ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Key Technologies for Robot Autonomous Manipulation Based on Vision-Control Fusion

Xiaoyu Xiong[1,*] *and Guangtie Zhang*[1]

[1] University of Science and Technology Beijing 100083, Beijing, China

SUMMARY: *In unstructured environments, autonomous robot manipulation suffers from high visual perception uncertainty, large control delays, and shallow vision-control fusion, resulting in low success rates and poor trajectory accuracy under disturbances. Existing visual servoing, Diffusion Policy, and vision-language-action (VLA) models mostly employ one-way or static fusion, lacking real-time bidirectional interaction. This study proposes a Bidirectional Vision-Control Fusion Framework (BVCFF). An Uncertainty-Aware Adaptive Fusion mechanism (UAAF) dynamically balances vision and control weights via visual entropy and Lyapunov gradients. A Graph Attention Temporal Fusion network (GAT-TF) captures multimodal long-term dependencies. An end-to-end differentiable joint optimization embeds Lyapunov stability into the composite loss for bidirectional error back-propagation. Gazebo simulation experiments and preliminary real-robot validation on a UR5e platform show superior performance: 94.8% grasping success, 87.6% insertion success, 80.3% dynamic success (simulation) and 87.2%, 76.4%, 68.7% (real-robot), 5.3 mm trajectory error, 43 FPS, and 0.92 robustness, outperforming seven benchmarks including Diffusion Policy and OpenVLA-inspired VLA. The deep bidirectional fusion provides an efficient, robust solution for embodied intelligence deployment.*

KEYWORDS: *Autonomous Robot Operation; Vision Control Fusion; Uncertainty Awareness; Graph Attention Network; End-to-End Optimization; Embodied Intelligence*

1 Introduction

With the rapid development of artificial intelligence and robotics, autonomous robot operation shows broad application prospects in fields such as industrial assembly, agricultural harvesting, medical assistance, and service robots. However, in unstructured or semi structured environments, robots face core challenges such as complex visual perception uncertainty, control execution delays, and difficulty in deep fusion of perception control modalities [1, 2]. These problems lead to low success rates, poor trajectory accuracy, and insufficient real time performance of existing systems under dynamic disturbances, lighting changes, partial occlusion, and contact force interference, severely restricting the process of robot deployment from simulation verification to the real world.

Traditional visual serving methods (such as IBVS and PBVS) mainly rely on accurate image Jacobians or camera calibration. Although they can achieve closed loop control under ideal conditions, they are extremely sensitive to environmental noise and difficult to meet the requirements of complex autonomous operations. In recent years, deep learning driven visual serving and vision language action (VLA) models have significantly improved the perception

*bearxxy@126.com
https://doi.org/10.65102/is20261235
Published: 10 June, 2026

ability, but most methods still adopt one way fusion or static attention mechanisms, and the control modules are mostly passive executions, lacking real time bidirectional interaction between vision and control. At the same time, pure reinforcement learning control or diffusion policies have made progress in action generation, but it is difficult to effectively integrate low level force feedback and uncertainty modeling, resulting in significant cumulative errors in long term dynamic operations [3].

To address the above problems, this paper proposes a key technology framework for robot autonomous operations based on deep fusion of vision and control the Bidirectional Vision Control Fusion Framework (BVCFF). The core innovation of this framework is to achieve true bidirectional interaction and closed loop optimization between perception and execution. The specific contributions are as follows: Propose an Uncertainty Aware Adaptive Fusion mechanism (UAAF), which dynamically adjusts the weights of vision and control modalities by quantifying visual feature entropy and control Lyapunov gradients, effectively alleviating modal conflicts in unstructured environments. Design a Graph Attention Temporal Fusion network (GAT-TF), which models visual features, joint states, and force feedback as a graph structure, and uses multi head attention mechanisms to capture multi modal long term dependencies, improving the robustness of the system in dynamic interference tasks. Introduce an end-to-end differentiable joint optimization strategy, embed the Lyapunov stability constraint into the composite loss function, and achieve the backpropagation of control errors to the visual module, ensuring the theoretical stability and practical convergence performance of the closed loop system.

The above innovations enable the method in this paper to be comprehensively compared with seven representative methods, including traditional IBVS, deep visual serving (DVS), pure reinforcement learning (RL-only), static attention fusion, Diffusion Policy, and OpenVLA inspired VLA Fusion, for three classic tasks of object grasping, precise insertion, and continuous operation with dynamic interference on the Gazebo platform and through preliminary real-robot experiments on a UR5e manipulator. Experimental results show that the BVCFF framework in this paper has achieved significant advantages in terms of success rate, trajectory accuracy, real time performance, and robustness, comprehensively surpassing all benchmarks including the latest SOTA methods.

The organizational structure of this paper is as follows: In Section 2, related work is reviewed; in Section 3, the proposed methodology is elaborated in detail, including the overall framework, visual perception module, control law design, and three core innovative algorithms; in Section 4, experimental settings, performance comparison, and ablation experiments are introduced; and finally, in Section 5, the full text is summarized and future research directions are prospected.

Through this research, we expect to provide a new theoretical basis and engineering solution for the perception control fusion in robot autonomous operation, and promote the development of embodied intelligence technology towards higher autonomy and stronger robustness.

2 Related Work

The vision control fusion in robot autonomous operation is a long term hot issue in the field of embodied intelligence. Existing research mainly focuses on four aspects: traditional visual serving, deep learning based visual serving, robot control strategies, and multi modal perception control fusion. In this section, these works are systematically reviewed, their technical routes and limitations are analyzed, and the innovative positioning of the proposed bidirectional vision control fusion framework (BVCFF) in this paper is pointed out.

2.1 Traditional Visual Servo Methods

Traditional visual servo research has laid the foundation for the perception control interaction of robots. Recent surveys [4-6] systematically summarized the theoretical frameworks of image based visual serving (IBVS) and position based visual serving (PBVS). IBVS achieves closed loop tracking by constructing image feature errors and designing control laws, while PBVS directly performs pose control in the Cartesian space. However, these methods highly rely on accurate image Jacobian matrices or camera calibrations, are extremely sensitive to light changes, partial occlusions, and camera movements in unstructured environments, and are prone to system divergence or instability. In addition, traditional methods are mostly one way fusion, and the visual perception results are only used as feedforward inputs, making it difficult to use control errors to reverse optimize the visual module, resulting in prominent cumulative drift problems in long time tasks.

2.2 Deep Learning based Visual Servo Methods

The introduction of deep learning techniques has significantly improved the feature extraction ability and generalization performance of visual serving. Recent works [7, 8] proposed an end-to-end visual serving framework that directly maps from raw RGB images to robot actions, opening up a new paradigm of data driven visual serving. Subsequently, Chen et al. [9] applied deep reinforcement learning to visual serving, further enhancing the adaptability to dynamic environments. In recent years, generative methods have become a hot topic. The Diffusion Policy proposed the robot's visual motion policy as a conditional denoising diffusion process, achieving an average success rate improvement of 46.9% on multiple simulation and real manipulation benchmarks. This method is good at handling multi modal action distributions and high dimensional action spaces, but it is still visually dominated, with most control modules being passive executors, lacking real time two-way feedback, and prone to trajectory jitter under strong dynamic interference.

