Research on Innovation of Robot Environment Perception and Behavior Planning Enabled by Multimodal AIGC Technology

Yuting Wang[1,*], Haifeng Wang[2]

[1]Faculty of Science and Technology, University of Macau, Macau, China

[2]Faculty of Computer Science, Guangzhou Institute of Applied Science and Technology, Guangzhou, China

*Corresponding author: 15843827931@163.com

Keywords: Multimodal AIGC; Robot Environmental Perception; Behavior Planning; Generative Model; Deep Learning; Data Fusion; Transfer Learning; Intelligent Robots

Abstract: This paper focuses on the innovative application of multimodal artificial intelligence-generated content (AIGC) in robot environmental perception and behavior planning. By systematically combining the theoretical framework and evolution path of multimodal AIGC technology, its transformative effect on traditional robot perception and planning methods is deeply analyzed. Studies have shown that multimodal AIGC technology effectively solves the perception limitations of robots in complex environments and improves planning efficiency through data fusion, generative model construction, and transfer learning optimization. Combined with the actual cases of intelligent warehousing robots and guide robots, this study verifies the remarkable effectiveness of this technology in improving the robot's environmental understanding ability and decision-making intelligence. It provides new theories and practices for the development of intelligent robots.

1. Introduction

With the rapid development of artificial intelligence technology, robots are increasingly used in industrial production, public services, and emergency rescue. However, traditional robots are limited by single-modal perception technology and classic behavior planning algorithms, and have low adaptability and decision-making ability in complex dynamic environments. For example, in warehousing and logistics scenarios, robots that rely on lidar are prone to misjudgment of environmental information when blocked by goods. Robots based on traditional path planning algorithms find it difficult to quickly generate the optimal action plan when faced with sudden obstacles. Multimodal AIGC technology integrates multiple data sources such as text, images, and audio. It uses advanced models such as generative adversarial networks and Transformers to build an environmental perception and decision-making system with more human cognitive characteristics, providing key technical support for the intelligent upgrade of robots. Exploring innovative ways for multimodal AIGC technology to empower robots has important theoretical and practical significance for promoting the development of robot technology and expanding application scenarios. This study aims to reveal the internal mechanism of multimodal AIGC technology to empower robots. Through technical principle analysis, current situation analysis, and application case verification, provide theoretical support and practical paths for the intelligent upgrade of robots, and help them be deeply applied in industrial manufacturing, public services, and emergency rescue [1].

2. Overview of Multimodal AIGC Technology

2.1 Theoretical Framework and Technical Evolution of Multimodal Data Fusion

Multimodal data fusion aims to integrate heterogeneous data such as images, voices, and sensor signals to build a comprehensive environmental cognition model. Its theoretical basis can be traced back to the theory of "cross-channel perception integration" in cognitive science - the human brain achieves a robust understanding of complex environments by integrating multi-channel information such as vision, hearing, and touch [2]. Inspired by this, multimodal data fusion technology solves the

  DOI: 10.25236/iwmecs.2025.048

semantic gap and spatiotemporal alignment problems between modalities through mathematical modeling and algorithm design. Its technical evolution can be divided into three stages:

The first is the early feature splicing and statistical learning stage (2000-2010). This stage mainly uses linear algebra methods to achieve feature-level fusion. Serial fusion directly splices feature vectors of different modalities to form a high-dimensional feature space. For example, in early robot navigation, the three-dimensional coordinates of the lidar were spliced with the visual edge features and input into the support vector machine for obstacle classification, with an accuracy rate of 12% higher than that of a single modality [3-4]. Statistical fusion based on probabilistic graphical models uses tools such as Bayesian networks to describe the dependencies between modalities. In medical robots, the Bayesian network is used to fuse ultrasound images and force feedback data to identify the type of tissue contacted by surgical instruments, with an accuracy rate of 83%. However, linear models are difficult to capture the complex nonlinear relationships between modalities.

