

Construction of Humanoid Robot and Generative Artificial Intelligence Information System Based on Deep Learning

Huijuan Sun^{1,a*} and Jian Shen^{2,b}

¹*School of Computer and Artificial Intelligence, Henan Finance University, Zhengzhou 451464, Henan, China*

²*Henan Education Examinations Center for Test Development and Assessment, Zhengzhou 451464, Henan, China*

With the rapid development of DL technology, artificial intelligence (AI) and robotics are gradually merging, especially the combination of humanoid robots and generative AI, which is becoming a key technology to promote the intelligent transformation of future industries. However, in traditional research, humanoid robots and generative AI technologies are developed in isolation, lacking effective collaboration and natural interaction, and lacking innovation in cross-domain application. To address these problems, in this study, an information system is constructed based on DL humanoid robots and generative AI. The system consists of three modules: a perception, decision and action module, and generation and interaction modules. The perception module collects and preprocesses data of multiple modalities and then fuses the multimodal data. The decision and action module achieves the decision-making function of the information system through the DQN model. The generation and interaction module introduces the Generative Pre-Trained Transformer (GPT) model to realize the generation and interaction functions of the information system. Experiments have shown that the image perception ability of the system constructed in this study is good. The average perception error of 100 areas in the experiment is 0.0361. The decision-making ability is strong, and the decision-making score for different scenarios exceeds 0.83. It can also maintain good consistency in the generated content while taking emotions into account during conversations. The robot equipped with the system proposed in this paper and the traditional robot were put into the service industry for comparative experiments. The results demonstrated that the robot equipped with the proposed system is superior to the traditional robot according to all scoring indicators. The information system of humanoid robots and generative AI based on DL constructed in this study breaks through the limitations of the isolated development of traditional humanoid robots and generative AI, through the strong integration of deep perception, decision-making and interaction. It significantly improves the intelligence level of robots, has broad application prospects, and provides new ideas and technical support for promoting industrial intelligent transformation and cross-domain application innovation.

Keywords: Deep learning, Humanoid robot, Generative Artificial Intelligence, Multimodal Perception, Deep Q network, Generative Pre-Trained transformer

1. INTRODUCTION

As AI [1–2] and robotics [3–4] continue to develop, deep learning (DL) [5–7] technology, as its core driving force, has

been widely used in various intelligent systems. In particular, the combination of humanoid robots and generative AI has become an important driving force for the intelligent transformation of industries. However, in traditional research, these two technologies tend to develop independently, without effective coordination and interaction. Humanoid robots focus more on the operation and perception of the physical world,

*Corresponding Author. Email: ^ahafushj@126.com, ^bshenjian@haecan.cn

while generative AI has made significant progress in natural language processing and content generation. Although each has its own unique advantages, their combination faces many technical challenges, including cross-domain integration, emotional interaction, and adaptability and security in actual application scenarios. Therefore, finding effective ways to integrate these two technologies and foster their collaboration in real-world applications has become a critical issue that requires immediate attention.

In recent years, there has been extensive research in the field of intelligent robots and AI within the academic community. For example, Lin et al. studied the willingness of customers to use AI robot equipment in hotel services, borrowed and verified the extended theory of acceptance of AI equipment use in hotel service environments, and the results showed that the theory is applicable [8]. Zemmar et al. integrated robots into a shielding layer to physically isolate medical staff and patients. The results showed that the use of robots and AI can reduce infectious contamination and assist in patient management in surgical environments when a large number of patients enter [9]. Grischke et al. reviewed the existing applications of dental robots and AI. The results indicated that machine learning and AI have been utilized in dental research to process vast amounts of data, aiding in dental decision-making through data-driven analysis algorithms. However, the implementation of robotic applications in this field remains limited [10]. Van Wynsberghe proposed a definition of sustainable AI, that is, sustainable AI is a movement that aims to promote changes throughout the life cycle of AI products to practice greater ecological integrity and social justice [11]. Huang and Rust created a strategic framework for leveraging AI to attract customers by offering various service benefits. This framework provides guidelines on how to utilize different AI technologies to engage customers, depending on the nature of the service task [12]. However, although these studies have made progress in their respective fields, they still have certain limitations. For example, when robots execute generative AI instructions, they often lack the ability to understand context and to provide real-time feedback; and when generative AI generates content, it usually fails to fully consider the robot's execution capabilities and its limitations in the physical environment. Therefore, the research at that time was mostly limited to the optimization of a single technology and lacked systematic integration across fields.

In order to overcome the above limitations, some scholars have begun to try to adopt different technology integration methods. For example, Chen Pei-Yu and Yuan-Chen Liu investigated the use of neural networks and AI in image recognition. In their study, students acted as instructors, training AI robots to identify objects. The findings indicated that teaching AI robots image recognition technology was more motivating for students to engage with science compared to traditional textbook-based learning [13]. Xue et al. explored the feasibility and challenges of combining multimodal fusion methods with physical human-computer interaction based on its characteristics, and proposed some research directions based on multimodal fusion methods [14]. Han et al. proposed a hierarchical framework to construct knowledge at the primitive, environmental, and

strategic levels, and used this knowledge as a training set for legged robots, thereby training a robot that can imitate animal or human behavior [15]. Xiao et al. developed a collaborative bilateral upper limb rehabilitation robot system based on mirror therapy and virtual stimulation. The system uses an adaptive proportional-integral-derivative controller based on a radial basis function neural network to improve the tracking performance of the patient side of the robot arm [16]. Although these fusion methods have advanced the technology to a certain extent, they still have certain shortcomings, especially in the context of deep integration of generative AI and humanoid robot technology. Problems such as situational adaptability and multi-tasking have not been fully addressed.