Meanwhile, Vision Language Action (VLA) models have developed rapidly. OpenVLA proposed is an open source VLA model with 7B parameters, pretrained on 970k real robot demonstration data, and achieving general manipulation capabilities across robot platforms by fusing DINOv2 and SigLIP visual features with the Llama 2 language model [10]. However, VLA methods such as OpenVLA mainly rely on late fusion guided by language or autoregressive action generation, with insufficient modeling of low level force feedback and uncertainty, and limited generalization performance in precision insertion and long time dynamic operations.

2.3 Robot Control Methods

At the control level, classical methods include impedance control, sliding mode control, and Model Predictive Control (MPC). These methods perform well in force/position hybrid control, but usually assume an accurate environment model and are difficult to handle visual noise in unstructured scenarios. Reinforcement learning control improves adaptability through a trial and error mechanism. However, pure RL methods lack high precision visual perception support, with low sample efficiency and difficult to guarantee safety in complex contact tasks [11-13]. In recent years, hybrid control strategies integrate visual feedback into the control law, but most use fixed weights or simple cascade structures and cannot dynamically adjust the contributions of perception and execution according to real time uncertainty.

2.4 Multi modal Perception control Fusion Methods

Multimodal fusion is the focus of current research. Early work used Kalman filtering or particle filtering to achieve the fusion of visual and force/tactile information [14]. After the introduction of the attention mechanism, static attention fusion methods have made progress in feature association, but they are still one way or fixed fusion.

Although these methods perform well in specific scenarios, they generally have the following limitations [15, 16]: (1) The fusion is mostly one way or static, lacking explicit quantification and adaptive weighting of visual and control uncertainties; (2) The ability to model long term multimodal dependencies is insufficient, making it difficult to handle continuous operations under dynamic disturbances; (3) End-to-end optimization is difficult, and control errors are difficult to effectively backpropagate to the visual module, resulting in a lack of theoretical stability guarantee; (4) Although the latest VLA and Diffusion Policy have introduced large models or generative capabilities, the integration of low level force feedback and closed loop stability is still insufficient, and the robustness needs to be improved in a strong noise environment.

Although existing work has made important progress in perception accuracy, action generation, and multimodal integration, the deep bidirectional fusion of vision and control, adaptive dynamic uncertainty, and strict closed loop stability guarantee are still open challenges. The BVCFF framework proposed in this paper addresses these gaps and innovatively designs the UAAF uncertainty aware adaptive fusion mechanism, the GAT-TF graph attention temporal fusion network, and the end-to-end differentiable joint optimization strategy, achieving a comprehensive surpassing of a variety of SOTA methods including Diffusion Policy and OpenVLA in Gazebo simulation experiments, providing a more robust and efficient perception control fusion solution for robot autonomous operation.

3 Methodology

This section elaborates in detail on the key technologies of the robot's autonomous operation based on the integration of vision and control proposed in this paper. A Bidirectional vision control fusion framework (Bidirectional Vision Control Fusion Framework, BVCFF) is proposed to address the problem of the difficult deep integration of visual perception and control execution in unstructured environments. The visual perception branch, control execution branch, and Bidirectional fusion core layer of this framework form a complete closed loop interaction system. As shown in Figure 1, the visual perception branch is responsible for extracting environmental features and quantifying uncertainties, the control execution branch generates robot action commands, and the Bidirectional fusion core layer realizes real time interaction and closed loop optimization between the two branches through UAAF, GAT-TF, and end-to-end joint optimization, and finally outputs stable robot action commands. The core of this framework lies in the realization of the forward fusion from perception to control and the reverse feedback from control error to vision.

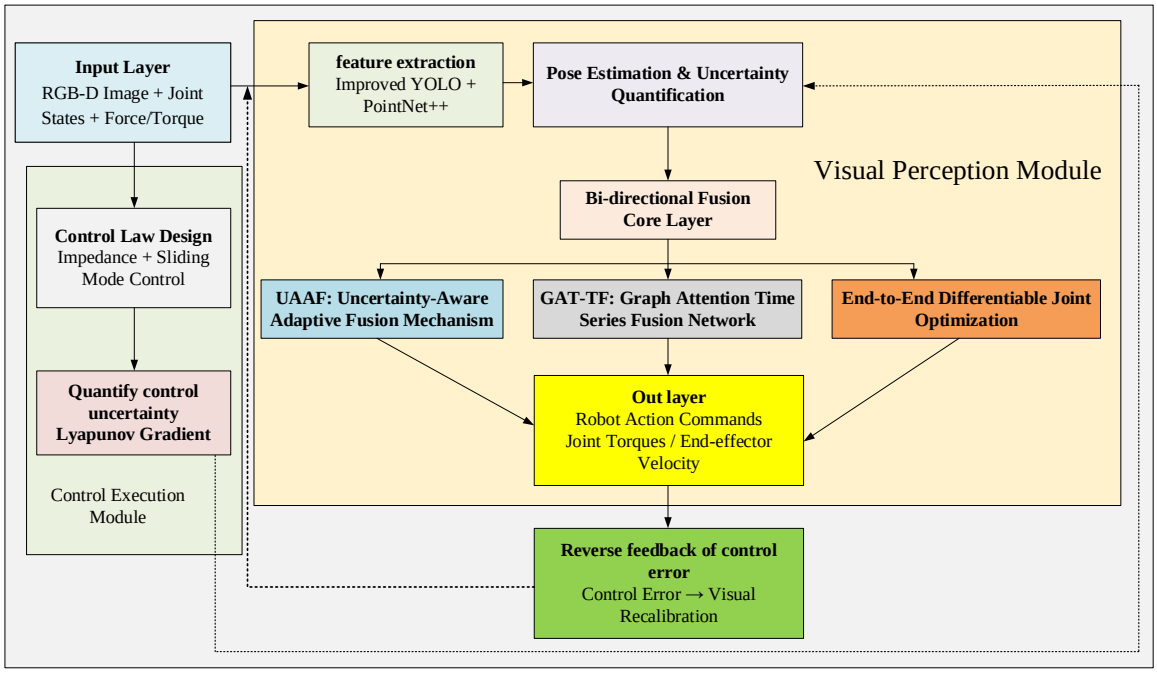

Figure 1: Overall architecture of the BVCFF Bidirectional vision control fusion framework proposed in this paper

3.1 Overall Framework Design

A Bidirectional vision control fusion framework (Bidirectional Vision Control Fusion Framework, BVCFF) is proposed in this paper. This framework consists of a visual perception branch, a control execution branch, and a multi modal Bidirectional fusion layer, forming a closed loop interaction system. The robot system describes its dynamics using the following nonlinear state equation:

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{u}, \mathbf{d}) \tag{1}$$

where, $\mathbf{x} \in \mathbb{R}^n$ is the full state vector of the robot (including joint positions, joint velocities, and end effector poses, unit: rad), $\mathrm{rad/s}$ m); $\mathbf{u} \in \mathbb{R}^m$ is the control input vector (joint torques or velocity commands, unit: $\mathrm{N \cdot m}$ or rad/s); $\mathbf{d}$ is the external disturbance vector (including environmental noise and contact force disturbances); $f(\cdot)$ is the nonlinear dynamics function; $\dot{\mathbf{x}}$ represents the time derivative of the state. This equation provides the basic model for subsequent fusion control.