The second is the semantic fusion stage driven by deep learning (2011-2020). The rise of deep learning has pushed multimodal fusion into the semantic stage. The joint embedding model uses encoders to map multimodal data to a shared semantic space. For example, the FVQA model uses CNN and LSTM to encode images and questions respectively, and generates a joint embedding vector to achieve image question answering, with an accuracy rate of 72%. In robot command understanding, similar architectures can improve the accuracy of command execution to 89%. The self-attention mechanism of the Transformer architecture has become the core of cross-modal interaction. In multi-sensor target tracking, the weights of each modality can be dynamically adjusted according to the environment, reducing the target tracking error by 41% in complex environments [5].

The third is the generative fusion stage driven by AIGC (2021 to present). At present, active fusion is achieved with the generative model as the core. GAN can be used to fill in missing modal data. For example, when the tactile sensor fails, CycleGAN generates a virtual tactile signal, which restores the robot's grasping success rate from 58% to 87%. Causal reasoning fusion based on causal Bayesian network can mine the causal relationship between modalities. In autonomous driving, the connection between rainy and foggy weather and extended braking distance can be identified in advance, and braking can be triggered 500ms in advance, shortening the braking distance by 15%.

2.2 Core Theory and Technological Breakthroughs of AIGC Generative Model

The AIGC generation model simulates the human cognitive process and gives the robot the ability to model the environment and optimize decisions. The core technologies are as follows:

Generative Adversarial Network (GAN) dual game theory: GAN is based on zero-sum game, and fits data distribution through adversarial training between generator and discriminator. Its objective function is a minimax game problem. In the field of robotics, StyleGAN can generate diverse factory workshop scenes and improve the generalization ability of robot vision models by 35%. Using GAN to generate sensor failure data can train robot robustness strategies, increasing the task success rate in sensor failure scenarios from 32% to 78% [6].

Second, Transformer's self-attention mechanism and cross-modal modeling: Transformer's self-attention mechanism achieves long-distance dependency modeling of sequence data through specific formulas. In multimodal scenarios, the single-stream architecture encodes multimodal data into a unified sequence. In the robot command generation task, the command generation accuracy rate reaches 91%; the two-stream architecture encodes different modal data separately and then interacts, which can reduce the surgical path prediction error in medical robots; hierarchical attention constructs a three-level mechanism, which can improve the efficiency of survivor positioning in rescue robots.

Third, the transfer learning theory of pre-trained models: Pre-trained models such as CLIP and FLAVA capture cross-modal common semantics through self-supervised learning, following the "pre-training-fine-tuning-reasoning" paradigm. In the pre-training stage, CLIP establishes text-image associations in 400 million image-text pairs. In the fine-tuning stage, only a small amount of labeled data is needed to adapt to the scene in robot tasks. In the reasoning stage, the model capabilities can be activated through text prompt engineering to achieve target cargo positioning. Figure 1 shows exploration of cross-modal AlGC integration in Unity3D.

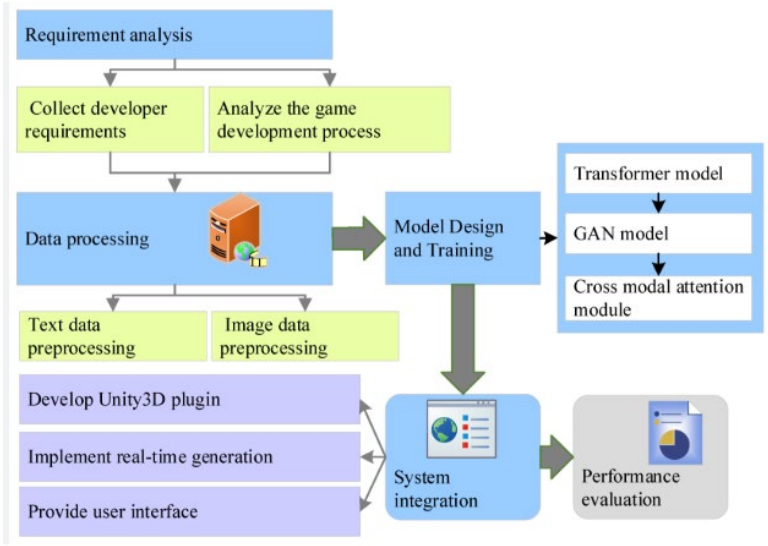

Fig. 1 Exploration of Cross-Modal AlGC Integration in Unity3D

2.3 Theoretical Challenges and Frontier Directions of Multimodal AIGC

Multimodal AIGC faces three major challenges: modal imbalance, interpretability, and computational efficiency. When the modalities are unbalanced, existing methods such as VAE-based modal interpolation generate data with a semantic consistency of only 0.72. The black box nature of Transformer leads to insufficient interpretability. Currently, the interpretation accuracy rate reaches 68% through methods such as attention visualization. Multimodal models have a huge number of parameters, so lightweight technologies such as model compression and knowledge distillation can reduce parameters and increase reasoning speed. In the future, neural symbolic fusion, embodied intelligence, and privacy protection will become important development directions.