In this study, a humanoid robot and generative AI information system is constructed based on DL, introducing the perception module, decision and behavior module, and generation and interaction module. In the perception module, data of multiple modalities are collected, different modal data are processed differently, and the data of multiple modalities are fused and processed [17]. In the decision and behavior module, the DQN model [18–20] is used to help the system make decisions based on the collected data. The inclusion of the GPT model [21–23] in the generation and interaction module enables the system to generate natural, fluent and emotional responses while understanding user input. The experiments show that the system proposed in this paper has good perception, decision-making and interaction capabilities. The system is installed on robots and compared with those of traditional robots in the service industry. The results show that the robot equipped with the proposed system superior to traditional robots according to several evaluation indicators, showing good application prospects. The significance of this study is that through the integration and innovation of DL technology, it addresses the problem of the isolated development of traditional humanoid robots and generative AI technology, and can be integrated into several domains. This provides a practical technical solution for achieving the intelligent transformation of the industry and promotes the development of robot technology so as to achieve higher intelligence and personalization.

2. SYSTEM ARCHITECTURE

The system architecture in this study is divided into three modules: perception module, decision and behavior module, and generation and interaction module.

2.1 Perception Module

2.1.1 Data Collection

In the data collection stage, the perception module first needs to collect raw data through various sensors. The data generated by these sensors usually have different formats. The sensors and data formats are shown in Table 1.

As shown in Table 1, the RGB camera collects visual data of the environment and outputs a three-dimensional matrix $X \in R^{H \times W \times C}$. H and W are the height and width of

Table 1 Data collection table.

Sensor Type	Collection data format	Data dimensions/features	Function and use
RGB Camera	$X \in R^{H \times W \times C}$	Image (width W, HeightH, Number of color channels C)	Provides visual information of the environment and capture images or video streams for object detection and scene understanding.
Depth Camera	$D \in R^{H \times W}$	Depth map (depth value for each pixel)	Obtains depth information of the environment for spatial perception, obstacle detection, and path planning.
Light Detection and Ranging	$P = \{p_1, p_2, \dots p_n\}$	Point cloud data ($p_i = (x_i, y_i, z_i) \in R^3$)	Acquires three-dimensional structure data in space for obstacle detection, environment modeling and positioning.
Microphone	$s(t) \in R$	Speech or audio signals (time series data)	Captures environmental sounds for speech recognition, sound source localization, and speech feedback generation.
Inertial Measurement Unit	$a(t) = (a_x(t), a_y(t), a_z(t))$ $w(t) = (w_x(t), w_y(t), w_z(t))$	Acceleration $a(t) \in R^3$ Angular velocity $w(t) \in R^3$	Captures the robot's motion state for posture estimation, navigation, and motion planning.

the image, respectively, and C is the number of color channels, which is 3, that is, three colors R, G, and B. The depth camera generates depth information for each pixel through structured light technology and outputs a depth map $D \in R^{H \times W}$. The laser radar device generates point cloud data $P = \{p_1, p_2, \dots p_n\}$ in three-dimensional space, and each point contains three-dimensional coordinates (x_i, y_i, z_i) . The microphone collects sound data, which is audio information that changes over time. The inertial measurement unit measures the robot's acceleration and angular velocity.

2.1.2 Data Preprocessing

The task of data preprocessing is to clean, normalize, and denoise the raw data so that they are suitable for subsequent analysis and feature extraction.

The data collected by the sensor is usually affected by noise, which may come from the environment, the sensor itself, or other external factors [24]. The goal of denoising is to retain the effective signal as much as possible and remove useless noise components. In this study, the Gaussian filter method is used to denoise the data[25], for which the formula is:

$$X_{denoised} = X_{raw} * G_{\sigma} \quad (1)$$

where X_{raw} is the original signal, $X_{denoised}$ is the denoised signal, G_{σ} is the standard deviation, σ is the Gaussian kernel, and $*$ represents the convolution operation.

Different types of sensor data have different ranges and scales. Standardization processing converts them into a unified scale, making subsequent model training and feature extraction more efficient. This study uses the zero mean method to standardize the data. The standardization formula is:

$$X_{norm} = \frac{X - \mu_X}{\sigma_X} \quad (2)$$

where μ_X and σ_X are the mean and standard deviation of the original input data respectively, and the standardized data X_{norm} has the property of mean 0 and standard deviation 1.

In a multi-sensor system, some sensors may have missing data due to failure, occlusion or other reasons [26]. In this study, the method used to process missing values is linear interpolation [27-29] for which the formula is:

$$x_k = x_i + \frac{(x_{i+1} - x_i)}{(t_{i+1} - t_i)}(t_k - t_i) \quad (3)$$

where t_i and t_{i+1} are the time points corresponding to x_i and x_{i+1} respectively, t_k is the time point corresponding to the missing value, and x_k is the interpolated data. However, if there are too many missing data or linear interpolation is not applicable, the missing time is filled with the mean:

$$X_{filled} = \mu_X \quad (4)$$

In a multi-sensor system, data from different sensors usually have different timestamps, and the goal of time synchronization is to align them to the same time axis. Assume that sensors S_1 and S_2 provide time series $X_1(t_1)$ and $X_2(t_2)$, respectively. In order to achieve time synchronization, S_2 interpolates the data to make it consistent with the S_1 timestamp. Usually, the time that needs to be synchronized is very close, so this study used linear interpolation for this synchronization. The formula is:

$$X_{2sync}(t_1) = X_2(t_2) + \frac{(t_1 - t_2)}{(t_1 - t_2 + \varepsilon)}(X_2(t_{i+1}) - X_2(t_i)) \quad (5)$$

where ε is a very small positive number, which is used to avoid zero division errors without affecting the results.

2.1.3 Feature Extraction

The specific method used for feature extraction depends on the type of sensor and the requirements of the task. This paper uses different feature extraction methods for different data.

