arXiv:2508.19191v2 [cs.RO] 27 Aug 2025

AutoRing: Imitation Learning-based Autonomous Intraocular Foreign Body Removal Manipulation with Eye Surgical Robot

Yue Wang[1†], Wenjie Deng[1†], Haotian Xue[2], Di Cui[2], Yiqi Chen[3], Mingchuan Zhou[4*], Haochao Ying[5*], and Jian Wu[6]

***Abstract*— Intraocular foreign body removal demands millimeter-level precision in confined intraocular spaces, yet existing robotic systems predominantly rely on manual teleoperation with steep learning curves. To address the challenges of autonomous manipulation—particularly kinematic uncertainties from variable motion scaling and Remote Center of Motion (RCM) point variation—we propose AutoRing, an imitation learning framework for autonomous intraocular foreign body ring manipulation. Our approach integrates RCM dynamic calibration to resolve coordinate system inconsistencies caused by intraocular instrument variation and introduces the RCM-ACT architecture, which combines action chunking transformers with real-time kinematic realignment. Trained solely on stereo visual data and instrument kinematics from expert demonstrations in a biomimetic eye model, AutoRing successfully completes ring grasping and positioning tasks without explicit depth sensing. Experimental validation demonstrates the successful implementation of end-to-end autonomy under uncalibrated microscopy conditions. The results provide a viable framework for developing intelligent eye surgical systems capable of complex intraocular procedures.**

I. INTRODUCTION

Intraocular foreign body removal requires submillimeter precision to safely remove fragments near delicate retinal tissues while minimizing iatrogenic damage [1]–[4]. Robotic assistance has shown great promise in enhanced control and stability to reduce physiological tremors [5]–[7]. However, current systems still rely on manual teleoperation, which introduces steep learning curves and operational inefficiencies. While automated removal offers the potential for consistent high-precision execution, key challenges include accurately identifying optimal grasping points (pose determination for action initiation) and continuously evaluating fragment adhesion states in real time (action status assessment and adaptive path planning), which have so far impeded the deployment of automated solutions.

Recent advances in large-scale imitation learning have significantly progressed the automating robotic manipulation tasks [8]–[11]. For instance, imitation learning was used for two Franka robots employed in the peg transfer task in the fundamentals of laparoscopic surgery [12] and has also been applied to path planning in orthopedic surgeries [13]. However, the application of imitation learning in micro-scale eye robotic systems has not been widely studied [14]–[16]. Such systems are typically equipped with micromanipulators and advanced intraocular imaging capabilities [17], generating rich multimodal datasets during procedures. These datasets include stereomicroscopic video streams and instrument kinematics. Leveraging this data enables the training of autonomous control policies for eye microsurgical tasks. In this paper, we explore intraocular foreign body removal operation, the complex yet representative task, to assess the potential of imitation learning in ophthalmic procedures.

However, applying imitation learning to eye surgical robot introduces two distinct technical challenges not commonly encountered in other robotic applications: 1) *Variable Motion Scaling.* Intraocular foreign body removal systems allows surgeons to adjust the control-to-motion ratio to transition between gross movements and fine manipulations. This variability introduces kinematic uncertainty, as the same control input may result in different instrument motions depending on the scaling factor. Consequently, imitation learning policies trained on raw control inputs without accounting for these scaling variations may fail to generalize effectively during actual surgical procedures. 2) *Remote Center of Motion Variation.* In ophthalmic robotic surgery, instruments are introduced into the eye through a fixed entry point in the sclera, known as the Remote Center of Motion (RCM). Maintaining precise instrument motion around this pivot is crucial to prevent tissue damage. However, slight positional deviations or variations in the RCM point can disrupt the spatial coordinate system, leading to inaccuracies in instrument positioning. Unlike traditional robotic systems, recalibration is complex due to anatomical constraints.

In this work, we propose AutoRing, an imitation learning framework for autonomous intraocular foreign body removal

[1]Yue Wang and Wenjie Deng are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310012, China. They are also with the State Key Laboratory of Transvascular Implantation Devices of the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China, and Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Hangzhou 310058, China. (E-mail: ywang2022@zju.edu.cn)

[2]Haotian Xue and Di Cui are with the Dessight Biomedical, Hangzhou 310023, China.

[3]Yiqi Chen is with the Center for Rehabilitation Medicine, Department of Ophthalmology, Zhejiang Provincial People's Hospital (Zhejiang Key Laboratory of Precision Medicine for Eye Diseases & Affiliated People's Hospital, Hangzhou Medical College), Hangzhou 310014, China.

[4]Mingchuan Zhou is with the School of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310009, China.