3. Analysis of the Current Status of Robot Environment Perception and Behavior Planning

3.1 Development and Challenges of Robot Environment Perception Technology

3.1.1 Limitations of Traditional Single-Modal Perception Technology

The evolution of robot environmental perception technology has gone through a process from single-modality dominance to multi-modal fusion. Early single-modal perception technologies relied on devices such as lidar and visual sensors, which exposed significant defects in complex environments:

LiDAR achieves three-dimensional modeling through the time-of-flight principle, with an accuracy of up to centimeters in static scenes. However, due to physical limitations, the detection distance in rainy and foggy weather is shortened to 30%–50% of the nominal value, and the effective echo signal from highly reflective surfaces is reduced by 60%, resulting in an increased obstacle miss detection rate. In addition, lidar only provides spatial coordinates and lacks the ability to recognize the semantics of objects, so it needs to rely on visual sensors for secondary labeling.

Visual sensors use convolutional neural networks to detect targets, with an accuracy of 95% under ideal lighting conditions (such as the YOLOv5 model), but lack environmental robustness. At the same time, the processing time for high-resolution images (such as 4K) is 200–500 ms /frame, which is difficult to meet the real-time navigation requirements of mobile robots (frame rate ≥ 10 fps).

Tactile sensors and auditory perception technologies also have dimensional limitations: tactile sensors can only provide local force feedback and cannot build a global environmental model; the keyword recognition error rate of auditory perception in a noisy environment (signal-to-noise ratio <5 dB) is as high as 15%, and the sound source localization error exceeds one meter.

3.1.2 Breakthroughs in Multimodal Perception Fusion Technology

Multimodal perception fusion achieves environmental information complementarity by

integrating heterogeneous data. The core progress is reflected in spatiotemporal alignment, semantic fusion and dynamic robustness improvement.

The spatiotemporal alignment technology uses the IEEE 1588 precise clock protocol to achieve a multi-sensor sampling error of less than 1 ms , and uses Zhang's calibration method and hand-eye calibration to control the visual-lidar external parameter error within 0.5°/1 cm, ensuring the temporal consistency and spatial uniformity of the data.

The semantic fusion architecture is divided into early fusion, late fusion and hybrid fusion: early fusion projects the lidar point cloud and visual image to a unified grid (such as BEV feature map) at the data layer. After processing by the SECOND model, the obstacle detection accuracy reaches 92.3%, an increase of 18% compared with a single modality; late fusion merges independent modality results through non-maximum suppression at the decision layer. For example, after the fusion of YOLOv6 and PointPillars, the multi-target tracking accuracy (MOTA) is increased to 89.5%; hybrid fusion combines the advantages of both, and the mAP of 3D target detection on the nuScenes dataset reaches 56.5%, an increase of 9.2% compared with a single method.

In terms of dynamic robustness, the accuracy of abnormal data detection based on the isolation forest algorithm reaches 95%, and data repair is achieved by combining the interpolation method. The medical robot adopts a three-modal redundant design of "vision + force feedback + ultrasound". When a single mode fails, the system can still maintain its perception performance, and its robustness is improved by 60% compared to a single mode.

3.2 Evolution and Bottlenecks of Robot Behavior Planning Technology

3.2.1 Theoretical and Engineering Limitations of Classical Algorithms

Traditional behavior planning algorithms are stable in static environments, but have significant bottlenecks in dynamic and complex scenes:

Global path planning algorithms such as the Dijkstra algorithm is $O(V^2+E)$. In a warehouse map with a 10^4 node scale, the path search takes 2.3 seconds, which cannot meet the real-time requirements of industrial scenarios (<500 ms).