For image data, this study selected Convolutional Neural Networks (CNN) [30–32] to extract features from image data,

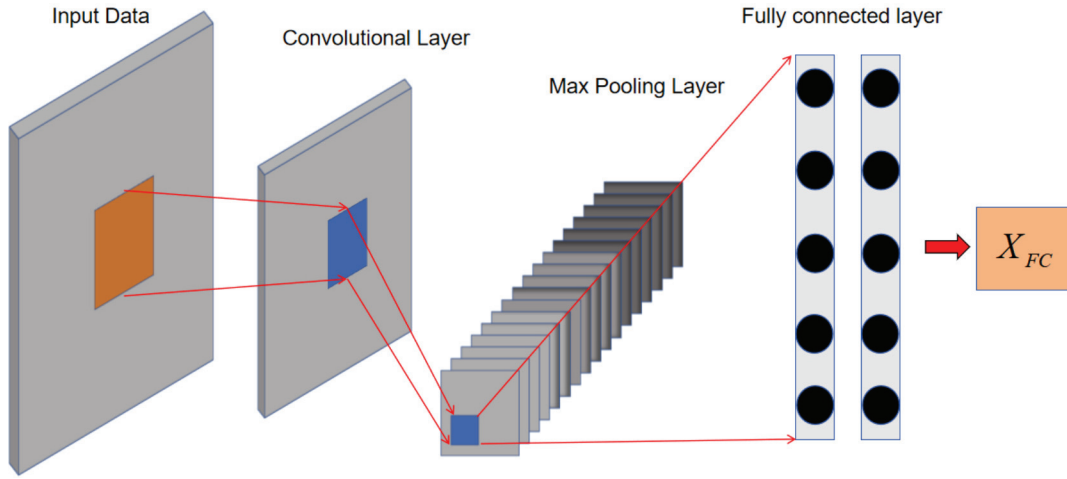


Figure 1 CNN structure diagram.

and extracted local and global features of the image by means of convolution and pooling operations. Each layer of CNN can extract features at different levels, usually from low-level edge and texture features to high-level object and scene features.

The image data in this study is $X \in R^{H \times W \times C}$ captured by the camera. First, through the convolution operation, the image is extracted by the convolution kernel:

$$X_{conv} = Conv(X, W) \quad (6)$$

where $Conv$ represents the convolution operation, and X_{conv} is the feature map after convolution.

The nonlinear function ReLU is applied to the feature map to enhance the network's expressive power:

$$X_{ReLU} = ReLU(X_{conv}) \quad (7)$$

Afterwards, the maximum pooling operation is used to further extract important image information:

$$X_{MaxPool} = MaxPool(X_{ReLU}) \quad (8)$$

The pooled feature map is flattened and processed through a fully connected layer to obtain the final high-level feature vector:

$$X_{FC} = FC(X_{MaxPool}) \quad (9)$$

The CNN structure diagram designed in this paper is shown in Figure 1.

For the three-dimensional spatial data $P = \{p_1, p_2, \dots, p_n\}$ collected by the laser radar device, this study uses PointNet[33-34] to extract features from the data. Each feature of the collected data is represented as $f(p_i)$. The global feature representation of the entire point cloud collected is obtained through PointNet:

$$f_{global} = MaxPool(f(p_1), f(p_2), \dots, f(p_n)) \quad (10)$$

The features of audio time series data are extracted so as to convert the audio signal into frequency domain representation and extract the frequency components and time-frequency features of the sound. The method used in this paper is Mel-Frequency Cepstral Coefficients (MFCC) [35]. The specific steps are:

- (1) Perform short-time Fourier transform (STFT) [36] on the audio. For the collected audio $s(t) \in R$, its STFT representation is:

$$S(f, t) = F\{s(t) \cdot \omega(t - t_0)\} \quad (11)$$

where F is the Fourier transform, $\omega(t - t_0)$ is the window function, $S(f, t)$ is the time-frequency representation, f is the frequency, and t is the time.

- (2) Map the spectrogram to the Mel scale.
- (3) Use discrete cosine transforms[37] to extract MFCC features. The extracted MFCC is expressed as:

$$MFCC(s(t)) = DCT(\log(MelSpectrogram(S(f, t)))) \quad (12)$$

where Mel-Spectrogram is the spectrum on the Mel frequency scale, and DCT is the discrete cosine transform.

2.1.4 Multimodal Fusion

After the data has undergone preprocessing and feature extraction, the features extracted from the data of different modalities are fused to obtain a comprehensive analysis of the data. This paper introduces the self-attention mechanism to the data in different modalities for feature-level fusion. This fusion method can effectively remove data redundancy and improve computational efficiency.

The self-attention mechanism can assign different weights to each modality based on the correlation between them. By calculating the correlation between each modality, the model can automatically learn which modalities are more important to the task, thereby assigning them greater weights. For the feature data extracted in the previous paper (image features named f_{image} , point cloud global features named f_{global} , and audio features named f_{audio}), the self-attention mechanism can generate a weight α to adjust the fusion of several features:

$$\alpha = Attention(f_{image}, f_{global}, f_{audio}) \quad (13)$$

Then, according to the weight α , feature fusion is performed by weighted summation:

$$f_{fused} = \alpha_1 f_{image} + \alpha_2 f_{global} + \alpha_3 f_{audio} \quad (14)$$

The data after feature fusion can provide better data input for system decision-making.

2.2 Decision and Behavior Module

In this module, the deep Q network (DQN) is used to help the system make decisions and plans based on the fused feature data obtained by the perception module. The core idea of DQN is to learn an approximate Q-value function through a deep neural network, and then make decisions through the Q-value function. The Q-value function is expressed as $Q(s, a)$, which represents the expected return that can be obtained after taking action a in state s . The DQN approximates the Q-value function through a deep neural network so that it can handle high-dimensional state space.