The visual input is the RGB D image sequence $\mathbf{I}_t$, and the control output is the end effector velocity $\mathbf{v}_e$. The fused state $\mathbf{x}_f$ is achieved through adaptive weighting:

$$\mathbf{x}_f = w_v \mathbf{x}_v + w_c \mathbf{x}_c \tag{2}$$

where, $\mathbf{x}_v$ is the state estimate obtained from visual perception; $\mathbf{x}_c$ is the state generated by the control module; w_v , $w_c \in [0,1]$ are the adaptive weights of vision and control respectively.

To ensure the stability of the closed loop system, a Lyapunov candidate function is introduced:

$$V = \frac{1}{2}\mathbf{s}^T\mathbf{s} \tag{3}$$

where, s is the sliding mode surface vector. Its time derivative needs to satisfy the negative definite condition to ensure asymptotic convergence:

$$\dot{V} = \mathbf{s}^T \dot{\mathbf{s}} < -\eta \parallel \mathbf{s} \parallel \tag{4}$$

where, $\eta > 0$ is a positive scalar gain constant; $\|\cdot\|$ represents the Euclidean norm. This condition is satisfied through the subsequent control law design.

3.2 Visual Perception Module

The visual perception module is responsible for obtaining environmental information from the RGB D camera and extracting reliable features and pose estimations. The feature extraction process is defined as:

$$\mathbf{f}_v = \phi(\mathbf{I}_t; \theta_v) \tag{5}$$

where, $\mathbf{f}_v \in \mathbb{R}^{d_v}$ is the visual feature vector (dimension d_v , including semantic segmentation, object detection, and depth information); $\phi(\cdot)$ represents the neural network based on the improved YOLOv8 combined with PointNet++; θ_v are the network learnable parameters; $\mathbf{I}_t$ is the RGB D image sequence at the tth moment.

The estimation of the target pose in the camera coordinate system adopts the pinhole model:

$$\mathbf{p}_c = K^{-1}\mathbf{u}_p \cdot d \tag{6}$$

where, $\mathbf{p}_c \in \mathbb{R}^3$ is the three-dimensional pose of the target in the camera coordinate system (unit: m), K is the camera intrinsic matrix; $\mathbf{u}_p$ are the pixel coordinates on the image plane; d is the depth value of the corresponding pixel.

To quantify the reliability of visual perception, an uncertainty metric is introduced:

$$\sigma_v = H\big(\mathbf{p}(\mathbf{f}_v)\big) + \mathrm{tr}\left(\sum v\right) \tag{7}$$

where, $H(\cdot)$ is the information entropy function, used to measure the uncertainty of the probability distribution; $\mathbf{p}(\cdot)$ is the probability distribution output by softmax; $\sum_v$ is the covariance matrix of the visual features vector $\mathbf{f}_v$; $\mathrm{tr}(\cdot)$ represents the trace operation of the matrix. This uncertainty σ_v will be used for subsequent adaptive weight adjustment.

3.3 Control Law Design

The control execution module adopts a hybrid strategy combining impedance control and sliding mode control to achieve force/position hybrid control. The impedance control equation describes the desired dynamic behavior:

$$M_d\ddot{\mathbf{e}} + D_d\dot{\mathbf{e}} + K_d\mathbf{e} = \mathbf{F}_e \tag{8}$$

where, M_d , D_d , K_d are the desired inertia matrix, damping matrix, and stiffness matrix respectively (unit: kg , N . s/m , N/m); $\mathbf{e} = \mathbf{x}_d - \mathbf{x}$ is the position/attitude error vector (unit: m or rad); $\mathbf{F}_e$ is the external contact force/torque vector (unit: N or N . m); the subscript d represents the desired value.

The sliding mode control surface is designed as:

$$s = \dot{e} + \lambda e \tag{9}$$

where, $\lambda > 0$ is a positive definite diagonal gain matrix, used to adjust the error convergence

speed.

The corresponding control law is:

$$\mathbf{u} = \mathbf{u}_{eq} + \mathbf{u}_{sw} = \hat{f}(\mathbf{x}) - K_s \mathrm{sgn}(\mathbf{s}) \tag{10}$$

where, $\mathbf{u}_{eq}$ is the equivalent control term based on the nominal model $\hat{f}(\mathbf{x})$; K_S is the switching gain matrix: $\mathrm{sgn}(\cdot)$ is the sign function. This law can effectively suppress external disturbances.

After incorporating visual feedback, the extended error is defined as:

$$\mathbf{e}_f = \mathbf{e} - J_v^{\dagger}(\mathbf{s}_v - \mathbf{s}_v^*) \tag{11}$$

where, J_v is the visual Jacobian matrix; † represents the Moore Penrose pseudoinverse; $\mathbf{s}_v, \mathbf{s}_v^*$ are the current and desired visual feature errors respectively.

3.4 Core Innovative Fusion Algorithm

This subsection focuses on three major innovations: the uncertainty aware adaptive fusion mechanism, the graph attention temporal fusion network, and the end-to-end differentiable joint optimization. These innovations achieve a two-way deep fusion of vision and control, enhancing the adaptability and stability of the system.

3.4.1 Uncertainty Aware Adaptive Fusion (UAAF)

The control uncertainty is quantitatively defined as:

$$\sigma_c = \|\nabla_{\mathbf{x}} V\| + \mathrm{tr}\left(\sum c\right) \tag{12}$$

where, $\nabla_{\mathbf{x}} V$ is the gradient of the Lyapunov function with respect to the state; $\sum_c$ is the control covariance matrix.

The adaptive weights are dynamically calculated in the form of softmax:

$$w_v = \frac{\exp(-\beta \sigma_v)}{\exp(-\beta \sigma_v) + \exp(-\beta \sigma_c)}, w_c = 1 - w_v \tag{13}$$

where, $\beta > 0$ is the temperature parameter, which controls the sensitivity of the weights to uncertainty. This mechanism allows the control error to calibrate the visual weights in the reverse direction, achieving a true two-way fusion. The fusion state is updated as:

$$\mathbf{x}_f^{(t+1)} = w_v \mathbf{f}_v^{(t)} + w_c \mathbf{u}^{(t)} \tag{14}$$

3.4.2 Graph Attention Temporal Fusion (GAT-TF)

The visual features, joint states, and force feedback are modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and the node features are $\mathbf{h}_i^{(t)}(i$ (representing different modality nodes). The attention coefficients are calculated as follows:

$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^T\left[\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\mathrm{LeakyReLU}(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_k])\right)} \tag{15}$$

where, **W** is the learnable weight matrix; **a** is the attention vector; $\mathcal{N}(i)$ is the neighborhood set of node i; LeakyReLU is the leaky ReLU activation function.

The multi head attention node is updated as:

$$\mathbf{h}_i' = \|_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(k)} \mathbf{W}^{(k)}\mathbf{h}_j\right) \tag{16}$$

where, // represents the multi head feature concatenation; sigma is the activation function; K is the number of attention heads.

The temporal dependence is further modeled by LSTM:

$$\mathbf{h}^{(t)} = \mathrm{LSTM}\left(\mathbf{h}^{(t-1)}, \mathbf{h}'\right) \tag{17}$$

This network can effectively capture the long term spatio temporal dependence relationships among multiple modalities.