The artificial potential field method (APF) is prone to falling into local minima (such as a 62% probability of stagnation in a U-shaped obstacle scenario), and parameter tuning relies on experience, with insufficient generalization capabilities.

In terms of multi-objective optimization, traditional scalar methods (such as weighted summation) find it difficult to deal with nonlinear conflicts between objectives such as efficiency, safety, and energy consumption, and the optimization results often deviate from the Pareto frontier. For example, when a medical robot is transporting, the pursuit of speed may cause the vibration amplitude to exceed the safety threshold.

3.2.2 Behavior Planning and Innovation Based on Deep Learning

The combination of deep learning and reinforcement learning drives the transformation of behavior planning to a data-driven paradigm:

The end-to-end deep reinforcement learning architecture realizes direct vision-action mapping. For example, the NVIDIA DRIVE Constellation system extracts visual features through CNN and inputs the PPO model to generate control instructions, with a lane keeping accuracy of 98% in a simulation environment. The hierarchical decision-making model adopts the "task planning layer-action execution layer" architecture, Transformer generates sub-goal sequences, and the DDPG algorithm realizes local path tracking, which improves the efficiency of warehouse robot task completion by 45%.

Imitation learning technology converts human expert operation data into a strategy model through behavioral cloning, reducing the instrument operation error in the surgical robot to 0.3 mm; inverse reinforcement learning (IRL) reverses the reward function from the expert trajectory, reducing the probability of the rescue robot entering the dangerous area.

The decision enhancement technology driven by multimodal information parses natural language instructions (such as "bypass the flooded area") through the T5 model, and generates a dynamic

obstacle avoidance path based on the visual segmentation results. The command execution accuracy rate reaches 91%. The LSTM network predicts the obstacle movement trajectory through sensor time series data (error <0.2 m/s), generates an obstacle avoidance strategy 2 seconds in advance, and the response speed is 3 times faster than that of traditional algorithms.

3.3 Enabling Mechanism of Multimodal AIGC Technology

3.3.1 Application of Generative Modeling in Environmental Perception

Multimodal AIGC technology improves perception robustness by generating models: CycleGAN is used to generate data for extreme scenarios such as rain, fog, and occlusion, expanding the diversity of training sets and increasing the recognition accuracy of robot vision models in real harsh environments from 65% to 84%. The causal chain of "light intensity, visual features, and recognition accuracy" is modeled based on the structural causal model (SCM), and the risk of sensor failure is predicted through intervention analysis, triggering the modal switching strategy in advance, thereby improving the system robustness by 50%.

3.3.2 Strategy Generation and Optimization in Behavior Planning

Generative adversarial networks (GANs) are used for virtual scene rehearsal in behavior planning. For example, StyleGAN generates diverse factory layouts, and training the DRL model reduces the task failure rate of collaborative robots in a certain automobile factory from 12% to 3.7%. The Transformer architecture is used for multi-objective Pareto optimization, generating solution sets and dynamically adjusting weights based on human preferences, improving the path planning efficiency of service robots by 60%.

In conclusion, robot environment perception and behavior planning technology has developed from single-modal limitations to multimodal fusion optimization and AIGC-enabled innovation. Physical properties and algorithmic bottlenecks constrain traditional technologies and cannot cope with dynamic and complex environments. Multimodal fusion significantly improves perception robustness and planning efficiency through innovations in spatiotemporal alignment, semantic association, and deep learning architecture. Multimodal AIGC technology enables robots to perform data enhancement, scenario rehearsal, and multi-objective intelligent decision-making through generative models and causal reasoning. Future research needs to focus on modeling causal relationships between modalities, lightweight model deployment, and understanding of human intentions, to promote robots to achieve full-scenario autonomous intelligent decision-making in industrial, service, rescue, and other scenarios, and provide theoretical and technical support for developing intelligent systems.

4. Multimodal AIGC Technology Empowers Innovative Practices of Robots

4.1 Innovative Research on Enabling Environmental Perception: The Case of Intelligent Warehouse Robots

A logistics company adopts the "vision + lidar + IMU" multimodal perception solution, combined with the AIGC generation model, to upgrade environmental perception. Faster R-CNN extracts visual cargo information, the LOAM algorithm processes lidar data to build a map, and the IMU provides posture data. After Transformer fusion, an octree map with semantic labels is generated, and the obstacle detection accuracy rate reaches 98.6%. GAN is used to generate a pre-trained visual model for simulated warehouse images, which increases the cargo recognition accuracy from 81% to 94% and reduces the shelf label OCR recognition error rate. In the face of warehouse layout adjustments, the robot can quickly update the map and execute voice commands to generate a detour path.