The basic operation process of DQN is as follows:

- (1) Initialization: The initial state can be set to s_0 , and the parameters θ_0 of the DQN can be initialized. θ_0 is the initial weight of the deep neural network. The parameters θ'_0 of the target network can be initialized. The target network is usually the same as the Q network, but it is updated regularly to maintain the stability of the update. The experience replay pool is initialized, and is used to store the historical experience of the system.

- (2) Interaction with the environment: In each operation, the system needs to interact with the environment to collect data and make decisions.

Assuming the current state is s_t , in state s_t , the current Q network $Q(s_t, a)$ is used to calculate the Q value of each action a , and the action with the largest Q value is selected. Assuming the selected action is a_t , after executing the a_t action, the environment returns the new action s_{t+1} and the corresponding reward r_t . Finally, the current experience (s_t, a_t, r_t, s_{t+1}) is stored in the experience replay pool.

- (3) Update the Q network: The goal of DQN is to make the Q value function accurately predict the long-term return of each state-action pair (s-a). Therefore, the Q network needs to be updated. To update, randomly extract a small batch of data $\{(s_i, a_i, r_i, s_{i+1})\}$ from the experience replay pool each time, calculate the maximum Q value $\max_{a'} Q_{\theta'}(s_{i+1}, a')$ under the next state s_{i+1} through the target network, and then calculate the target Q value y_i :

$$y_i = r_i + \gamma \max_{a'} Q_{\theta'}(s_{i+1}, a') \quad (15)$$

where γ is the discount factor that determines the impact of future rewards on the current Q value.

The current Q network can be used to estimate the Q value $Q_{\theta}(s_i, a_i)$, and then the parameters θ of the Q network are updated by minimizing the error between the target Q value and the current Q value:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N (Q_{\theta}(s_i, a_i) - (r_i + \gamma \max_{a'} Q_{\theta'}(s_{i+1}, a')))^2 \quad (16)$$

where N is the batch size, the gradient of $L(\theta)$ with respect to parameter θ is calculated using the back-propagation algorithm, and the parameter θ is updated using the gradient descent method:

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} L(\theta) \quad (17)$$

where α is the learning rate, which controls the step size of the update.

- (4) Update the target network: In order to ensure the stability of the system operation, the parameters θ of the Q network are copied to the target network θ' every certain number of steps. In this way, the high-frequency fluctuation problem of the Q value during operation is alleviated:

$$\theta' \leftarrow \theta \quad (18)$$

2.3 Generation and Interaction Module

In this study, the GPT model is selected to build the generation and interaction module, with the goal of enabling the system to understand the user's input and generate appropriate natural language responses. The specific construction steps are as follows.

2.3.1 Input Preprocessing Context Understanding

The user's input is usually text or voice input. Text is processed directly, and voice is converted to text for processing. Assuming that the current user's input is x_t , in order to ensure that the generation and interaction modules can understand the context and history of the conversation, $H = \{h_1, h_2, \dots, h_n\}$ needs to be introduced. Each h_i is generated by the model or input by the user in the i -th round of dialogue. Therefore, in this module, the dialogue history and current input are concatenated and passed to the GPT model to help the model capture the context of the dialogue. The input format is:

$$Input_t = [H, x_t] \quad (19)$$

2.3.2 Generation Process of GPT Model

The GPT model is based on the Transformer architecture. The core part is the self-attention mechanism that captures the dependencies between words in the input sequence. The GPT model is an autoregressive model that generates the next word step by step and uses the word as the condition for the next moment.

After processing the input $Input_t$, GPT generates the output through the following steps:

- (1) $Input_t$ is first converted into a vector representation in the embedding layer:

$$E_t = \text{Embedding}(Input_t) \quad (20)$$

where E_t is the word vector representation of the input text, and these vectors are passed from the embedding layer to the self-attention layer of the Transformer.

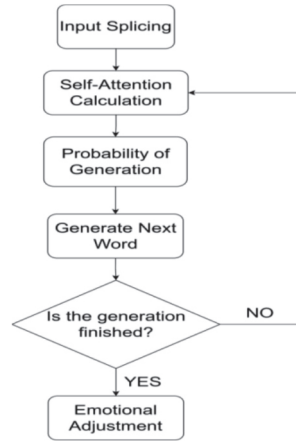


Figure 2 GPT generation flow chart.

- (2) In each layer of the Transformer, the input vector passes through the self-attention mechanism, which calculates the degree of dependence of each word on other words based on the relationship between the current word and all words. The calculation formula for the self-attention mechanism is:

$$Attention(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (21)$$

where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key, which is used to scale the stability of the calculation. In Transformer, the input vector is calculated in parallel through multiple self-attention heads, allowing the model to focus on different parts of the input, thereby obtaining a more comprehensive output.

- (3) The output after the self-attention mechanism is passed to the feedforward upgrade network for nonlinear transformation. The formula for this process is:

$$H_t = FFN(Attention(Q, K, V)) \quad (22)$$

where FFN represents the feedforward neural network, and H_t represents the current hidden state.

- (4) Finally, the GPT model transforms the hidden state H_t into a probability distribution through a linear transformation. The distribution represents the probability of all possible words generated at the current position. The conditional probability of generating each word is:

$$P(y_t | x_1, x_2, \dots, x_t) = \text{softmax}(W_h H_t + b) \quad (23)$$

where W_h is the weight matrix of the hidden state, b is the bias term, and the softmax function converts it into a probability distribution for generating the next word. Based on this probability distribution, the model can select the most likely word as the output with:

$$y_t = \arg \max P(y_t | x_1, x_2, \dots, x_t) \quad (24)$$

where y_t is the word with the highest probability and is also the next output.

2.3.3 Emotion and Tone Adjustment

In actual interaction, it is necessary to understand the context of the language, and to adjust the generated response based on sentiment analysis and tone. For example, when it is detected that the user is emotionally excited, the model may need to generate soothing sentences; if the user communicates in a humorous way, the robot can generate a humorous response.

