Research on Robust Reinforcement Learning Methods for Embodied Perception of Robotic Arms in Unstructured Environments

Zhiyao Wang
Hangzhou Dianzi University ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, China
**Corresponding Author*

Abstract: Embodied perception is a core capability for robotic arms to operate autonomously in unstructured environments. However, traditional methods struggle to effectively distinguish between lighting changes, sensor noise accumulation, and object occlusions due to complex disturbances, potentially leading to operational errors or task failure. To address these challenges, this paper proposes a robust optimization framework for embodied perception in robotic arms based on deep reinforcement learning, systematically evaluating the contributions of each module through ablation studies. The problem of perceptual robustness is modeled as a Markov Decision Process (MDP), with a multimodal attention fusion module designed to incorporate signal-to-noise ratio-aware perception, and contrastive learning employed to enhance feature discriminability. Perceptual confidence is introduced as a state variable, enabling a dynamic perception gain modulation mechanism that achieves end-to-end co-optimization of perception weights and action policies. Additionally, an intrinsic motivation exploration strategy based on prediction error is integrated to alleviate the challenge of sparse rewards in unstructured scenarios. A large-scale parallel training environment is built on the NVIDIA IsaacGym platform, using the UR5e robotic arm to generate diverse task scenarios involving extreme lighting conditions, texture variations, and random occlusions for grasping, placing, and assembly tasks. Ablation results demonstrate that the dynamic perception gain mechanism significantly improves success rates under both normal and strong interference conditions; the intrinsic motivation exploration strategy effectively enhances exploration efficiency in sparse-reward settings; and the multimodal attention fusion module provides a solid foundation for gain adjustment by estimating weight distributions. The complete framework outperforms standard baselines such as DDPG across all three task categories, maintaining low perceptual error levels even under extreme occlusion conditions.

Keywords: Unstructured Environment; Robotic Arm; Embodied Perception; Reinforcement Learning; Robustness Optimization

1. Introduction

Robotic manipulators are increasingly widely applied in industrial manufacturing, logistics warehousing and other fields. However, when deployed in unstructured environments, factors such as sudden illumination changes, target occlusion, and accumulated sensor noise often lead to unstable output of the perception system, which in turn affects the success rate and safety of operations. According to the research trends in recent years, the development of embodied intelligence has gradually shifted from algorithm verification under controlled environments to the exploration of adaptability in complex scenarios, and the continuous and reliable interaction of agents in dynamic environments is regarded as a fundamental problem that embodied systems must address to achieve practical applications, which provides a macro-level reference for the research on perceptual robustness of robotic manipulators[1,2]. At the perceptual level, multi-modal data fusion is generally recognized as an effective direction to improve the accuracy of task understanding[3]. Visual, force and distance sensors have respective characteristics in terms of information type and disturbance pattern, and their overall perception capability can theoretically be enhanced through complementarity. Nevertheless, there is still a

lack of effective solutions for automatically adjusting the fusion strategy when the reliability of a certain modality declines. Some studies have explored processing methods based on uncertainty quantification from the perspective of visual navigation, attempting to maintain system performance in the presence of perceptual noise, which provides insights for explicitly incorporating uncertainty into the decision-making loop[4]. At the control level, deep reinforcement learning has been applied to robotic manipulator visual servoing, joint motion optimization, dynamic obstacle avoidance and other tasks, initially demonstrating its potential for coping with uncertainty[5-7]. However, most of these works assume that observation signals are reliable, and when the perceptual input itself is distorted due to environmental interference, there is still limited discussion on whether policy networks can effectively distinguish signals from noise. Some studies have pointed out that explicitly modeling perceptual uncertainty is a key problem to improve the robustness of motion planning, and this view is also instructive for the robotic manipulator field. In addition, relevant explorations on adaptive prescribed performance control attempt to enhance anti-interference capability by constraining tracking errors, but the exploration efficiency problem caused by sparse rewards remains prominent in unstructured environments[8,9]. Research in the multi-agent communication field discusses coordination strategies from the perspective of information interaction, which provides an external perspective for addressing the coupling problem between perception and control.

Overall, existing studies have made respective progress in the directions of multi-modal fusion, reinforcement learning control and uncertainty modeling, but there are still limited attempts to integrate these directions into a unified framework, which enables perceptual quality to become an integral part of policy learning rather than an external prerequisite. This study attempts to incorporate perceptual confidence into the state space of reinforcement learning, and constructs a perceptual robustness optimization method for robotic manipulators oriented to unstructured environments by combining dynamic gain adjustment and intrinsic motivation exploration, so as to provide preliminary reference for related research.