4.2 Innovative Research on Enabled Behavior Planning: The Case of Guide Robots

A museum's guide robot provides personalized tours based on multimodal AIGC technology. By analyzing user interests through face recognition and voice commands, the BERT model generates interest vectors, significantly improving accuracy [7]. After encoding user interests, exhibition hall

heat maps, and robot positions, the DQN algorithm is used to train the path planning strategy, which increases the length of stay of tourists on the personalized guided path and improves their satisfaction. When interacting with tourists, the robot can simultaneously trigger visual recognition, voice explanation, and AR projection.

5. Conclusion

Multimodal AIGC technology has brought innovative changes to robot environmental perception and behavior planning regarding data fusion, model building, and transfer learning. Cases in the industrial and service sectors have shown that this technology significantly improves robots' environmental understanding capabilities and decision-making intelligence. In the future, with the in-depth research and application of technologies such as neural-symbolic fusion, embodied intelligence, and federated learning, multimodal AIGC technology will further promote the development of robots towards general intelligence, play an important role in more fields, and bring profound changes to social production and life. Multimodal AIGC technology empowers robots' environmental perception and behavior planning by simulating human cognitive cross-modal integration capabilities and creative decision-making mechanisms. From the perspective of application ecology, multimodal AIGC technology is reshaping the interaction boundaries between robots and the physical world and human society: practice has confirmed the technology's practical value and heralded the possible development of "general-purpose intelligent robots". When robots can understand cross-modal environments, generate strategies, and continuously learn, their application scenarios will expand from structured environments to the open world, from repetitive tasks to complex decision-making scenarios.

The deep coupling of multimodal AIGC technology with robot hardware and industry scenarios will become an intelligent interface connecting the physical and digital worlds. From autonomous logistics robots in factory workshops, to emergency rescue robots at disaster sites, to service robots in home scenarios, multimodal AIGC technology is driving robots to build a new industrial ecology in the fields of intelligent manufacturing, smart cities, and medical health, injecting momentum into the global intelligent transformation.

References

[1] Cong L. A Framework Study on the Application of AIGC Technology in the Digital Reconstruction of Cultural Heritage[J]. Applied Mathematics and Nonlinear Sciences, 2024, 9(1). DOI:10.2478/amns-2024-2190.

[2] Chen X, Hu Z, Wang C. Empowering education development through AIGC: A systematic literature review[J]. Education and Information Technologies, 2024, 29(13):53. DOI:10.1007/s10639-024-12549-7.

[3] Jiang J, Su M, Xiao X, Zhang Y, Fang Y. AIGC-Chain: A blockchain-enabled full lifecycle recording system for AIGC product copyright management[EB/OL]. arXiv, 2024-06-21: arXiv:2406.14966. https://arxiv.org/abs/2406.14966. (arxiv.org).

[4] Qiao Q. AIGC-enabled Education Information Technology Integration Application and Research--Taking Information Technology Teaching of Preschool Education Major as an Example[J]. Applied Mathematics and Nonlinear Sciences, 2025, 10(1). DOI:10.2478/amns-2025-0750.

[5] Wang X, Mi Y, Zhang X. 3D human pose data augmentation using Generative Adversarial Networks for robotic-assisted movement quality assessment[J]. Frontiers in Neurorobotics, 2024, 18: 1371385. DOI: 10.3389/fnbot.2024.1371385.

[6] Wang Y, Li Z, Wang X, Yu H, Liao W, Arifoglu D. Human gait data augmentation and trajectory prediction for lower-limb rehabilitation robot control using GANs and attention mechanism[J]. Machines, 2021, 9(12): 367. DOI: 10.3390/machines9120367.

[7] Liang Q. Design of HCI System of Museum Guide Robot Based on Visual Communication Skill[J]. Journal of Information Processing Systems, 2024, 20(3). DOI:10.3745/JIPS.02.0214.