During the generation process, sentiment adjustment is achieved by changing the generation probability distribution:

$$P(y_t | H_t, x_t, Sentiment_t) = \text{softmax}(W_h H_t + b + \lambda \cdot Sentiment_t) \quad (25)$$

where λ is the weight of sentiment adjustment, and $Sentiment_t$ is the sentiment information. In this way, the system adjusts the output content according to the sentiment state.

Ultimately, the generated response is not just a single word, but a coherent sentence or dialogue. After each round of generation, GPT can iteratively generate subsequent content based on the generated vocabulary sequence $(y_1, y_2, \dots, y_n)$ until the end symbol is generated. The GPT generation process is shown in Figure 2.

3. EVALUATION AND ANALYSIS

3.1 Perception Error Analysis

In order to analyze the image perception results, the system's image perception ability is quantified and the image perception accuracy is evaluated. By means of experiments, the image perception results and the real images are compared and analyzed, the absolute error between the two is calculated, and the error heat map is drawn.

Figure 3 shows the error heat map of the image perception result and the actual image. There are 10×10 areas in total, with each area representing the overlap of the image perception result and the real image. The value represents the absolute error between the image perception result and the actual image in the area. The area with larger error can appear

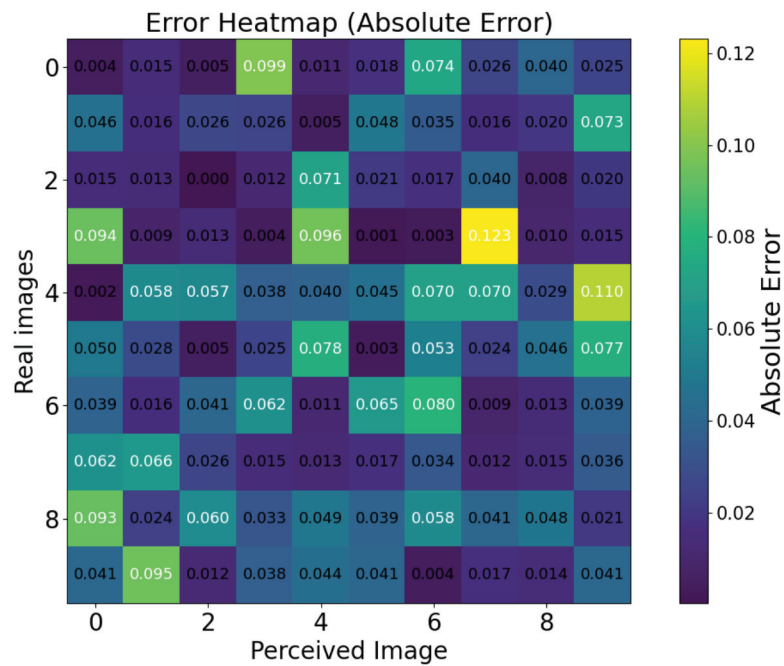


Figure 3 Error heat map.

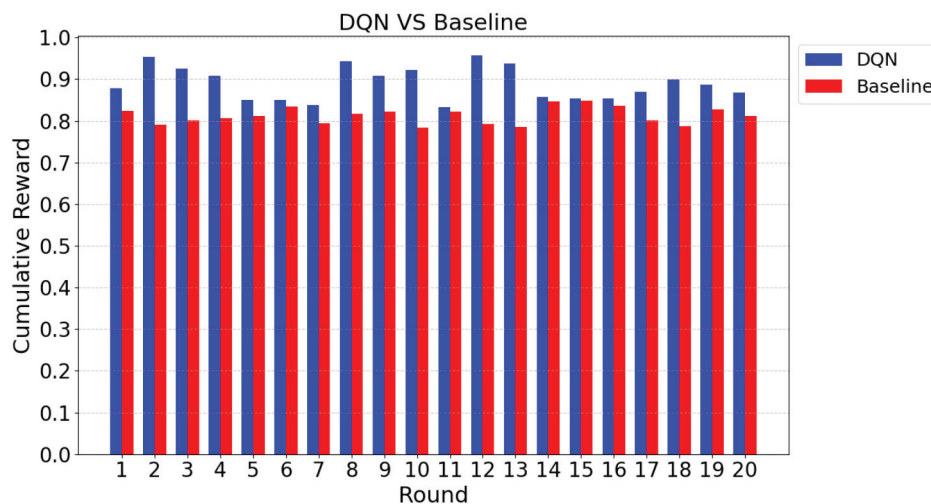


Figure 4 Decision-based score comparison chart.

bright, and the area with smaller error can appear dark. The absolute errors in most areas are below 0.05, indicating that the system's perception in most areas is relatively accurate. However, there are some areas with large errors (higher than 0.1), indicating that the system's perception in these areas has large deviations, and further analysis of the causes and improvements are needed. The average analysis of the error values in the heat map shows that the average absolute error of 100 regions is 0.0361, which shows that the overall image perception ability of the system is relatively good.

3.2 Decision Quality Analysis

In order to evaluate the decision-making ability of the system, twenty different scenarios were selected for the system's decision-making module. The cumulative reward of each

scenario was normalized into the interval [0,1] as a score. At the same time, a benchmark strategy was introduced for comparison with the DQN model of the decision module of this system.

Figure 4 is a decision-based score comparison chart showing the different decision scores for two decision models in twenty different scenarios. Figure 4 shows that the decision score of the DQN model used in this paper is roughly maintained in the range of 0.83–0.96, while the decision score of the baseline strategy is in the range of 0.78–0.85. Compared with the baseline strategy, DQN has a decision score of more than 0.85 in 18 out of 20 scenarios, and 8 out of these 18 scenarios have a decision score of more than 0.9. This shows that the DQN model can learn effective strategies in different scenarios to maximize the cumulative reward, indicating that the decision module proposed in this paper can make better decisions in different scenarios.