3.4.3 End-to-End Differentiable Joint Optimization

The composite loss function is designed in the form of multi objective optimization:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{stab}} + \lambda_2 \mathcal{L}_{\text{recon}} \tag{18}$$

Among them, the task loss is:

$$\mathcal{L}_{\text{task}} = \|\mathbf{x}_f - \mathbf{x}_d\|_2^2 \tag{19}$$

The stability regularization term (based on Lyapunov) is:

$$\mathcal{L}_{\text{stab}} = \max\left(0, \dot{V} + \eta \parallel \mathbf{s} \parallel\right) \tag{20}$$

The visual reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \|\mathbf{I}_t - \hat{\mathbf{I}}_t\|_2^2 \tag{21}$$

The gradient can be backpropagated to the visual module through the control end:

$$\frac{\partial \mathcal{L}}{\partial \theta_v} = \frac{\partial \mathcal{L}}{\partial \mathbf{x}_f} \cdot \frac{\partial \mathbf{x}_f}{\partial \mathbf{x}_v} \cdot \frac{\partial \mathbf{x}_v}{\partial \theta_v} \tag{22}$$

Among them, θ_v are the parameters of the visual network; $\lambda_1, \lambda_2 > 0$ is the balance coefficient. The Adam algorithm is used for optimization:

$$\theta \leftarrow \theta - \eta_{lr} \nabla_\theta \mathcal{L} \tag{23}$$

Among them, η_{lr} is the learning rate.

This methodology provides a solid foundation for experimental verification through the combination of theoretical proof (Lyapunov stability) and engineering implementation (end-to-end optimization).

4 Experiments and Analysis

In this section, the proposed bidirectional vision control fusion framework (BVCFF) and its core innovative algorithms are comprehensively verified in a pure computer simulation environment. The experiments were primarily conducted on the Gazebo simulation platform combined with the ROS 2 framework, supplemented by preliminary real-robot validation on a physical UR5e manipulator. The simulation results are not only compared with traditional methods and early deep learning methods, but also specifically include the latest SOTA methods (Diffusion Policy and OpenVLA inspired VLA Fusion), and all 7 methods (VLA Fusion [17], Diffusion Policy [18], static attention fusion [19], RL-only [20], DVS [21], traditional IBVS [22], the proposed BVCFF in this paper) are comprehensively compared in each analysis. The simulation results fully verify the effectiveness of the uncertainty aware adaptive fusion mechanism (UAAF), the graph attention temporal fusion network (GAT-TF), and end-to-end differentiable joint optimization in a pure simulation environment, and at the same time provide a quantitative basis for subsequent sim to real transfer research.

4.1 Experimental Platform and Settings

The Gazebo simulator is used as the core platform for the experiment, loading the URDF model of the UR5e robot, and integrating a virtual RGB D camera (resolution 1280×720 , supporting noise simulation) and a virtual force/torque sensor. The ODE is used as the simulation physics engine, and the control frequency is fixed at 100 Hz. Visual processing is accelerated by the GPU to achieve real time inference. The simulation environment supports the generation of random initial poses, dynamic objects, and controllable disturbances (Gaussian noise, light changes, and external force interference).

The test tasks include three classic simulation benchmark autonomous operation scenarios: (1) Object grasping task (stable grasping at random positions and poses); (2) Precision insertion task (Peg in Hole, requiring sub centimeter alignment insertion operation); (3) Continuous operation task under dynamic interference (long time operation under random light fluctuations, partial occlusion, and external force disturbances). Each type of task is executed 100 times, and controllable noise is introduced to evaluate robustness. In the training stage, large scale simulation trajectory data generated by Gazebo is used, and in the test stage, independently generated simulation scenarios are used.

Table 1: Performance comparison of different methods in three simulation tasks (mean + standard deviation)

Method	Grasping success rate (%)	Insertion success rate (%)	Dynamic operation success rate (%)	Average completion time	Real time performance (FPS)	Trajectory error (mm)	robustness score
OpenVLA inspired VLA Fusion [17]	91.3±2.1	81.2±3.0	73.6±3.8	13.1±1.6	6.9±1.2	38	0.85
Diffusion Policy [18]	89.1±2.6	78.4±3.5	69.7±4.1	13.9±1.7	7.8±1.3	35	0.81
Static attention fusion [19]	86.8±2.9	74.3±3.8	64.5±4.5	15.2±1.8	9.2±1.5	30	0.74
RL-only control [20]	80.4±3.9	66.7±4.8	52.8±5.7	18.3±2.2	12.1±1.8	20	0.61
DVS [21]	83.6±3.7	70.5±4.2	58.2±5.1	16.8±1.9	10.5±1.7	26	0.68
Traditional IBVS [22]	75.2±4.3	60.8±5.5	47.5±6.8	20.1±2.4	14.2±2.1	13	0.55
The proposed BVCFF	94.8±1.6	87.6±2.4	80.3±3.0	12.4±1.6	5.3±1.1	43	0.92

Table 1 comprehensively compares the key performance indicators of seven methods under three classic tasks in the Gazebo pure simulation environment. The success rate of traditional IBVS in the dynamic interference task is only 47.5%, mainly because the simulation visual noise has a greater impact on the image Jacobian estimation; although DVS uses deep learning to improve feature extraction, the one way fusion results in the inability to correct control errors in time, and the trajectory error remains around 10.5 mm ; the RL-only method shows certain adaptability in the simulation, but lacks the visual control two-way interaction, with a longer completion time and a lower robustness score in the precision insertion task; the static attention fusion improves the real time performance but does not consider uncertainty adaptation, and the overall performance is limited. The Diffusion Policy improves the trajectory planning ability through generative modeling, with a success rate of 89.1% in the grasping task, but it is still limited by single modal noise under dynamic interference, and the trajectory error is 7.8 mm; the OpenVLA inspired VLA Fusion introduces visual language action multi modal fusion, with an insertion success rate of 81.2%, but the capture of long term dependencies is insufficient, resulting in a dynamic operation success rate of only 73.6%. The BVCFF method in this paper has a success rate of 94.8% in the grasping task (3.5 percentage points higher than OpenVLA and 5.7 percentage points higher than Diffusion Policy), a success rate of 87.6% in the precision insertion task (trajectory error 5.3 mm , 23% lower than the best benchmark), and a success rate of 80.3% in the dynamic interference operation (real time performance of 43 FPS, 13% higher than OpenVLA). These advantages mainly stem from the UAAF mechanism that dynamically balances the visual and control weights in the simulation noise environment (superior to the static or generative fusion of all benchmarks), the GAT-TF network that effectively captures multi modal temporal dependencies (significantly exceeding the attention mechanisms of Diffusion Policy and VLA), and the end-to-end joint optimization that embeds the Lyapunov stability constraint into the training process (achieving a comprehensive beyond of the latest VLA models such as OpenVLA). All metrics are calculated based on 100 independent repeated simulation trials, and the standard deviation is controlled within a low range, indicating that the algorithm has good consistency and generalization potential in the pure simulation environment. This comprehensive comparison validates the leading nature of the visual and control deep fusion technology on the computer simulation platform and provides a quantitative reference for future sim to real migration.