2. Problem Modeling and Research Methods

2.1 Framework Overview

The problem of embodied perception robustness of robotic arms in unstructured environments is essentially a sequential decision-making problem of seeking the optimal action strategy in partially observable and dynamically changing states. Traditional methods treat perception and control as separate modules, where the perception system outputs a fixed format of pose estimation and the control system calculates the motion trajectory accordingly. This feedforward information transmission architecture has a fundamental flaw: when sudden light changes or occlusion cause visual input degradation, the perception module indiscriminately outputs low-quality estimates, and the control module lacks the ability to distinguish perception quality and can only generate actions based on incorrect information, ultimately resulting in grasping failure or collision accidents.

To fundamentally address the disconnection between perception and control, this study formalizes the robustness problem of embodied perception in robotic arms as a Markov decision process (MDP) and defines tuples $M=(S, A, P, R, \gamma)$. The key difference from existing reinforcement learning methods lies in the way the state space and action space are defined. The state space S not only contains traditional kinematic state quantities, but also introduces perceptual confidence vectors. It is defined as follows:

$$s_t=\left[\theta_t,\dot{\theta}_t,x_t,\dot{x}_t,c_{vision},c_{force},c_{lidar}\right]\in\mathbb{R}^{n_s} \quad (1)$$

Here, $\theta_t\in\mathbb{R}^6$ is the joint Angle vector, $\dot{\theta}_t\in\mathbb{R}^6$ is the joint Angle acceleration, $x_t\in\mathbb{R}^7$ is the end effector pose (position $p\in\mathbb{R}^3$, quaternion $q\in\mathbb{R}^4$), $\dot{x}_t\in\mathbb{R}^6$ is the end velocity. c_{vision}, c_{force}, $c_{lidar}\in[0,1]$ is for visual, force, and lidar perception confidence estimates, respectively. When directly incorporated c into the state space, perceptual quality is no longer an external premise of the system but an internal variable that the policy network can directly observe and utilize, providing a basis for the co-optimization of perception and control.

The setting of the action space A also reflects the innovative ideas of this study. In traditional methods, the motion space only includes joint torque or position increments. In this study, it is expanded to a combined space of motion control instructions and perceptual gain parameters:

$$a_t=[\tau_t,\alpha_{vision},\alpha_{force},\alpha_{lidar}]\in\mathbb{R}^9 \quad (2)$$

Here, $\tau_t\in\mathbb{R}^6$ is the expected joint torque, and $\alpha_{vision},\alpha_{force},\alpha_{lidar}\in[0,1]$ is the perceptual gain coefficients for each of the three modalities, satisfying the normalization constraints $\sum_i \alpha_i=1$. By simultaneously outputting action and gain, the policy network can autonomously adjust its dependence on each sensor mode based on the current perceptual confidence level.

The design of the reward function R takes into account both task completion and the quality of the perception policy:

$$R(s_t,a_t)=R_{task}(s_t)+\omega_1 R_{perception}(s_t,a_t)-\omega_2 C_{energy}(\tau_t) \quad (3)$$

Where R_{task} is the sparse positive reward for task success (grab success +1.0, place success +0.5); $R_{perception}$ is perceptual quality rewards, defined as the weighted sum of perceptual confidence at the current moment $R_{perception}=0.5c_{vision}+0.3c_{force}+0.2c_{lidar}$, encouraging the strategy to maintain a perceptual state of high confidence; C_{energy} is energy consumption penalties, $C_{energy}=//\tau_t//^2$ suppress unnecessary vigorous movement. $\omega_1=0.1$ and $\omega_2=0.01$ is the balance coefficient. The state transition probability $P(s_{t+1}\mid s_t,a_t)$ is implicitly defined by the simulation physics engine based on action and environmental parameters, with a discount factor γ of 0.99.