[5]Haochao Ying is with the School of Public Health and Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310058, China. (E-mail: haochaoying@zju.edu.cn)

[6]Jian Wu is with the State Key Laboratory of Transvascular Implantation Devices of the Second Affiliated Hospital and School of Public Health, Zhejiang University School of Medicine, Hangzhou 310009, China. He is also with Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Hangzhou 310058, China.

[†]Equal contribution.

[*]Corresponding Authors: Haochao Ying and Mingchuan Zhou.

surgical manipulation. Our method trains control policies through expert demonstrations of grasping and placing foreign body in a biomimetic eye model, utilizing solely instrument kinematics and stereo visual data. To address the unavoidable variability in motion scaling inherent to microsurgical operations, we bypass kinematic ambiguity by directly training policies on instrument actuator-level displacements rather than raw control signals, ensuring generalization across magnification levels while remaining extensible to diverse microsurgical platforms. For the critical RCM variation challenge, we devise a dynamic coordinate alignment strategy: real-time recalibration of instrument positions against three anatomical fiducial markers converts variation-corrupted kinematic data into a unified global coordinate system via iterative rotation matrix updates. This software-defined spatial anchoring mitigates cumulative errors from RCM point deviations without requiring hardware recalibration, preserving submillimeter positioning accuracy. By achieving submillimeter-scale precision in instrument-ring interactions without explicit depth sensing (0.12 mm mean error in grasping localization), our study demonstrates critical feasibility for automated intraocular foreign body removal procedures under real-world kinematic uncertainties.

Our work makes three primary contributions: (i) The first imitation learning system enabling fully autonomous intraocular foreign body removal manipulation using eye surgical robot, demonstrating submillimeter accuracy; (ii) A novel RCM-based state representation resolving motion scaling uncertainties in microsurgical interfaces; (iii) Successful validation of end-to-end visuomotor policies for intraocular manipulation under uncalibrated microscopy conditions.

II. RELATED WORK

We analyze two related advanced areas: eye surgical robotic systems and imitation learning for surgical automation, concluding each with identified research gaps.

A. Eye Surgical Robotic Systems

Eye surgical automation has progressed through three distinct technological phases, each addressing specific limitations of prior systems. Initial first-generation platforms like the IRISS robot achieved 0.2 mm precision through mechanical Remote Center of Motion (RCM) constraints [18], yet necessitated complete human trajectory control due to lacking intraoperative sensing. Second-generation systems overcame this limitation by integrating micron-scale optical coherence tomography (OCT), exemplified by Ebrahimi et al. [19] whose vision-force fusion algorithm enhanced instrument localization accuracy by 37% without preoperative calibration. Current third-generation solutions demonstrate partial autonomy across three technical dimensions: (1) tissue interaction control via spectral OCT-guided robot reducing retinal tears by 41% through real-time elastography (10 mN tension detection) [20], (2) workflow optimization with systems like Preceyes Surgical System (PSS) cutting operative time by 28% [21], and (3) instrument guidance through Zhou et al. [22] hybrid robot achieving 12 μm needle placement accuracy via software RCM control, complemented by their 98.7% accurate deep learning-based needle detection [23]. However, these advancements remain fragmented—no system integrates all three dimensions simultaneously, existing autonomy operates in isolated surgical phases, and clinical validation is confined to porcine models. This fragmentation motivates our investigation into cross-modality integration to establish the first end-to-end automation framework for human clinical applications.

B. Imitation Learning in Robotics

Imitation learning has been widely explored in robotic manipulation tasks, enabling robot to learn from expert demonstrations by replicating human actions. A common method in early research was behavioral cloning, which predicts actions directly from observations. As deep learning and generative modeling advanced, new architectures like ConvNets and Vision Transformers (ViTs) [24] were employed for image processing, while recurrent networks and transformers [25]–[27] helped incorporate temporal dependencies in the observations. In addition, generative approaches, including energy-based models, diffusion models, and VAEs, were introduced to better capture complex motion patterns from demonstrations [28]–[31]. Among these advancements, Action Chunking with Transformers (ACT) [32] has emerged as a landmark framework for learning complex, long-horizon manipulation tasks from demonstrations. Diverging from traditional behavioral cloning, which predicts a single action at each timestep, the core idea of ACT is to predict a sequence of future actions, or an "action chunk." This approach leverages the powerful self-attention mechanism of Transformers to capture long-term temporal dependencies in the observation data, resulting in smoother and more coherent robot trajectories. While ACT has demonstrated remarkable success in tabletop manipulation tasks within structured environments, its direct application to microsurgical domains like intraocular procedures reveals a key limitation: its reliance on a consistent coordinate frame. The unique challenge of a shifting Remote Center of Motion (RCM) in ophthalmic surgery, which is not present in typical robotic setups, compromises the precision of the standard ACT framework and motivates the adaptations presented in this work.

