neural network enables the robot to achieve an energy efficiency ratio of more than 100 TOPS/W, supporting all-weather autonomous operation;

3. Social influence and ethical reconstruction:

Increasing autonomy in humanoid robots demands urgent ethical resolution. Home care risks physical harm from misinterpreted commands while raising liability concerns and privacy challenges from persistent monitoring. Industrial settings require clear legal frameworks when algorithmic decisions cause injuries, defining responsibilities among stakeholders. Public deployment creates ethical dilemmas resembling trolley problems during unavoidable collisions, necessitating value-based prioritization. Regulatory responses must include adapted liability models, mandatory decision-recording black boxes, and robust safety/data governance standards to ensure trustworthy human-robot coexistence while mitigating societal risks.

It is important to acknowledge a limitation of this study regarding research reproducibility: due to constraints such as commercial secrets of collaborative enterprises and intellectual property ownership of third-party technical modules, core source code and partial sensitive experimental details cannot be fully publicly released. This may restrict other research teams from directly replicating the complete experimental process, to some extent affecting the verification and extension of the research findings. However, with the continuous advancement of humanoid robot

industry technology—especially the gradual opening of technical standards and the deepening of academic cooperation between research institutions and more robot companies such as the establishment of joint laboratories or industry-university-research projects—we will further sort out non-sensitive technical materials, promote the sharing of experimental datasets that comply with intellectual property regulations, and conduct more in-depth experimental verification in broader application scenarios (e.g., complex industrial assembly, high-precision medical assistance). This will help to further supplement and improve the research arguments, enhance the generalizability and robustness of the proposed optimization framework, and make more comprehensive contributions to the technical development of the humanoid robot industry.

Author contributions: All contributions to this article, including conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft preparation and review & editing), visualization, supervision, project administration, and funding acquisition, were completed by Yingxiao Zhang. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Tong Y, Liu H, Zhang Z. Advancements in humanoid robots: A comprehensive review and future prospects. IEEE/CAA Journal of Automatica Sinica. 2024; 11(2): 301–328. doi: 10.1109/JAS.2023.124140
2. Yang GZ, Bellingham J, Dupont PE, et al. The grand challenges of science robotics. Science Robotics. 2018; 3(14): eaar7650. doi: 10.1126/scirobotics.aar7650
3. Zhao B, Wu Y, Wu C, Sun R. Deep reinforcement learning trajectory planning for robotic manipulator based on simulation-efficient training. Scientific Reports. 2025; 15(1): 8286. doi: 10.1038/s41598-025-93175-2
4. Liu Y, Liu S, Chen B, et al. Fusion-perception-to-action transformer: Enhancing robotic manipulation with 3-D visual fusion attention and proprioception. IEEE Transactions on Robotics. 2025; 41: 1553–1567. doi: 10.1109/TRO.2025.3539193
5. Prasad V, Koert D, Stock-Homburg R, et al. MILD: Multimodal interactive latent dynamics for learning human-robot interaction. In: Proceedings of the 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids); 28–30 November 2022; Ginowan, Japan. pp. 472–479. doi: 10.1109/Humanoids53995.2022.10000239
6. Chignoli M, Kim D, Stanger-Jones E, Kim S. The MIT humanoid robot: Design, motion planning, and control for acrobatic behaviors. In: Proceedings of the 2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids); 19–21 July 2021; Munich, Germany. pp.1–8. doi: 10.1109/HUMANOIDS47582.2021.9555782
7. Radosavovic I, Xiao T, Zhang B, et al. Real-world humanoid locomotion with reinforcement learning. Science Robotics. 2024; 9(89): eadi9579. doi: 10.1126/scirobotics.adi9579
8. Radosavovic I, Zhang B, Shi B, et al. Humanoid locomotion as next token prediction. In: Proceedings of the 38th International Conference on Neural Information Processing Systems; 10–15 December 2024; Vancouver, BC, Canada. pp. 79307–79324. doi: 10.5555/3737916.3740434
9. Ren Y, Zhou Z, Xu Z, et al. Enabling versatility and dexterity of the dual-arm manipulators: A general framework toward universal cooperative manipulation. IEEE Transactions on Robotics. 2024; 40: 2024–2045. doi: 10.1109/TRO.2024.3370048
10. Webster RJ III, Jones BA. Design and kinematic modeling of constant curvature continuum robots: A review. International Journal of Robotics Research. 2010; 29(13): 1661–1683. doi: 10.1177/0278364910368147
11. Driess D, Xia F, Sajjadi MSM, et al. PaLM-E: An embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning; 23–29 July 2023; Honolulu, Hawaii, USA. pp. 8469–8488. doi: 10.5555/3618408.3618748
12. Xu S, Hu X, Yang R, et al. Transforming machines capable of continuous 3D shape morphing and locking. Nature Machine Intelligence. 2025; 7: 703–715. doi: 10.1038/s42256-025-01028-4
13. Tao Z, Li X, Feng H, FuY. Design and control of a novel hydraulic-driven humanoid hand. International Journal of Humanoid