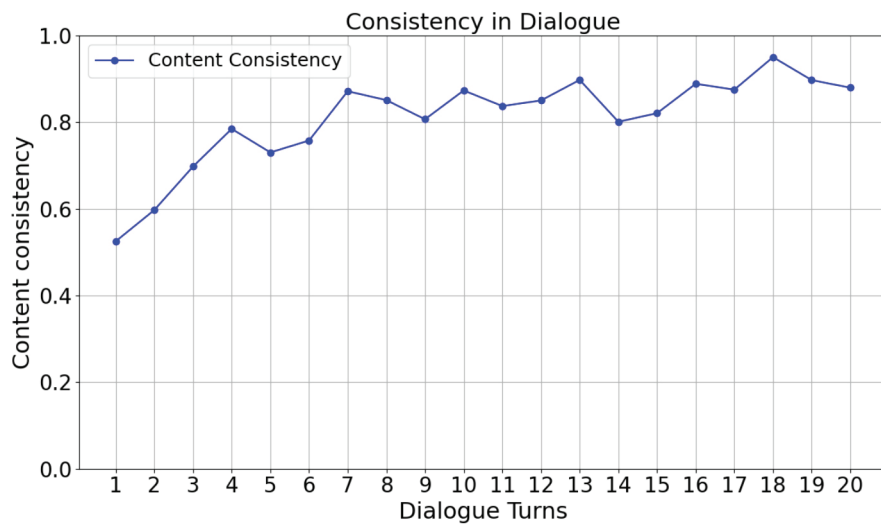


Figure 5 Content consistency analysis chart.

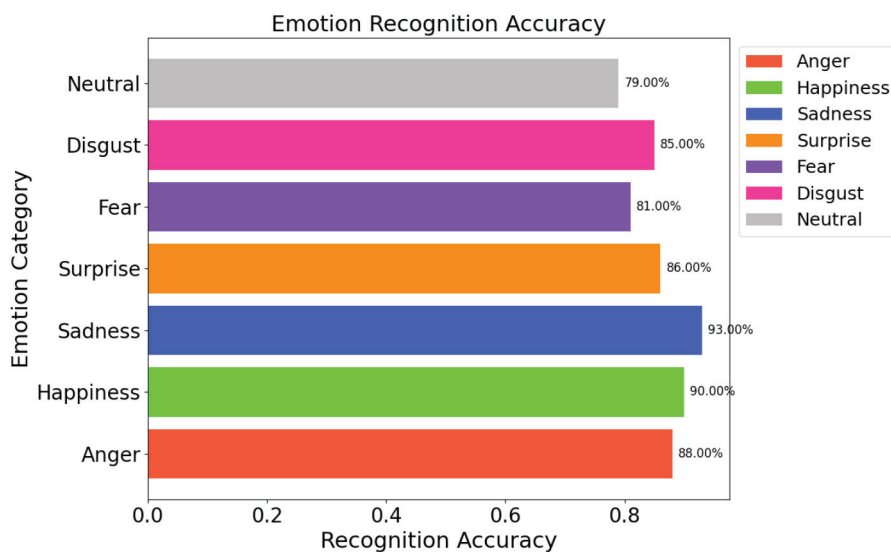


Figure 6 Accuracy of emotion recognition.

3.3 Generation and Interaction Analysis

In order to evaluate the functions of the system's generation and interaction modules, experiments were conducted to randomly conduct conversations and use cosine similarity to calculate the similarity between semantics. In this way, the similarity between the generated text and the context was compared to determine the content consistency between the texts, thereby evaluating the system's generation and interaction functions.

Figure 5 is an analysis chart for content consistency from the beginning to the end of the conversation. Early in the conversation, the content consistency between the text generated by the system and the context is low, which may be mainly because there is not much text to learn at the beginning of the conversation. However, after the first few rounds of conversation, an upward trend in content consistency can be observed. In the middle of the conversation, the system's response becomes more consistent and although the consistency fluctuates, it is still increasing overall. At

the end of the conversation, the content consistency is basically determined, indicating that at this time the system's interaction model for the conversation has matured and can efficiently and stably generate content to complete the conversation.

In addition to analyzing the consistency of the generated content, the study also analyzed the recognition accuracy of the emotional state of the generated content and the real emotion to determine whether the system's recognition of the emotion in the sentence is consistent with the actual emotion.

Figure 6 shows the accuracy analysis of seven different types of emotion recognition. The system has the highest accuracy for the recognition of emotions such as sadness, with a recognition accuracy of 93%. Next are the emotions of happiness and anger, with recognition accuracies of 90% and 88%, respectively. Surprise and disgust emotions have recognition accuracies of 86% and 85%, respectively. Finally, fear and neutral emotions have recognition accuracies of 81% and 79%, respectively. Through numerical analysis, it was found that the system's recognition accuracy for most emotions is

Comparison Between Deep Learning-based Robot and Traditional Robot

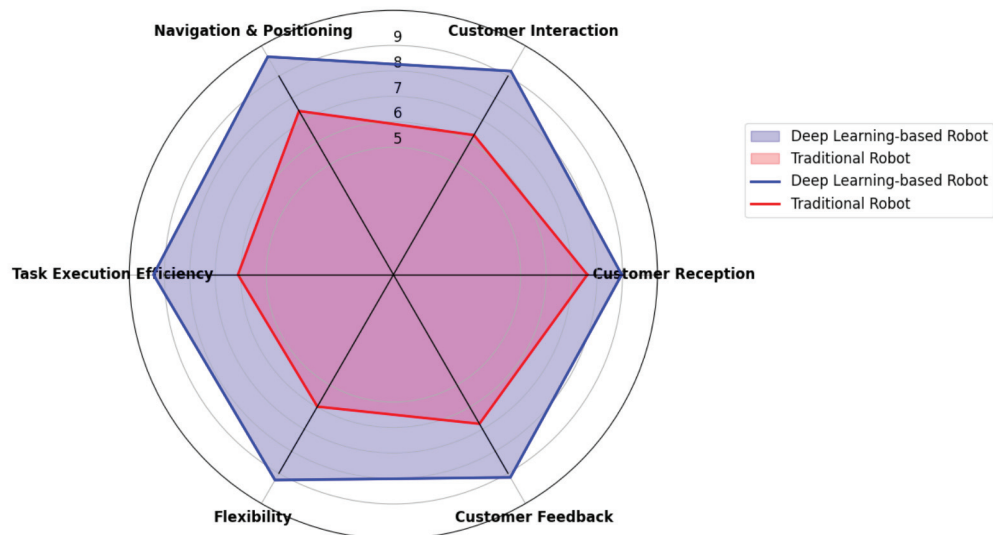


Figure 7 Comparison of robot scoring indicators.