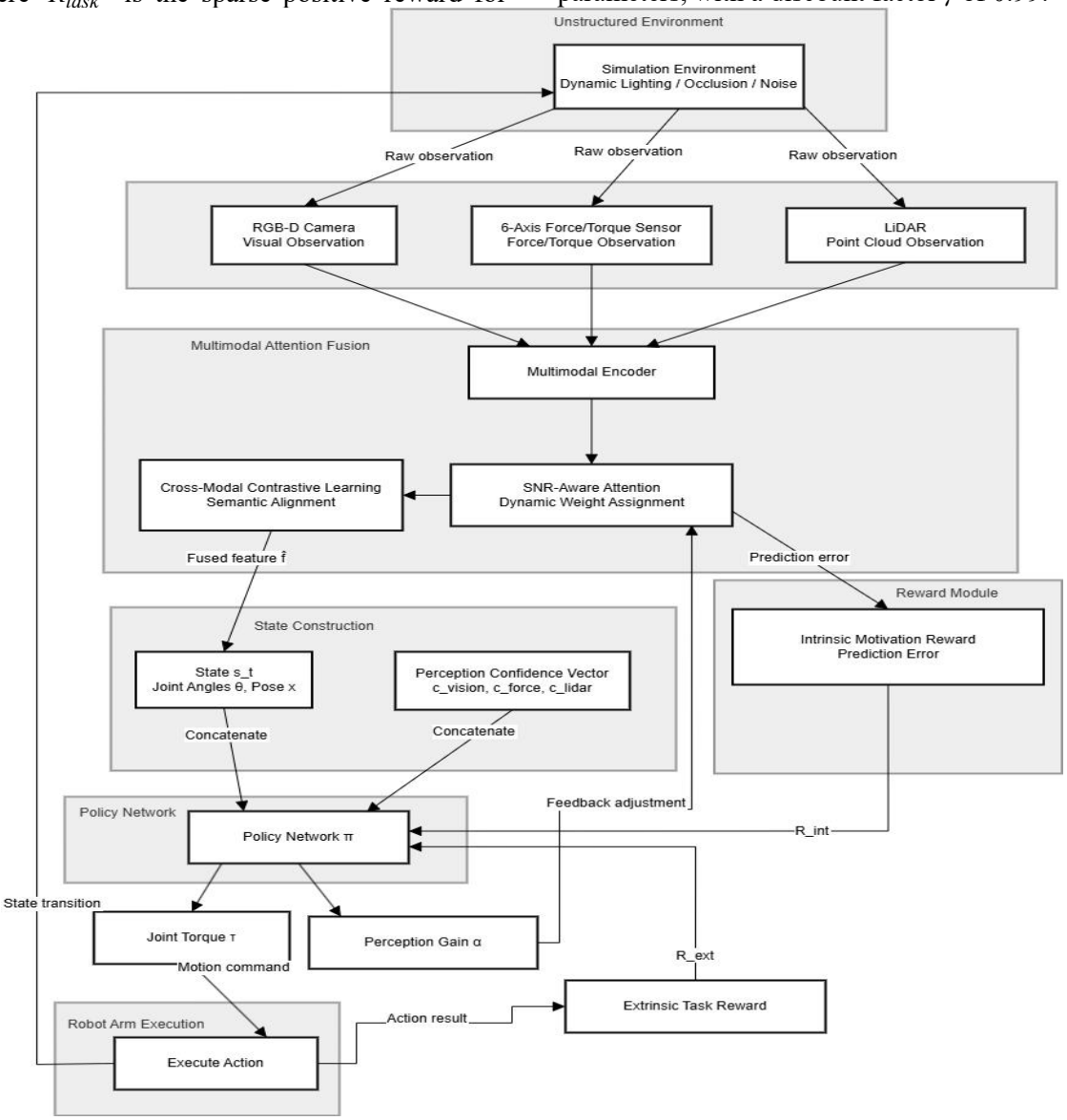

Figure 1. The Overall Research Framework

The overall research framework is shown in Figure 1. After the environmental states are extracted by the multimodal encoder, they are concatenated with the perceptual confidence vector and input into the policy network. The policy network simultaneously outputs the action and perceptual gains, and the gain parameters feedback adjust the modal fusion weights at the

next moment to form a closed-loop optimization.

2.2 Dynamic Perception Gain Mechanism

In the current mainstream deep reinforcement learning-based robotic arm control schemes, the sensor configuration is mostly determined in the design stage and remains unchanged throughout the task execution. For instance, in works such as visual servoing and single-arm intelligent control, the weights of the visual cameras remain fixed throughout the entire task, and the fusion ratios among different sensors also remain constant[6,7]. In terms of modeling perceptual uncertainty, existing studies have explicitly modeled perceptual noise at the planning level, but the connection between the perception module and the policy network is still feedforward[9]. The assessment results of perception quality are only used to adjust the confidence threshold and have not yet been fed back to the perception acquisition strategy itself. This architecture can still be maintained when the perception conditions are relatively stable, but when a certain modality significantly degrades, the system may have difficulty proactively adjusting its reliance on other modalities, thereby potentially affecting the reliability of decision-making.

In this study, perceptual confidence c_i is introduced as a state variable into the reinforcement learning loop, and the perceptual gain vector α is output by the policy network as part of the action. The physical intuition of this design is that the perceived quality of each mode of the robotic arm is constantly changing when performing operational tasks in an unstructured environment. For example, visual confidence c_{vision} may drop sharply when approaching highly reflective metal parts due to specular reflection; When touching soft objects, the force signal may show a low signal-to-noise ratio due to the contact force being too small. In such cases, fixedly relying on visual information and ignoring tactile feedback can easily lead to positioning bias. The dynamic gain mechanism enables the policy network to assess the status of each sensor in real time and adjust the dependency accordingly.