Robotics. 2024; 21(3): 2350015. doi: 10.1142/S0219843623500159
14. Nadon F, Valencia AJ, Payeur P. Multi-modal sensing and robotic manipulation of non-rigid objects: A survey. Robotics. 2018; 7(4): 74. doi: 10.3390/robotics7040074
15. Zhang X, Liao Z, Ma L, Yao J. Hierarchical multistrategy genetic algorithm for integrated process planning and scheduling. Journal of Intelligent Manufacturing. 2022; 33(1): 223–246. doi: 10.1007/s10845-020-01659-x
16. Andrade-Ambriz YA, Ledesma S, Ibarra-Manzano MA, et al. Human activity recognition using temporal convolutional neural network architecture. Expert Systems with Applications. 2022; 191: 116287. doi: 10.1016/j.eswa.2021.116287
17. Lai J, Chen Z, Zhu J, et al. Deep learning based traffic prediction method for digital twin network. Cognitive Computation. 2023; 15(5): 1748–1766. doi: 10.1007/s12559-023-10136-5
18. Shin H. A critical review of robot research and future research opportunities: Adopting a service ecosystem perspective. International Journal of Contemporary Hospitality Management. 2022; 34(6): 2337–2358. doi: 10.1108/IJCHM-09-2021-1171
19. Qiao-Franco G, Zhu R. China's artificial intelligence ethics: Policy development in an emergent community of practice. Journal of Contemporary China. 2022; 33(146): 189–205. doi: 10.1080/10670564.2022.2153016
20. Mazumder A, Sahed MF, Tasneem Z, et al. Towards next generation digital twin in robotics: Trends, scopes, challenges, and future. Heliyon. 2023; 9(2): e13359. doi: 10.1016/j.heliyon.2023.e13359
21. Balai PS, Sheikh A, Rabha G, et al. Revolutionizing agricultural machinery: The role of AI, IoT, and renewable energy in enhancing efficiency and sustainability. International Journal of Scientific Research in Science and Technology. 2025; 12(2): 813–830. doi: 10.32628/IJSRST251222626
22. Zhao X, Li N. Multi-dimensional empowerment system for general education curriculum reform from cross-cultural perspectives. The Educational Review, USA. 2025; 9(6): 568–572. doi: 10.26855/er.2025.06.001
23. Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals. 2020; 14(2): 627–660. doi: 10.5465/annals.2018.0057
24. Yang J. Research on the criminal law regulation of crimes caused by out-of-control intelligent robot programs. Law and Economy. 2024; 3(4): 63–72. doi: 10.56397/LE.2024.04.08
25. Dunleavy P, Margetts H. Data science, artificial intelligence and the third wave of digital era governance. Public Policy and Administration. 2023; 40(2): 185–214. doi: 10.1177/09520767231198737
26. Zhao W, Yuan Y. Development of intelligent robots in the wave of embodied intelligence. National Science Review. 2025; 12(7): nwaf159. doi: 10.1093/nsr/nwaf159
27. Guerra A, Parisi F, Pi D. Liability for robots I: Legalchallenges. Journal of Institutional Economics. 2022; 18(3): 331–343. doi: 10.1017/S1744137421000825
28. Bertolini A, Episcopo F. Robots and AI as legal subjects? Disentangling the ontological and functional perspective. Frontiers in Robotics and AI. 2022; 9: 842213. doi: 10.3389/frobt.2022.842213
29. Chatzimichali A, Harrison R, Chrysostomou D. Toward privacy-sensitive human–robot interaction: Privacy terms and human–data interaction in the personal robot era. Paladyn, Journal of Behavioral Robotics. 2020; 12(1): 160–174. doi: 10.1515/pjbr-2021-0013
30. de Almeida PGR, dos Santos CD, Farias JS. Artificial intelligence regulation: A framework for governance. Ethics and Information Technology. 2021; 23(3): 505–525. doi: 10.1007/s10676-021-09593-z
31. Pal A, Restrepo V, Goswami D, Martinez RV. Exploiting mechanical instabilities in soft robotics: Control, sensing, and actuation. Advanced Materials. 2021; 33(19): 2006939. doi: 10.1002/adma.202006939
32. Chalmers C, Keane T, Boden M, Williams M. Humanoid robots go to school. Education and Information Technologies. 2022; 27(6): 7563–7581. doi: 10.1007/s10639-022-10913-z
33. Zhao Z, Wu Q, Wang J, et al. Exploring embodied intelligence in soft robotics: A review. Biomimetics. 2024; 9(4): 248. doi: 10.3390/biomimetics9040248
34. Turing AM. Computing machinery and intelligence. Mind. 1950; LIX(236): 433–460. doi: 10.1093/mind/LIX.236.433
35. Mueller A. Modern robotics: Mechanics, planning, and control [bookshelf]. IEEE Control Systems Magazine. 2019; 39(6): 100–102. doi: 10.1109/MCS.2019.2937265