over 85%, except for fear and neutral emotions. This may be partly because fear and neutral emotions are not easy to recognize, but it also shows that the system designed in this study has certain differences in terms of recognizing different emotions. Although the recognition is relatively accurate, there is still room for improvement.

3.4 Robot Comparison

In order to further analyze the effect of the system designed in this paper on robots, an experiment was conducted to score the performance of robots equipped with the system proposed in this paper (robots based on DL) and that of traditional robots in the service industry. There are six scoring indicators: customer reception, customer interaction, navigation and positioning, task execution efficiency, flexibility, and customer feedback.

Figure 7 shows a comparison of the six scoring indicators of DL-based robots and traditional robots in the service industry. The blue area is the score of the DL-based robots proposed in this study, and the red area is the score of traditional robots. DL-based robots can provide customers with a more human-like and natural interactive experience, while maintaining high efficiency in a multi-tasking environment, and responding to customers quickly. This shows that robots based on DL have advantages over traditional robots and may gradually replace traditional robots in the service industry in the future.

4. CONCLUSIONS

This study examined and constructed a humanoid robot and generative AI information system based on DL. On the basis of multimodal data collection and fusion, the DQN model and GPT model are introduced as the decision model and generative interaction model of the system respectively. The system achieves multimodal data collection and fusion,

and has good perception, decision-making and interactive generation capabilities. Compared with traditional robots, the system installed on the robot has obvious advantages. This study has contributed to the integration of robots and AI in the future, although it has several shortcomings. After experimental analysis, the perception module designed in this study is not ideal in terms of perceiving the actual situation. There are perceptions with large errors, and the recognition accuracy of some emotions in the generative interaction module is low. Therefore, the aim of future research is to further improve the perception module and the generative interaction model, optimize the perception error to the minimum, and ensure that the optimization effect is balanced so that all emotions can be recognized accurately.

FUNDING

This work was supported by the Science and Technology Research Project of Henan Province (No. 262102321076, 252102211110).

REFERENCES

1. Mintz, Yoav, and Ronit Brodie. "Introduction to AI in medicine". *Minimally Invasive Therapy & Allied Technologies* 28.2 (2019): 73–81.
2. Kaul, Vivek, Sarah Enslin, and Seth A. Gross. "History of artificial intelligence in medicine". *Gastrointestinal Endoscopy* 92.4 (2020): 807–812.
3. Abbott, Jake J., Eric Diller, and Andrew J. Petruska. "Magnetic methods in robotics". *Annual Review of Control, Robotics, and Autonomous Systems* 3.1 (2020): 57–90.
4. Delmerico, Jeffrey, Mintchev, Stefano, Giusti, Alessandro, Gromov, Boris, Melo, Kamilo, Horvat, Tomislav, et al. "The current state and future outlook of rescue robotics". *Journal of Field Robotics* 36.7 (2019): 1171–1191.