Perceptual confidence c_i is calculated by confidence estimators that are independent of each mode. Visual confidence c_{vision} assesses feature internal consistency using the scaled dot product in the self-attention mechanism:

$$c_{vision}=\sigma\left(W_v^T\cdot\mathrm{Softmax}(f_v\cdot f_v^T/\sqrt{d})\right) \quad (4)$$

Where f_v is the visual feature vector, d is its dimension, σ is the sigmoid function. The principle of this formula is that occlusion or blurring disrupts the local structure of the feature map, resulting in an abnormal distribution of self-attention weights, which then c_{vision} decreases. Force perception confidence c_{force} is based on signal variance estimation, and lidar confidence c_{lidar} is based on the proportion of effective point clouds. The policy network outputs the joint torque as well as the gain vector, and the overall optimization objective can be expressed as:

$$\max_\pi \mathrm{E}_\pi\left[\sum_{t=0}^{T}\gamma^t\left(R_{task}(s_t)+\lambda\sum_i\alpha_{i,t}\cdot c_{i,t}\cdot R_{modality,i}\right)\right] \quad (5)$$

When the confidence of a certain mode decreases, the product term $\alpha_{i,t}\cdot c_{i,t}$ guides the strategy to reduce the gain weight of that mode through the gradient signal. For example, when vision is severely degraded ($c_{vision}\approx 0$), $\alpha_{vision}\cdot c_{vision}$ tends to zero, forcing the network to shrink α_{vision} and increase α_{force}, driving the robotic arm to perform touch detection actions to obtain tactile information.

To illustrate the technical differences more clearly between this method and the traditional fixed-gain method, Table 1 conducts a comparative analysis from four dimensions.

Table 1. Comparison Between Dynamic Perceptual Gain Mechanism and Traditional Fixed Gain Methods

Comparison Dimensions	Traditional fixed-gain methods	The proposed dynamic gain mechanism
Gain adjustment mode	Predefined, the task remains unchanged throughout	Online adaptive adjustment, policy network output
Perception quality perception	No ability to assess perceived quality	Real-time estimation of confidence levels in each mode c_i
Control coupling	Perception is independent of the control loop	End-to-end collaborative optimization of perception and control
Failure response strategies	Unable to respond to modal degradation	Automatically switch to a reliable mode and trigger a detection action

2.3 Multimodal Attention Fusion Module

Multimodal fusion has garnered certain attention in the field of embodied intelligence. A review work on vision-language-action models systematically sorts out relevant fusion methods, and a study on indoor thermal environment control also attempts to incorporate multimodal perception with reinforcement learning for modeling[10,11]. Nevertheless, from the perspective of existing schemes, the fusion of heterogeneous sensor data still mostly adopts equal weight allocation or simple concatenation, and insufficient consideration has been given to the differences in signal-to-noise ratio and variations in information contribution of different sensors in dynamic environments. When environmental conditions fluctuate significantly, such a fixed fusion strategy may constrain the overall adaptability of the perception system.

The core of the multimodal attention fusion module designed in this study lies in the introduction of an attention mechanism for environmental signal-to-noise ratio awareness. The signal-to-noise ratio estimator g_{SNR} takes the original signal statistical features of each mode as input and outputs signal-to-noise ratio estimated value SNR_i:

$$SNR_i=g_{SNR}\left(\mu_i,\sigma_i^2,H_i\right) \tag{6}$$

Where μ_i and σ_i^2 are respectively the mean and variance of the signal, and H_i is the information entropy. Visual modes estimate clarity by extracting the proportion of high-frequency components using the Laplacian operator, force modes estimate contact stability by signal power spectral density, and lidar estimates scan quality by point cloud density variance. Attention weights β_i are generated by a softmax function with a temperature coefficient:

$$\beta_i=\frac{\exp(SNR_i/\tau)}{\sum_{j\in\{vidion\,,force\,,lidar\}}\exp(SNR_j/\tau)} \tag{7}$$

The temperature coefficient τ=0.1 is used to control the sharpness of the attention distribution: the smaller it is, the more concentrated the weight obtained by the high signal-to-noise ratio mode will be. When the visual sensor is severely degraded by specular reflection, the SNR_{vision} drops sharply which causes $\beta_{vision}\approx0$ and the network automatically shift attention to the force and lidar modes.

To further enhance feature discrimination, cross-modal contrastive learning losses are introduced in the module. For multimodal observation pairs (positive sample pairs) from the same target object, minimize their cosine distance in the high-dimensional hidden space; For observation pairs (negative sample pairs) of different objects, maximize their distances. The loss function is:

$$\mathscr{L}_{contrast}=-\ln\frac{\exp\left(\mathrm{sim}(z_i^{vision},z_i^{force})/\kappa\right)}{\sum_{j=1}^{B}\exp\left(\mathrm{sim}(z_i^{vision},z_j^{force})/\kappa\right)} \tag{8}$$

Where z_i^{vision} and z_i^{force} represent the projections of the i-th sample in the visual and force feature Spaces respectively, $\mathrm{sim}(\cdot,\cdot)$ is cosine similarity, κ=0.07 is temperature parameter, B is batch size. The intuitive meaning of the design is that the visual appearance of the same object should be similar to the tactile feedback in the semantic space, and the force features can provide anchoring constraints even if the light and viewing Angle changes cause the visual features to drift, helping the encoder learn modality invariant target representations. The fused composite feature vector $\hat{f}$ is derived from the weighted sum of the modal features by attention weights:

$$\hat{f}=\beta_{vision}\cdot f_{vision}+\beta_{force}\cdot f_{force}+\beta_{lidar}\cdot f_{lidar} \tag{9}$$