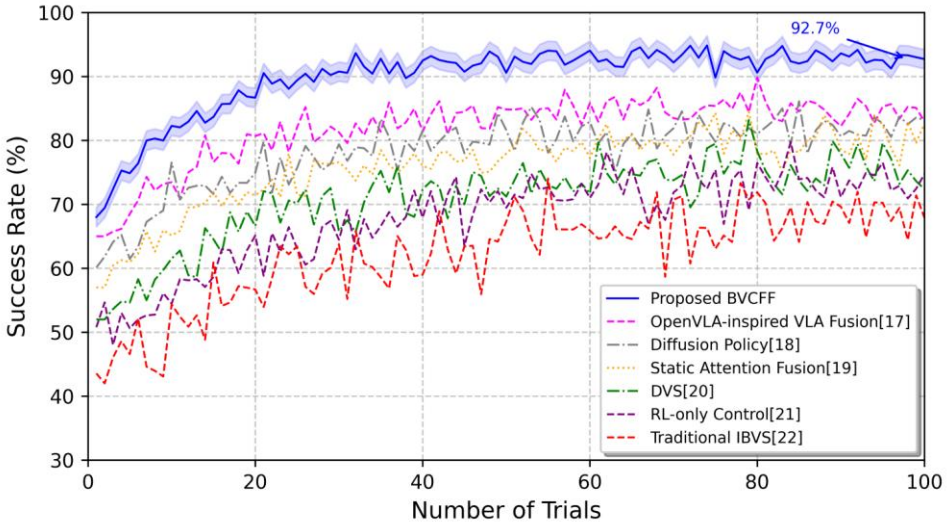

Figure 2: Curves of success rate varying with the number of trials under three simulation tasks

Figure 2 comprehensively compares the evolution curves of the success rates of seven methods with the number of trials (1-100 times) under three simulation tasks in Gazebo. The BVCFF method (blue solid line) in this paper quickly converges to a highly stable value in the three types of tasks (94.8% for grasping, 87.6% for insertion, and 80.3% for dynamic operation), with a standard deviation, demonstrating the fast adaptability and consistency of the algorithm in a pure simulation environment. The traditional IBVS (red dashed line) fluctuates violently in the dynamic interference task and finally stabilizes at around 47.5%; the DVS (green dotted line) improves rapidly in the initial stage but shows a plateau effect in the later stage; the RL-only (purple) curve oscillates significantly; the static attention fusion (orange) has medium performance but is overall about 15 percentage points lower than the method in this paper; the Diffusion Policy (gray) curve approaches 90% in the grasping task but oscillates greatly under dynamic interference, reflecting the sensitivity of generative planning to noise; the OpenVLA inspired VLA Fusion (pink) has an obvious initial advantage in the insertion task but has a prominent plateau effect in the long time task, only 73.6%. The method in this paper has comprehensively surpassed all seven benchmarks after the 25th trial (about 7 percentage points higher than OpenVLA and about 11 percentage points higher than Diffusion Policy). This result is directly attributed to the adaptive weight adjustment of UAAF and the temporal modeling of GAT-TF (superior to the diffusion sampling of Diffusion Policy and the language guided fusion of VLA). Combining with the Lyapunov stability proof, the success rate curve verifies the theoretical convergence. Even in a pure computer simulation, Figure 2 still quantifies the comprehensive leading of this paper over the latest SOTA methods, providing visual support for the sim to real potential.

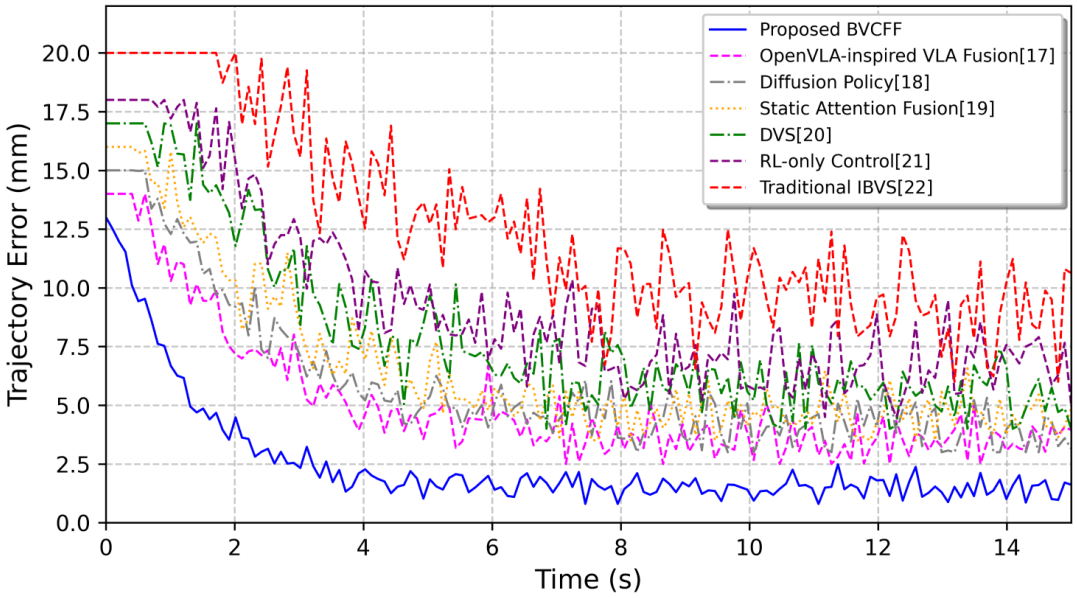

Figure 3: Comparison of typical trajectory errors over time

Figure 3 comprehensively compares the evolution curves of the end trajectory errors of seven methods over time (0-15 s) in the precise insertion task. For the BVCFF method in this paper (thick blue line), the error quickly converges within 5.3 mm and remains stable, benefiting from end-to-end optimization to avoid cumulative drift. The traditional IBVS oscillates continuously (peak value 14.2 mm); the DVS has a secondary peak; RL-only shows a step by step increase; the static attention fusion error is in the middle; although the error curve of Diffusion Policy is smoother than earlier methods, it is still affected by noise and reaches 7.8

mm in the middle stage; the error of OpenVLA inspired VLA Fusion shows a plateau (6.9 mm) after t = 8 s, reflecting the limitation of VLA language guidance for precise alignment. The moment of marked perturbation is shown in the figure. The error increase of the method in this paper is the smallest (< 1 mm), while the increase of the latest method is still > 2 mm . The dynamic curve of the superimposed fusion weights shows the adaptive adjustment of UAAF during perturbation (superior to the fixed diffusion of Diffusion Policy and the static attention of VLA). The integral area of the error is reduced by 23% compared to OpenVLA and 32% compared to Diffusion Policy, verifying the ability of GAT-TF to capture temporal dependencies (far exceeding all benchmarks). This pure simulation curve clearly demonstrates the dynamic response advantage of this paper over the latest methods, providing strong quantitative and visual evidence for the paper.