Although end-to-end imitation learning theoretically holds the potential to address the complexities of surgical tasks, its effective application in the surgical domain remains an open question. Recent efforts have begun to apply imitation learning in surgical contexts [14]–[16], such as Kim et al. [33], which introduced the Surgical Robot Transformer (SRT) for surgical tasks using approximate kinematics data from the da Vinci system, overcoming challenges related to inaccurate proprioception and hysteresis. However, this approach is not applicable in intraocular foreign body removal surgery, as such procedures require precise contact with real-world objects and accurate localization. Moreover, challenges persist in applying imitation learning to fine-scale surgical tasks, particularly those requiring variable-scale manipulation.

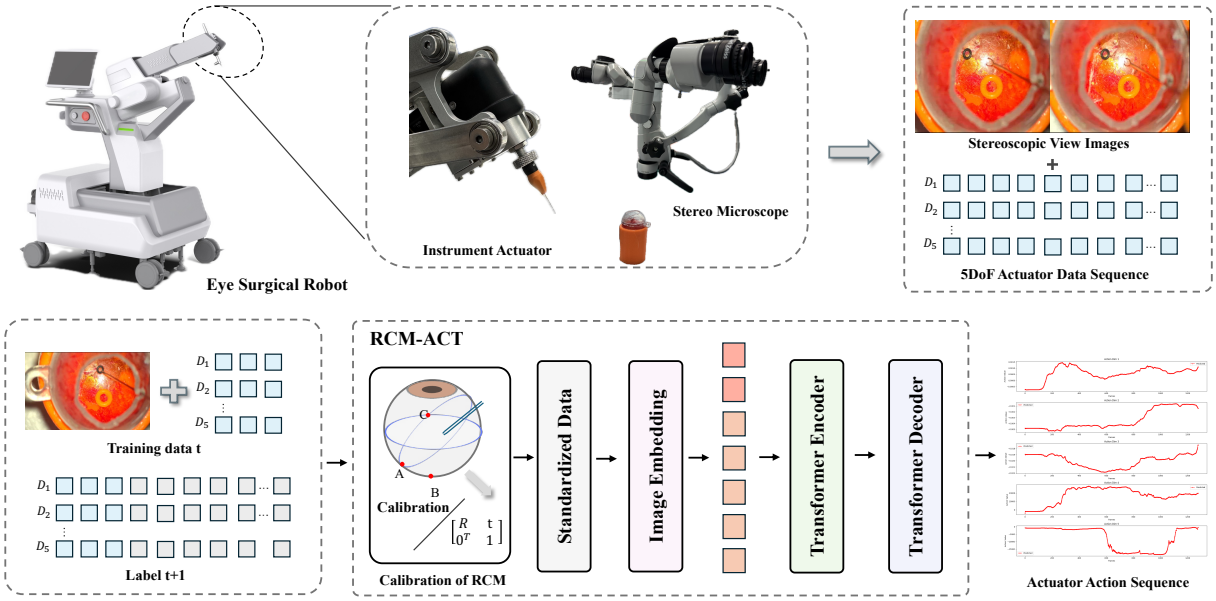


Fig. 1: Overview of RCM-ACT Framework: Instrument Actuation with Data Acquisition and Multimodal Training

III. METHOD

A. Task Definition

We use the classic ring grasp-and-place operation to simulate intraocular foreign body removal [34]. A robotic manipulator is used to manipulate a 3D eye model $\mathcal{M}$, consisting of two rings: a black ring R_{black} and a larger orange ring R_{orange}. The task involves automatically grasping the black ring R_{black} and placing it onto the orange ring R_{orange} with high precision. At each timestep t, the robot receives an observation o_t, which includes camera images of the model $\mathcal{M}$, showing the positions of the rings $\mathbf{p}_{R_{\text{black}}}[t]$ and $\mathbf{p}_{R_{\text{orange}}}[t]$, as well as the robot's end-effector pose $\mathbf{p}_t = (x_t, y_t, z_t)$ and orientation $\mathbf{r}_t$. The robot also receives proprioception x_t, representing the current arm pose. Based on these inputs, the robot computes an action a_t to move the end-effector to grasp the black ring R_{black} and place it on the orange ring R_{orange}. The task is learned from demonstrations, where each trajectory $\tau = \{(o_1, x_1, a_1), \ldots, (o_T, x_T, a_T)\}$ consists of a sequence of observations, proprioception, and actions. The robot aims to minimize misalignment and collision, ensuring precise placement of the black ring on the orange ring.