2.4 Intrinsic motivation-driven exploration strategies

In unstructured environments, key perceptual states including the edge contours of occluded objects, surface features at specular reflection angles, and the mechanical responses at interfaces of different materials are often difficult to be accessed incidentally by intelligent agents. Existing studies on dynamic obstacle avoidance and adaptive prescribed performance control primarily rely on extrinsic task rewards to drive policy learning, and generally achieve satisfactory performance under conditions with dense reward signals[7,12]. However, when the feedback provided by the environment is relatively sparse, the policy tends to remain in the already explored region, and states that are critical to perceptual robustness may not be sufficiently explored. This limits the generalization capability of the policy to a certain extent.

To address this issue, this study introduces an intrinsic motivation exploration module based on prediction error. The theoretical basis of this module is the curiosity-driven exploration mechanism: the agent has an intrinsic desire to explore state regions whose dynamic laws are difficult to predict. In practice, build an auxiliary

prediction network $f_\phi(s_t,a_t)$ that takes the current state and action as input to predict the state at the next moment s_{t+1}. Prediction error is defined as a measure of state novelty:

$$e_t=//f_\phi(s_t,a_t)-s_{t+1}\|_2^2 \quad (10)$$

Intrinsic rewards are defined as:

$$R_{intrinsic}=\eta\cdot(1-\exp(-\beta\cdot e_t)) \quad (11)$$

Where η=0.1 is the intrinsic reward scaling factor, β=1.0 is the saturation rate parameter. Using an exponential saturation form ensures that the intrinsic reward is bounded and prevents it from dominating the total reward signal. The prediction network f_ϕ is updated synchronously with the policy network, and the loss function is the mean square error:

$$\mathscr{L}_{pred}=\frac{1}{B}\sum_{i=1}^{B}\|f_\phi(s_i,a_i)-s_{i+1}\|_2^2 \quad (12)$$

The total reward function is composed of external task rewards and intrinsic motivation rewards:

$$R_{total}(s_t,a_t)=R_{extrinsic}(s_t,a_t)+\lambda_t\cdot R_{intrinsic}(s_t,a_t) \quad (13)$$

The mixture coefficient $\lambda_t=\lambda_0\cdot\exp(-\delta t)$ decays exponentially over time, initial value is λ_0=0.5, decay factor is $\delta=10^{-4}$. This decay design allows for extensive exploration led by intrinsic motivation in the early stages of training, and as the strategy gradually masters the laws of environmental dynamics and the prediction error naturally decreases, the influence of external task rewards gradually increases, guiding the strategy to converge from the exploration stage to the task optimization stage.

The difference between this approach and traditional exploration strategies lies in its goal-oriented nature. ϵ -greedy or entropy regularization methods explore the state space in a uniform random manner and are often less efficient. The mechanism based on prediction error drives the robotic arm to actively search for regions that are difficult for the perception model to predict - regions that typically correspond to high-information states such as boundary occlusions, material transitions, and nonlinear contacts, where the experience accumulated is most crucial for improving perception robustness.

2.5 Algorithmic Implementation

This section provides the complete pseudo-code of the algorithm to clearly demonstrate the integration and execution process of each module.

Algorithm: Dynamic Perception Gain Depth Deterministic Policy Gradient (DPG-DDPG)

Input: Range of environmental parameters, number of training rounds M, step size per round T

Output: Policy network parameter θ^π, Q network parameter θ^π

Initialize the policy network π(s|θ^π) and the Q network Q(s,a|θ^Q)

Initialize the target network π' and Q', θ^{π'}←θ^π, θ^{Q'}←θ^Q

Initialize the multimodal encoder f_v, f_f, f_l

Initialize the signal-to-noise ratio estimator g_SNR and predict the network f_φ

Initialize the experience replay pool D with capacity N=10^6

for episode = 1 to M do

Randomly sampling combinations of environmental parameters (illumination L, occlusion rate O, noise σ)

Reset the environment to obtain initial multimodal observations o_0

Fusion encoding → Initial state s_0

for t = 0 to T-1 do

a_t = [τ_t, α_t] ← π(s_t) + explore noise

Perform a_t to obtain the next observation o_{t+1} and the external

reward r_{ext}

c_{vision} ← Visual self-attention assessment

c_{force} ← Force perception variance assessment

c_{lidar} ← Point cloud density assessment

s_{t+1} ← Multimodal fusion coding (o_{t+1})

e_t = f_φ(s_t, a_t) - s_{t+1} \ n ²

r_{int} = η·(1-exp(-β·e_t))

r_{total} = r_{ext} + λ_t·r_{int}

Store transfer samples (s_t, a_t, r_{total}, s_{t+1}) to D

Randomly sample batch B from D = {(s_i,a_i,r_i,s_{i+1})}

y_i = r_i + γ·Q'(s_{i+1}, π'(s_{i+1}))