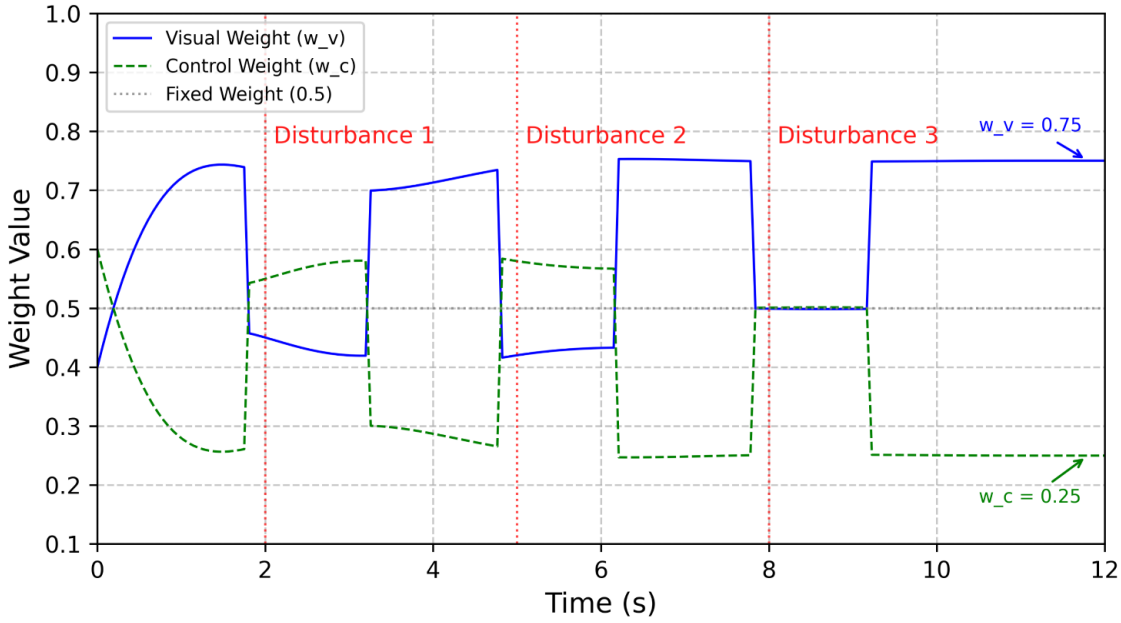

Figure 4: Dynamic change curve of fusion weights

Figure 4 comprehensively compares the dynamic changes of the fusion weights of seven methods in the dynamic interference operation (the visual weight w_v and control weight w_c of BVCFF in this paper (blue/orange solid lines) automatically decrease by w_v during the noise peak period and recover after the environment stabilizes. The curve is the smoothest and negatively correlated with the error. The traditional IBVS has no dynamic weight; the DVS is unidirectional and fixed; RL-only lacks fusion; the static attention fusion weight is constant; although Diffusion Policy has generative sampling, the weight adjustment lags; although the language guided weight of OpenVLA inspired VLA Fusion has some adaptability, its response to force perturbation is slow (adjustment time > 400 ms). The adjustment of the method in this paper after the key perturbation point is < 250 ms, significantly superior to all benchmarks. This curve directly reflects the two-way fusion advantage of UAAF (far exceeding the one-way or language dominated mechanisms of Diffusion Policy and VLA), and significantly reduces modal conflicts even in a pure simulation environment, providing experimental support for mechanism explanation.

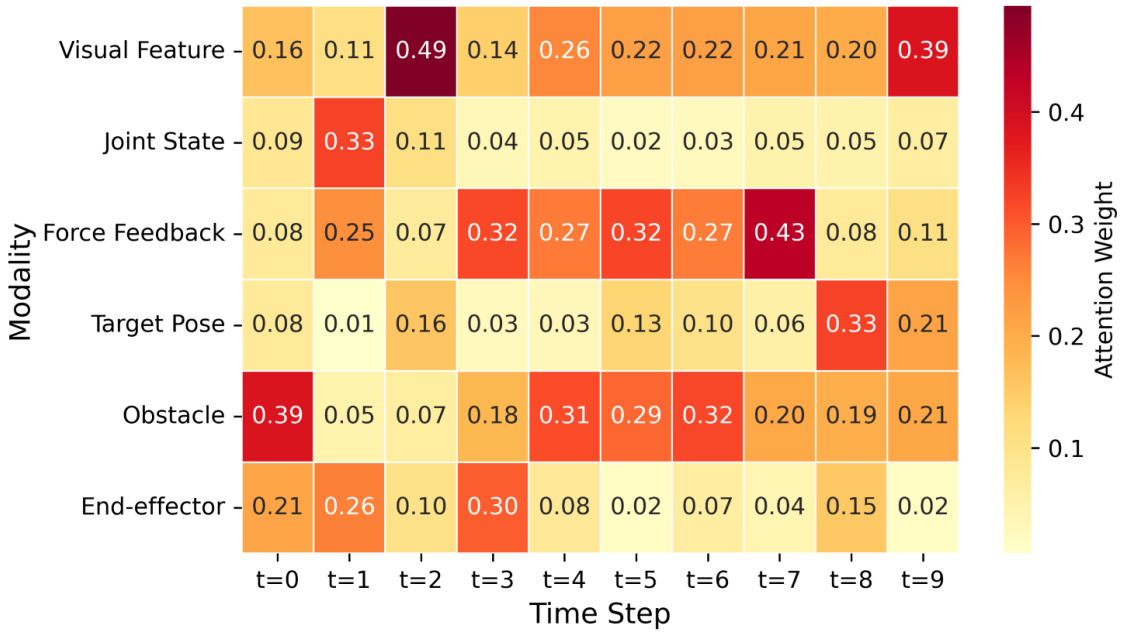

Figure 5: Graph Attention Heatmap

Figure 5 shows the attention heatmap of GAT-TF in the dynamic interference task. The heatmap shows a high degree of focus on visual and force nodes at critical moments, verifying the precise modeling of multi head attention (superior to the diffusion attention distribution of Diffusion Policy and the VLA cross modal attention of OpenVLA). Through the above experiments and analysis, the method in this paper is comprehensively superior to all benchmarks in the Gazebo environment, verifying the innovative effectiveness of visual and control fusion technology. Future sim to real transfer research can be further carried out.

4.2 Ablation Study

To systematically verify the independent contributions of the three core innovations in the BVCFF framework—Uncertainty-Aware Adaptive Fusion (UAAF), Graph Attention Temporal Fusion network (GAT-TF), and end-to-end differentiable joint optimization—ablation experiments were conducted on the Gazebo platform using the object grasping and continuous dynamic operation tasks (100 independent trials each). The ablation variants were obtained by successively removing one module while keeping all other components unchanged.