5. Mater, Adam C., and Michelle L. Coote. "Deep learning in chemistry". *Journal of chemical information and modeling* 59.6 (2019): 2545–2559.
6. Christin, Sylvain, ÉricHervet, and Nicolas Lecomte. "Applications for deep learning in ecology". *Methods in Ecology and Evolution* 10.10 (2019): 1632–1644.
7. Janiesch, Christian, Patrick Zschech, and Kai Heinrich. "Machine learning and deep learning". *Electronic Markets* 31.3 (2021): 685–695.
8. Lin, Hongxia, Oscar Hengxuan Chi, and DoganGursoy. "Antecedents of customers' acceptance of artificially intelligent robotic device use in hospitality services". *Journal of Hospitality Marketing & Management* 29.5 (2020): 530–549.
9. Zemmar, Ajmal, Andres M. Lozano, and Bradley J. Nelson. "The rise of robots in surgical environments during COVID-19". *Nature Machine Intelligence* 2.10 (2020): 566–572.
10. Grischke, Jasmin, Jahannsmeier, Lars, Eich, Lukas, Griga, Leif, Haddadin, Sami. "Dentronics: Towards robotics and artificial intelligence in dentistry". *Dental Materials* 36.6 (2020): 765–778.
11. Van Wynsberghe, Aimee. "Sustainable AI: AI for sustainability and the sustainability of AI". *AI and Ethics* 1.3 (2021): 213–218.
12. Huang, Ming-Hui, and Roland T. Rust. "Engaged to a robot? The role of AI in service". *Journal of Service Research* 24.1 (2021): 30–41.
13. Chen, Pei-Yu, and Yuan-Chen Liu. "Impact of AI Robot Image Recognition Technology on Improving Students' Conceptual Understanding of Cell Division and Science Learning Motivation". *Journal of Baltic Science Education* 23.2 (2024): 208–220.
14. Xue, Teng, Wang, Weiming, Ma, Jin, Liu, Wenhai, Pan, Zhenyu, Han, Mingshuo. "Progress and prospects of multi-modal fusion methods in physical human–robot interaction: A review". *IEEE Sensors Journal* 20.18 (2020): 10355–10370.
15. Han, Lei, Zhu, Qingxu, Sheng, Jiapeng, Zhang, Chong, Li, Tingguang, Zhang, Yizheng, et al. "Lifelike agility and play in quadrupedal robots using reinforcement learning and generative pre-trained models". *Nature Machine Intelligence* 6.7 (2024): 787–798.
16. Xiao, Wei, Chen, Kai, Fan, Jiaming, Hou, Yifan, Kong, Weifei, Dan, Guo. "AI-driven rehabilitation and assistive robotic system with intelligent PID controller based on RBF neural networks". *Neural Computing and Applications* 35.22 (2023): 16021–16035.
17. Debie, Essam, Rojas, Raul Fernandez, Fidock, Justin, Barlow, Michael, Kasmarik, Kathryn, Anavatti, Sreenatha. "Multi-modal fusion for objective assessment of cognitive workload: A review". *IEEE transactions on cybernetics* 51.3 (2019): 1542–1555.
18. Li, Jianxin, Chen, Yiting, Zhao, XiuNiao, Huang, Jinyu. "An improved DQN path planning algorithm". *The Journal of Supercomputing* 78.1 (2022): 616–639.
19. Yang, Yang, Li Juntao, and Peng Lingling. "Multi-robot path planning based on a deep reinforcement learning DQN algorithm". *CAAI Transactions on Intelligence Technology* 5.3 (2020): 177–183.
20. Wang, Zhirun, Zhao, Wenjing, Zhai, Aiping, He, Peng, Wang, Dong. "DQN based single-pixel imaging". *Optics Express* 29.10 (2021): 15463–15477.
21. Chavez, Martin R., Butler, Thomas S., Rekawek, Patricia, Heo, Hye, Kinzler, Wendy L. "Chat generative pre-trained transformer: why we should embrace this technology". *American journal of obstetrics and gynecology* 228.6 (2023): 706–711.
22. Zhang, Yanhui, Pei, Haolong, Zhen, Shihan, LiQian, Liang, Fengchao. "Chat generative pre-trained transformer (ChatGPT) usage in healthcare". *Gastroenterology & Endoscopy* 1.3 (2023): 139–143.
23. Hasani, Amir M., Singh, Shiva, Zahergivar, Aryan, Ryan, Beth, Nethala, Daniel, Bravomontenegro, Gabriela, et al. "Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports". *European Radiology* 34.6 (2024): 3566–3574.
24. Ke Wang and Weihua Zhu. Eye tracking and gesture recognition system for intelligent human-computer interaction. *Engineering Intelligent Systems*. 2025, 33(4).
25. Fernandez, Manuel P., Rossini,Laureano A. Bulus, Cruz, José L., Andres, Miguel V.,Caso,Pablo A. Costanzo. "High-speed and high-resolution interrogation of FBG sensors using wavelength-to-time mapping and Gaussian filters". *Optics Express* 27.25 (2019): 36815–36823.
26. Haixia Guo. Address query and signal processing system based on artificial intelligence sensor array in time domain. *Engineering Intelligent Systems*. 2025, 33(4).
27. Niedzielski, Tomasz, and MichałHalicki. "Improving linear interpolation of missing hydrological data by applying integrated autoregressive models". *Water Resources Management* 37.14 (2023): 5707–5724.
28. Xu, Pengfei, and YinjieJia. "SNR improvement based on piecewise linear interpolation". *Journal of Electrical Engineering* 72.5 (2021): 348–351.
29. Nevskii, Mikhail Viktorovich. "On the Minimal Norm of a Projection Operator for Linear Interpolation on an-Dimensional Ball". *Mathematical Notes* 114.3 (2023): 415–418.
30. Li, Zewen, Liu, Fan, Yang, Wenjie, Pen, Shouheng, Zhou,Jun. "A survey of convolutional neural networks: analysis, applications, and prospects". *IEEE Transactions on Neural Networks and Learning Systems* 33.12 (2021): 6999–7019.
31. Cong, Iris, Soonwon Choi, and Mikhail D.Lukin. "Quantum convolutional neural networks". *Nature Physics* 15.12 (2019): 1273–1278.
32. Dubey, Shiv Ram, Chakraborty, Soumendu, Roy, Swalpa Kumar, Mukherjee, Snehasis, Singh, Satish Kumar, Chaudhuri, BidyutBaran. "diffGrad: an optimization method for convolutional neural networks". *IEEE Transactions on Neural Networks and Learning Systems* 31.11 (2019): 4500–4511.
33. Kuttan, Manjunath Omana, Zhou, Kai, Steinheimer, Jan, Redelbach, Andreas, Stoecker, Horst. "An equation-of-state-meter for CBM using PointNet". *Journal of high energy physics* 2021.10 (2021): 1–25.
34. Ding Chuanwei, Zhang Li, Chen Haoyu, Hong Hong, Zhu Xiaohua, "Fioranelli Francesco. Sparsity-based human activity recognition with PointNet using a portable FMCW radar". *IEEE Internet of Things Journal* 10.11 (2023): 10024–10037.
35. Hasan, MdRakibul, MdMahbub Hasan, and Md Zakir Hossain. "How many Mel-frequency cepstral coefficients to be utilized in speech recognition? A study with the Bengali language". *The Journal of Engineering* 2021.12 (2021): 817–827.
36. Cai, Haiyan, Jiang, Qingtang, Suter, Bruce W. "Analysis of adaptive short-time Fourier transform-based synchrosqueezing transform". *Analysis and Applications* 19.01 (2021): 71–105.
37. Begum, Mahbuba, Jannatul Ferdush, and Mohammad Shorif Uddin. "A Hybrid robust watermarking system based on discrete cosine transform, discrete wavelet transform, and singular value decomposition". *Journal of King Saud University-Computer and Information Sciences* 34.8 (2022): 5856–5867.