Update Q Network: L_Q = (1/B)·∑(y_i-Q(s_i,a_i))²

Update Policy Network: L_π = -(1/B)·∑Q(s_i, π(s_i))

Updated forecast network: L_{pred} = (1/B)·∑||f_φ(s_i,a_i)

-s_{i+1}||²

Update all parameters using the Adam optimizer

θ^{Q'} ← τθ^Q + (1-τ)θ^{Q'}

θ^{π'} ← τθ^π + (1-τ)θ^{π'}

s_t ← s_{t+1}

end for

λ_t ← λ_0·exp(-δ·episode)

end for

2.6 Experimental Design

The experimental platform is built on NVIDIA IsaacGym Preview 4 and uses the PhysX 5.1 physics engine for high-fidelity rigid body dynamics simulations. Run 1024 standalone environment instances simultaneously on a single NVIDIA RTX 4090 card through GPU parallelization capabilities. The robotic arm model is a Universal Robots UR5e six-degree-of-freedom collaborative robot, with kinematic and dynamic parameters set according to official specifications; The end effector is the Robotiq 2F-85 adaptive parallel gripper, with a maximum clamping force of 235 N and a stroke of 85 mm.

The sensor simulation configuration is shown in Table 2. The visual sensor is a wrist-mounted RGB-D camera with a resolution of 640×480, a field of view of 57°×43°, and a sampling frequency of 30 Hz; The force sensor is a six-dimensional force/torque sensor on the wrist, with a resolution of 0.1N and 0.005Nm, and a frequency of 100Hz; The lidar is located at the end, with a measurement range of 0.05-2 m, an angular resolution of 1°, and a frequency of 20 Hz.

Table 2. Configuration Parameters of the Simulation Sensor

Sensor type	Installation location	Parameters	Frequency
RGB-D camera	Wrist	640×480, FoV 57°×43°	30 Hz
Six-dimensional force/torque sensor	Wrist	Resolution 0.1 N/0.005 Nm	100 Hz
Lidar	End	Range 0.05-2 m, angular resolution 1°	20 Hz

Table 3. Configuration of Simulation Parameters for Unstructured Environments

Parameters	Range of distribution	Sampling method
Ambient light intensity	0-10,000 lux	Uniform distribution
Direction of light	[0°,360°] azimuth, [0°,90°] elevation	Evenly distributed
Occlusion ratio	0%-70% target area	Uniform segmentation
Speed of motion of the obstruction	0-0.3 m/s	Normal distribution (μ=0.1,σ=0.05)
Background texture type	Metal, glass, wood, fabric, check	Uniform dispersion
Standard deviation of visual noise	0.01-0.05	Uniform distribution
Force shift noise	0-0.5 N	Normal distribution (μ=0.2,σ=0.1)

Unstructured environmental features were simulated by randomly sampling multi-dimensional parameters per round, with parameter configurations shown in Table 3.

Task scenarios include three types of typical operations: Grab task - grab the target object from a random position on the workbench and lift 0.2m; Placement task - putting the grabbed object into the specified container; Assembly task - Insert the part into the corresponding shape of the base slot hole. Target object types cover 25 geometries, including cubes (side length 0.03-0.08 m), cylinders (radius 0.02-0.05 m, height 0.05-0.12 m), and irregular polyhedrons, with material properties covering different surface characteristics such as metals, plastics, and woods.

The evaluation metrics are defined as follows: Task success rate (SR) is the ratio of the number of successful attempts to the total number of attempts; Strong Interference Success Rate (SR-hard) is specifically used to count the performance of difficult scenarios with occlusion rates exceeding 50%; Perceptual errors are evaluated using root mean square error (RMSE) to measure the deviation between the predicted pose and the true value; The average number of steps completed reflects the efficiency of the strategy; The robustness decay coefficient is defined as the rate at which success decreases as the intensity of interference increases.

To systematically evaluate the independent contributions of each module, the experiment adopted an ablation study design and constructed the following comparison baselines: (a) DPG-DDPG (complete method); (b) NO-DPG (No dynamic perception gain, gain fixed as equal weight $\boldsymbol{\alpha_i=1/3}$); (c) NO-ATTN (No attention fusion, with simple splicing as an alternative); (d) NO-IM (No intrinsic motivation, using only external rewards); (e) Standard-DDPG (without any of the modules introduced in this study). Each group of experiments was repeated under five different random seeds (42, 123, 456, 789, 1024), with mean and standard deviation recorded. Statistical

significance was performed using the two-tailed t-test, with significance levels set as $p<0.05$.