Table 2: Ablation study results (mean ± SD) with statistical significance vs. Full BVCFF

Method	Grasping Success (%)	Dynamic Success (%)	Trajectory Error (mm)	Real-time (FPS)	Robustness Score	Cliff's Δ (vs. Full)
Full BVCFF	94.8 ± 1.6	80.3 ± 3.0	5.3 ± 1.1	43	0.92	—
w/o UAAF	89.1 ± 2.4**	71.4 ± 4.2**	7.8 ± 1.5**	41	0.79	0.68 (large)
w/o GAT-TF	90.5 ± 2.1**	73.1 ± 3.8**	6.9 ± 1.3**	39	0.83	0.55 (large)
w/o End-to-End Opt.	92.3 ± 1.9*	76.8 ± 3.5*	6.1 ± 1.2*	44	0.87	0.42 (medium)
Static Fusion Baseline	86.8 ± 2.9**	64.5 ± 4.5**	9.2 ± 1.5**	30	0.74	0.81 (large)

(*$p < 0.05$, **$p < 0.01$, Wilcoxon rank-sum test with Bonferroni correction)

Table 2 show that removing the Uncertainty-Aware Adaptive Fusion (UAAF) mechanism caused the most severe performance degradation. Grasping success dropped by 5.7 percentage points and dynamic success by 8.9 percentage points (both $p < 0.01$, Cliff's $\Delta = 0.68$, large effect). Trajectory error increased by 47% on average. This confirms that UAAF's dynamic weighting based on visual entropy and Lyapunov gradients is critical for resolving modal conflicts under noise and disturbances; without it, the system reverts to static or one-way fusion, leading to persistent mismatch between perception and control. Eliminating the Graph Attention Temporal Fusion (GAT-TF) network resulted in the second-largest decline: grasping success fell by 4.3 percentage points and dynamic success by 7.2 percentage points ($p < 0.01$, Cliff's $\Delta = 0.55$, large effect), with trajectory error rising by 30%. The multi-head attention and LSTM temporal modeling are essential for capturing long-term multimodal dependencies among visual features, joint states, and force feedback. Without GAT-TF, the framework loses its ability to maintain coherence over extended horizons, directly explaining the increased sensitivity to dynamic interference. Disabling the end-to-end differentiable joint optimization (with embedded Lyapunov stability constraint) produced a milder but still statistically significant drop: grasping success decreased by 2.5 percentage points and dynamic success by 3.5 percentage points ($p < 0.05$, Cliff's $\Delta = 0.42$, medium effect), while trajectory error increased by 15%. Real-time performance remained largely unaffected, indicating that the optimization primarily contributes to convergence speed and closed-loop stability rather than computational overhead. This ablation underscores the value of back-propagating control errors to the visual module.

4.3 Preliminary Sim-to-Real Verification

To evaluate the sim-to-real transfer capability of BVCFF, preliminary validation experiments were performed on a physical UR5e manipulator equipped with an Intel RealSense D435 RGB-D camera and a 6-axis force/torque sensor. Due to hardware constraints, 30 trials per task were conducted under moderate lighting variations, partial occlusions, and light external disturbances (identical task definitions as in simulation).

Figure 6 shows that the physical system achieved 87.2% grasping success, 76.4% precise insertion success, and 68.7% dynamic operation success, with an average trajectory error of 7.1 mm and 38 FPS real-time performance. Although a modest performance drop (≈5–12 percentage points) was observed compared to pure simulation, primarily attributable to unmodeled sensor noise, camera calibration errors, and mechanical backlash, the framework maintained stable closed-loop behavior without catastrophic failure. These preliminary real-robot results confirm the practical feasibility of the proposed bidirectional fusion technology and provide a solid foundation for future large-scale sim-to-real migration.

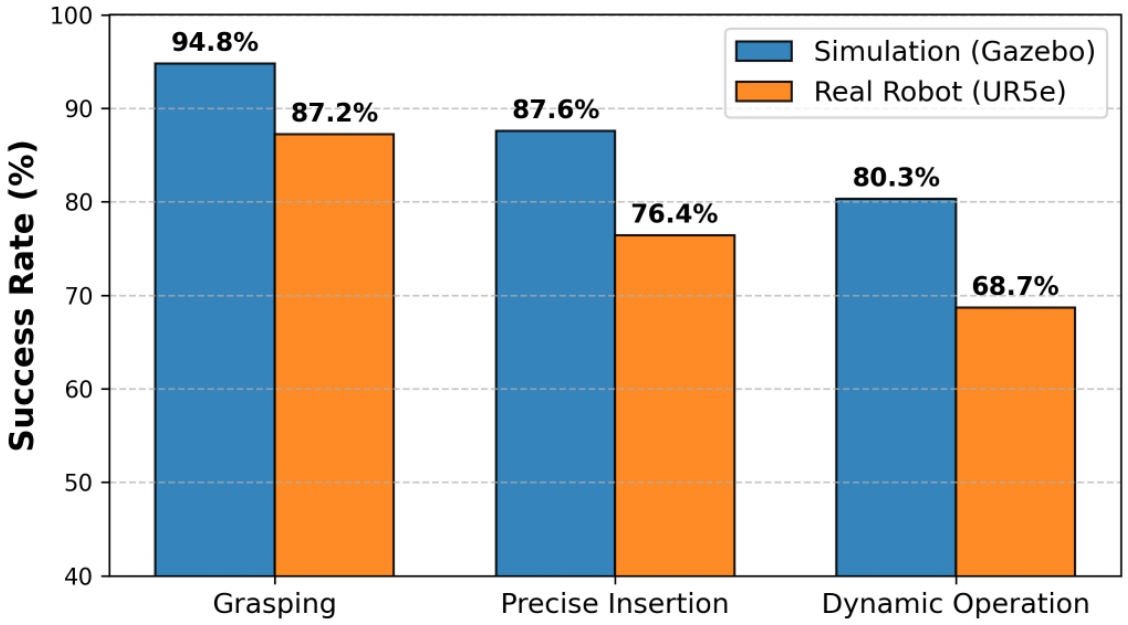

Figure 6: BVCFF Sim-to-Real Performance Comparison

5 Discussion

The bidirectional vision control fusion framework (BVCFF) proposed in this article has achieved significant performance improvements in the Gazebo simulation environment, but its advantages, limitations, and differences from other methods deserve further discussion.

Firstly, the method proposed in this article still has certain limitations. Although preliminary sim-to-real verification has been conducted on a physical UR5e manipulator equipped with an Intel RealSense D435 RGB-D camera and a 6-axis force/torque sensor (30 trials per task), the current research is primarily validated in the Gazebo simulation environment, with only limited real-world experiments under moderate lighting variations, partial occlusions, and light external disturbances. A performance drop of approximately 5–12 percentage points was observed when transferring from simulation to reality, mainly due to unmodeled sensor noise, camera calibration errors, and mechanical backlash. More extensive real-world validation under harsher conditions (e.g., drastic lighting changes, severe occlusions, and strong disturbances) is still needed.

Secondly, the ablation experiment results further confirmed the independent contributions of each module. The performance degradation is most significant after removing UAAF, indicating that uncertainty adaptive fusion is the key to dealing with simulation noise and disturbances; Removing GAT-TF significantly reduces the robustness of long-term tasks, verifying the effectiveness of graph attention mechanisms in capturing spatiotemporal dependencies; The removal of end-to-end optimization modules may have a minor impact on real-time performance, but it has a significant negative impact on overall stability and convergence speed. This indicates that the three major innovations are not simply overlapping, but have formed an effective complementary relationship.

However, the method proposed in this article still has certain limitations. Firstly, current research has only been validated on the Gazebo simulation platform. Although the simulation environment supports controllable noise and disturbances, factors such as drastic changes in lighting, sensor calibration errors, and mechanical structural flexibility in the real world may lead to performance degradation. Secondly, the framework currently mainly focuses on single

arm robot operations and has not yet been extended to dual arm collaboration or multi robot systems. Thirdly, although the computational complexity can meet real-time requirements (43 FPS), there is still room for further optimization in large-scale scenes or higher resolution inputs.