B. RCM Dynamic Calibration

In each task demonstration, during the data acquisition process, there exists a shift in the position of the point of insertion into the eye, causing a deviation in the position of the Remote Center of Motion (RCM). This results in a variation of the RCM point, meaning that the coordinate system for each data collection is inconsistent. To address this issue, we propose a dynamic calibration method for the RCM. RCM dynamic calibration is only performed when the RCM point varies or shifts. This variation typically occurs when the instrument is restarted or upon its entry into the eye, which can cause a shift in the point of insertion. The calibration is completed once at the beginning of each complete task episode, such as a full demonstration or a single autonomous execution, rather than continuously at every timestep.

The dynamic calibration is achieved by using the robotic instrument to reach three fixed points $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ within the workspace. By obtaining the positions of these three fixed points in the current coordinate system at each timestep, we are able to compute the corresponding rotation matrix $\mathbf{R}_t$ that aligns the local coordinate system to a global one. This rotation matrix $\mathbf{R}_t$ allows the transformation of the collected data from varying coordinate frames into a common reference frame. The process of reaching the three fiducial markers for calibration was performed under the same RCM constraint as the surgical manipulation itself. The robot's motion was pivoted around the established RCM point at the eye's entry, ensuring that the calibration procedure accurately reflects the kinematics of the actual task.

Mathematically, we can express this process as follows:

1) **Collection of Fixed Points:** At each timestep t, the instrument reaches three fixed points that are not co-planar $\mathbf{p}_1(t), \mathbf{p}_2(t), \mathbf{p}_3(t)$. These points are defined in the current coordinate system $\mathcal{C}_t$, where each point $P_i(t)$ is represented as:

$$p_i(t) = [x_i(t), y_i(t), z_i(t)]^T, \quad i = 1, 2, 3 \qquad (1)$$

2) **Transformation Calculation:** The rotation matrix $\mathbf{R}_t$ is computed using these three fixed points, and the translation vector d, is computed using these three corresponding point pairs. The goal is to align the local coordinate system $\mathcal{C}_t$ with a global coordinate system $\mathcal{C}_0$. This can be achieved by comparing the relative orientations of the fixed points in both coordinate

Algorithm 1: RCM-ACT Training with Calibration

Input : Raw demonstration dataset $\mathcal{D} = \{\tau_1, \tau_2, \dots\}$, chunk size k, weight β
Output: Updated model parameters θ, ϕ

Phase 1: Dataset Calibration
for *each demonstration episode* τ *in* $\mathcal{D}$ **do**
  Compute the single rotation matrix R_τ for this episode by reaching three fixed points;
  for *each timestep* t *in* τ **do**
    Realign observation o_t and proprioception x_t to the global coordinate system $\mathcal{C}_0$ using R_τ;
  end
end
Let $\mathcal{D}_{\text{calibrated}}$ be the dataset of all realigned episodes;
Phase 2: Model Training
Initialize encoder $q_\phi(z|a_{t:t+k}, \overline{o_t})$ and decoder $\pi_\theta(\hat{a}_{t:t+k}|o_t, z, \mathcal{C}_0)$;
for *each training step* **do**
  Sample a batch of trajectories $(o_t, a_{t:t+k})$ from $\mathcal{D}_{\text{calibrated}}$;
  Sample z from $q_\phi(z|a_{t:t+k}, \overline{o_t})$;
  Predict $\hat{a}_{t:t+k}$ from $\pi_\theta(\hat{a}_{t:t+k}|o_t, z, \mathcal{C}_0)$;
  Compute reconstruction loss: $L_{\text{reconst}} = \text{MSE}(\hat{a}_{t:t+k}, a_{t:t+k})$;
  Compute regularization loss: $L_{\text{reg}} = D_{\text{KL}}(q_\phi(z|a_{t:t+k}, \overline{o_t}) \parallel \mathcal{N}(0, I))$;
  Update θ, ϕ with ADAM optimizer using loss $L = L_{\text{reconst}} + \beta L_{\text{reg}}$;
end

systems. The rotation matrix $\mathbf{R}_t$ is found by solving:

$$p_i(t) = R_t \cdot p_i(0) + d_t, \quad i = 1, 2, 3 \tag{2}$$

where $\mathbf{p}_i(0)$ is the position of the fixed point $\mathbf{p}_i$ in the global coordinate system $\mathcal{C}_0$.