2.7 Experimental Results and Analysis

Table 4 presents a comparison of the comprehensive performance of each method on the three tasks. The complete DPG-DDPG method achieved the best results on all four indicators. Compared with Standard-DDPG, the grab success rate increased by 25.3 percentage points and the assembly success rate increased by 28.9 percentage points. Under strong interference conditions, the success rate of DPG-DDPG (68.9%) was 37.4 percentage points higher than that of Standard-DDPG (31.5%), indicating that the robustness advantage of the proposed framework in unstructured environments becomes more significant as interference intensifies.

Table 4. Comparison of the Overall Performance of Each Method Across Three Tasks

Methods	Grab success rate (%)	Placement success rate (%)	Assembly success rate (%)	Strong interference success rate (%)
Standard-DDPG	62.3±3.1	58.7±2.8	45.2±4.2	31.5±4.8
No-IM	71.8±2.4	67.3±3.1	54.6±3.5	42.3±4.1
No-Attn	76.4±2.1	72.1±2.6	60.8±3.2	49.7±3.8
No-DPG	79.2±1.8	74.5±2.3	63.4±2.9	53.2±3.5
DPG-DDPG	87.6±1.5	83.8±1.9	74.1±2.4	68.9±2.7

Table 5. Quantification of the Independent Contribution of Each Module in Ablation Experiments

Comparison Paths	Involving modules	Regular contribution (Δ%)	Strong interference contribution (Δ%)
No-IM → No-Attn	Intrinsic Motivation Exploration	+4.6	+7.4
No-Attn → No-DPG	Multimodal attention fusion	+2.8	+3.5
No-DPG → DPG-DDPG	Dynamic perception gain	+8.4	+15.7

Table 5 quantifies the independent contributions of each module through ablation analysis. The dynamic perception gain mechanism made the most significant contribution, independently contributing an 8.4-percentage-point increase in success rate in the complete method and increasing to 15.7 percentage-points under strong interference conditions. This is because the perception quality fluctuates sharply in strong interference scenarios, and the fixed gain method cannot respond in time, while the dynamic gain mechanism ensures decision stability by adjusting sensor dependencies in real time. The intrinsic motivation exploration module contributed more under strong interference (7.4%) than in regular scenarios (4.6%), indicating that key perceptual regions with a large amount of sparse rewards require active exploration to cover. Although the independent contribution of the multimodal attention fusion module is relatively small (2.8%), it is the basis on which the dynamic gain mechanism works - without high-quality modal weight estimates, gain tuning will lose its basis.

Table 6 shows variations in perceptual error at different occlusions. As the occlusion rate increased from 0% - 20% to 60% - 80%, the positioning error of Standard-DDPG rose sharply from 3.2 mm to 22.1 mm (an increase of 590%), while that of DPG-DDPG increased only from 2.4 mm to 9.3 mm (an increase of 288%). In the 40% - 60% occlusion range, the error of DPG-DDPG (5.7mm) was 62.0% of that of No-DPG (9.2mm), demonstrating that the dynamic gain mechanism effectively suppressed the effects of visual degradation by switching to force-sensing and lidar modes. It is worth noting that even in the face of 60% - 80% severe occlusion, the 9.3mm error of DPG-DDPG still meets the general positioning accuracy requirements for grasping tasks (about 10-15mm), while the 22.1mm of Standard-DDPG far exceeds this threshold.

Table 6. Comparison of Perception Errors of Various Methods Under Different Occlusion Intensities

Occlusion rate	Standard-DDPG (mm)	No-DPG (mm)	DPG-DDPG (mm)
0% - 20%	3.2	2.8	2.4
20% - 40%	7.5	5.6	3.8
40% - 60%	14.3	9.2	5.7
60% - 80%	22.1	14.8	9.3

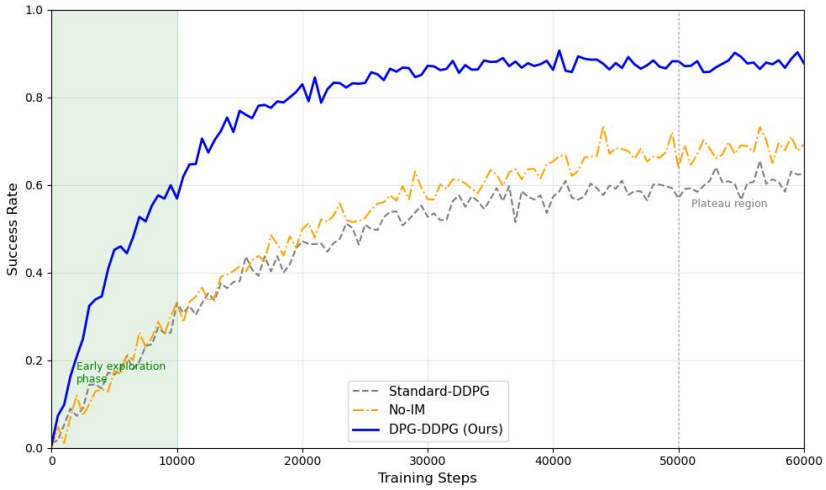

Figure 2. Comparison of Training Convergence Curves for Each Method

The convergence curves during training (Figure 2) further indicate that Standard-DDPG converges slowly, enters a plateau after approximately 50,000 steps, and ultimately achieves a success rate of less than 65%; The No-IM variant is similar to the Standard-DDPG in the early stages of training, but grows weakly in the later stages due to a lack of active exploration motivation; DPG-DDPG shows the characteristic of extensive exploration followed by rapid convergence in the early exploration phase (0-10,000 steps) driven by intrinsic motivation.