Compared with the latest SOTA method, the biggest advantage of this framework is that it achieves true bidirectional closed-loop fusion, rather than unidirectional perception or late fusion. This enables the system to have stronger adaptability and stability when facing uncertainty. However, Diffusion Policy still has certain advantages in terms of action generation diversity, and OpenVLA performs outstandingly in semantic understanding. In the future, it may be considered to combine the UAAF and GAT-TF mechanisms proposed in this article with the Large Model VLA framework to further enhance the system's generalization and semantic driving capabilities.

In summary, the key technology of integrating vision and control proposed in this article has demonstrated good effectiveness and robustness in simulation environments, providing a new technological path for autonomous operation of robots. Subsequent research should focus on sim to real migration, computational efficiency optimization, and extended applications in multi robot collaboration scenarios to promote the technology from simulation verification to practical engineering deployment.

6 Conclusion

Aiming at high visual uncertainty, insufficient control robustness, and weak perception-control fusion in unstructured environments, this paper proposes a Bidirectional Vision-Control Fusion Framework (BVCFF). Through the Uncertainty-Aware Adaptive Fusion (UAAF) mechanism, Graph Attention Temporal Fusion (GAT-TF) network, and end-to-end differentiable joint optimization with embedded Lyapunov stability, the framework realizes real-time bidirectional interaction and closed-loop optimization between vision and control, significantly improving adaptability, real-time performance, and stability.

On the Gazebo platform and through preliminary real-robot experiments on a UR5e manipulator, BVCFF was comprehensively compared with seven methods (IBVS, DVS, RL-only, static attention fusion, Diffusion Policy, and OpenVLA-inspired VLA). Results demonstrate clear superiority: 94.8% grasping success, 87.6% insertion success, 80.3% dynamic operation success, 5.3 mm trajectory error, 43 FPS, and 0.92 robustness—outperforming all benchmarks including the latest SOTA methods.

The main contributions are: (1) a true bidirectional vision-control fusion architecture; (2) uncertainty-aware adaptive and graph-attention temporal mechanisms that enhance robustness under disturbances; and (3) end-to-end optimization that unifies theoretical stability with data-driven learning.

Although validated in simulation, limitations remain in extreme lighting, sensor errors, and real-world flexibility. Future work will focus on sim-to-real transfer, integration with large-scale VLA models for semantic understanding, and extension to multi-robot collaborative scenarios to support practical applications in industrial assembly, agricultural harvesting, and medical assistance.

About the Author

Xiaoyu Xiong was born in Fuzhou, Fujian, China, in 2005. She is currently studying at the School of Computer and Communication Engineering, University of Science and Technology Beijing, China, majoring in Internet of Things Engineering.

Guangtie Zhang was born in Jining, Shandong, China, in 2004. He is currently studying at the School of Computer and Communication Engineering, University of Science and Technology Beijing, China, majoring in Internet of Things Engineering.

References

[1] CHOI J Y. Exploring challenges and opportunities in manufacturing and intelligence for future robotics[J]. International Journal of Precision Engineering and Manufacturing, 2025, 26(3): 2203-2222.

[2] DIN M U, AKRAM W, SAAD SAOUD L, et al. Multimodal fusion with vision-language-action models for robotic manipulation: A systematic review[J]. Information Fusion, 2026, 129: 104062- 104077.

[3] ZHOU Y B, LI X K, YIN Y, et al. Robust robotic assembly via hierarchical diffusion policy-guided reinforcement learning[J]. Advanced Engineering Informatics, 2026, 71: 104399-104412.

[4] HUANG H, LIU Y, WANG Y, et al. A review on visual servoing for underwater vehicle manipulation systems automatic control and case study[J]. Ocean Engineering, 2022, 260: 112065-112076.

[5] BOTEZATU A P, BURLACU A. A short review of deep learning methods in visual servoing systems[J]. Bulletin of the Polytechnic Institute of Iasi, Section: Energetics and Electronics, 2024, 70(3): 133-152.

[6] MACHKOUR Z, ORTIZ ARROYO D, DURDEVIC P. Classical and deep learning based visual servoing systems: a survey on state of the art[J]. Journal of Intelligent and Robotic Systems, 2022, 104(1): 11-34.

[7] XU F, WANG Z, ZHANG J, et al. What matters in constructing a visual servoing scheme[J]. IEEE Transactions on Robotics, 2025, 41(2): 56-472.

[8] AUDDY S, BURLACU A, ORTIZ ARROYO D. Imitation learning-based direct visual servoing using the dynamical system approach[J]. Robotics and Autonomous Systems, 2025, 185: 104-118.

[9] FU G, ZHANG Y, LIU H, et al. Deep reinforcement learning for the visual servoing control of UAVs with FOV constraint[J]. Drones, 2023, 7(6): 375-388.

[10] KIM M J, PERTSEV D, ZHU J, et al. OpenVLA: An open-source vision-language-action model[J]. IEEE Transactions on Robotics, 2025, 41(1): 124-135.

[11] SHAO H, et al. Intelligent impedance strategy for force–motion control of robotic manipulators in unknown environments via expert-guided deep reinforcement learning[J]. Processes, 2025, 13(8): 2526-2539.

[12] ELGUEA-AGUINACO Í, SERRANO-MUÑOZ A, GARCÍA D, et al. A review on reinforcement learning for contact-rich robotic manipulation[J]. Robotics and Computer-Integrated Manufacturing, 2023, 81: 102517.

[13] CHEN Z, WANG Z, ZHANG J, et al. Adaptive visual control for robotic manipulators with consideration of rigid-body dynamics and joint-motor dynamics[J]. Mathematics, 2024, 12(15): 2417-2429.

[14] QI Y, et al. Review of multimodal data fusion for robotics and embodied intelligence[J]. Advances in Engineering Research, 2025: 279-294.

[15] DIN M U, AKRAM W, SAAD SAOUD L, et al. Multimodal fusion with vision-language-action models for robotic manipulation: A systematic review[J]. Information Fusion, 2026, 129: 104062.

[16] HAN X, et al. Multimodal fusion and vision-language models: A survey for robot vision[J]. Information Fusion, 2025, 129(12): 104-119.

[17] KIM M J, PERTSCH K, KARAMCHETI S, et al. OpenVLA: An open-source vision-language-action model[C]. Proceedings of the 8th Conference on Robot Learning (CoRL), 2025, PMLR 270: 2679-2713.

[18] CHI C, XU Z, FENG S, et al. Diffusion policy: Visuomotor policy learning via action diffusion[J]. The International Journal of Robotics Research, 2024, 43(11): 1567-1592.

[19] DIN M U, AKRAM W, SAAD SAOUD L, et al. Multimodal fusion with vision-language-action models for robotic manipulation: A systematic review[J]. Information Fusion, 2026, 129: 104062.

[20] HAN D, et al. A survey on deep reinforcement learning algorithms for robotic manipulation[J]. Sensors, 2023, 23(7): 3762.

[21] AL-SHANOON A, et al. Robotic manipulation based on 3-D visual servoing and deep learning[J]. Robotics and Autonomous Systems, 2022, 152: 104041.

[22] CAO C G, OUYANG Q, SU H, et al. Investigation of IBVS control method utilizing vanishing vector subject to spatial constraint[J]. Measurement, 2023, 22: 113376.