The experimental data above show that the proposed method outperforms the comparison baseline in four dimensions: overall performance, module contribution, convergence speed, and anti-interference ability. The dynamic perception gain mechanism, multimodal attention fusion module, and intrinsic motivation exploration strategy work in synergy at three levels: perception-control synergy, multimodal fusion, and exploration strategy, respectively, to form a highly robust robotic arm operation solution for unstructured environments.

3. Conclusions and Prospects

This study addresses the issue of insufficient robustness of robotic arm perception in unstructured environments and explores an optimization framework that integrates multimodal perception and deep reinforcement learning. Experimental results show that the dynamic perception gain mechanism brings about an 8 percentage point increase in success rate in normal scenarios, and the improvement is more significant under strong interference conditions; The multimodal attention fusion module provides a reference for weight estimation of gain adjustment; Intrinsic motivation exploration strategies alleviate the problem of sparse rewards to some extent. The complete framework outperformed the contrast methods in success rates on all three tasks and maintained relatively stable positioning accuracy under extreme occlusion conditions. Overall, the initial exploration of this study provides a possible idea for the robotic arm to shift from "passive perception" to "active adaptive decision-making", which needs to be further verified and improved in the future in virtual-real migration and actual deployment.

References

[1] Sun, C., Yuan, X., Wang, Y., & Liu, W. (2025). Embodied Intelligence-based Autonomous Unmanned System Technology. *Journal of Automatica Sinica*, *51*(4), 762-777.

[2] Zheng, N., Yang, M., Jiang, W., Sun, H., & Ding, N. (2025). Development Trends and Prospects of Embodied Intelligence. *Engineering Sciences*, 1-13.

[3] Li, H., Chen, Y., Cui, W., Liu, W., Liu, K., Zhou, M., Zhang, Z., & Zhao, D. (2026). A Survey of Vision-Language-Action Models for Embodied Manipulation. *Journal of Automatica Sinica*, *52*(1), 18-51.

[4] Wang, N. (2023). *Embodied Intelligent Visual Navigation Based on Entropy-Estimated Energy Models* Daliann Maritime University].

[5] Mao, J., Wang, Z., Zhou, X., Xia, F., & Zhang, C. (2025). Design and Experimental Verification of a Dynamic Obstacle Avoidance Algorithm for Robotic Manipulators Based on Deep Reinforcement Learning. *Experimental Technology and*

Management(4).

[6] Shao, Y., Zhou, H., Fan, X., Xu, Q., & Liu, N. (2024). An Intelligent Control Algorithm for Single Robotic Manipulators Based on Deep Reinforcement Learning. *Journal of Tianjin University of Technology*, *40*(6), 70-77.

[7] Yuan, Q., Qi, J., & Yu, H. (2025). Intelligent Visual Servo Control of Robotic Manipulators Based on Deep Reinforcement Learning. *Computer Integrated Manufacturing System*(3).

[8] You, X., Zhang, Y., Qiang, H., Xiang, G., Liao, Y., Guo, B., Zhong, Y., Fang, H., Zhao, T., & Dian, S. (2026). *An Adaptive Prescribed Performance Control Method for Robotic Manipulators Based on Reinforcement Learning*.

[9] Hu, C., Du, B., Wang, Z., Zhang, Q., Zhang, R., & Gao, H. (2025). Perceptual Uncertainty-Aware Motion Planning for Autonomous Driving Based on Adaptive Heuristic Reinforcement Learning. *Intelligent Transportation Systems, IEEE Transactions on*, *26*(8), 12676-12687.

[10] Dong, J., Zhou, L., Sun, Q., Xu, J., & Tang, Y. (2026). Bio-Inspired Intelligent Robotics Technology: From Morphological Structure to Embodied Intelligence. *Information and Control*, 1-19.

[11] Zou, Q., Tang, Y., Gao, B., Zhao, X., & Zhang, Z. (2025). Multi-agent Thinking Semi-multi-turn Communication Network Based on Deep Reinforcement Learning. *Journal of Control Theory and Applications*, *42*(3), 553-562.

[12] Xie, X., & Zhang, D. (2025). An Industrial Embodied Agent Demonstration Learning System Based on the Internet of Things. *Robot Technique and Application*(6), 53-56.