3) **Recalibration of the Coordinate System:** After computing $\mathbf{R}_t$, we can then apply it to all subsequent data collected at timestep t. Specifically, for any observation o_t and proprioception x_t taken at timestep t, we can recalibrate the data to the global coordinate system $\mathcal{C}_0$ by applying the inverse of the rotation matrix:

$$p' = R_t^T(p - d_t), \tag{3}$$

By recalibrating each new data collection point using $\mathbf{R}_t$, we ensure that all data can be transformed into a consistent coordinate system, allowing for more reliable task learning and accurate performance evaluation across different demonstrations.

C. RCM-ACT Architecture

In current imitation learning methods, tasks are typically modeled based on the relative position of the robotic manipulator, and actions are executed accordingly. The standard ACT framework excels at imitation learning by predicting future action chunks, but its performance presumes a stable and consistent motion base for the robot. This assumption is violated in the intraocular surgical environment, where a significant challenge arises from the variation of the robot's motion base—the Remote Center of Motion (RCM)—due to slight shifts in the instrument's scleral entry point. This RCM variation makes it difficult for an unmodified ACT policy to maintain the submillimeter accuracy required for the task. To address this issue, we propose the RCM-ACT model, as illustrated in Figure 1, which tightly integrates our RCM Dynamic Calibration module with the ACT framework. Before the policy network predicts a future action sequence, we first leverage the dynamic calibration to transform all kinematic data from varying coordinate frames into a unified global coordinate system, $\mathcal{C}_0$. This software-defined spatial anchoring effectively mitigates the kinematic uncertainty caused by RCM point deviations, allowing the model to learn a robust manipulation policy that is invariant to such shifts.

To train the RCM-ACT model for a new task, we begin by collecting human demonstrations via rostopic for data transmission and dataset recording. The action a_t at each timestep is defined as the target state vector for the robot's end-effector, represented by $a_t = [x, y, z, \theta, \dot{g}]^\top$, where $[x, y, z]$ represents the Cartesian coordinates, θ is the gripper's rotation angle, and $\dot{g}$ denotes the gripper's opening/closing state. This target state is then translated into low-level actuator commands by the robot's internal controller. Instead of mapping joystick or controller inputs to robot actions—a mapping that is directly affected by the scaling factor—our policy learns a direct mapping from the current state (composed of stereo visual data and instrument kinematics) to the desired next state (represented as instrument actuator-level displacements). The input to the system includes stereo imagery from a stereo microscope, which provides real-time visual feedback from two camera views. In addition to the visual data, the input also contains three values representing the cartesian coordinates of the current position and the rotation encoder value of the gripper. These proprioceptive and visual observations are essential for the task as they offer real-time sensory data about the robot's state and environment.

Once the data is collected, we train the RCM-ACT model to predict future actions based on the current observations. The predicted actions correspond to the target joint positions for robotic arms at the next timestep. The ACT model, which is trained to predict the next sequence of actions based on current states, mimics how a human operator would proceed given the current observations. These predicted target joint positions are then used to guide the robotic arms in achieving the desired configurations.

The RCM-ACT model combines the traditional ACT framework with RCM dynamic calibration. The ACT model consists of an encoder $q_\phi(z|a_{t:t+k}, \overline{o_t})$, which learns a latent variable z conditioned on the action sequence $a_{t:t+k}$ and the observation sequence $\overline{o_t}$, where $\overline{o_t}$ represents the observation without image data. The decoder $\pi_\theta(\hat{a}_{t:t+k}|o_t, z, \mathcal{C}_0)$ generates the predicted future actions $\hat{a}_{t:t+k}$ based on the current

observations o_t ,the latent variable z and the global coordinate system $\mathcal{C}_0$. The model is trained to minimize the reconstruction loss $L_{\text{reconst}} = \text{MSE}(\hat{a}_{t:t+k}, a_{t:t+k})$, which measures the error between the predicted and actual actions, as well as the regularization loss, which is the Kullback–Leibler (KL) divergence between the latent space distribution and a standard Gaussian distribution $L_{\text{reg}} = D_{\text{KL}}(q_\phi(z|a_{t:t+k}, \overline{o_t}) \parallel \mathcal{N}(0, I))$, which ensures that the latent space is regularized to follow a Gaussian distribution.

During the training process, the robot's position and orientation are recalibrated using RCM dynamic calibration to ensure that the data is aligned with the global coordinate system. This calibration step is crucial for compensating for the potential variation in the motion base point, ensuring that the model can execute the task reliably. Once the model is trained, it is used to predict future actions based on the current state and the predictions are executed by the robot to perform the task.

The integration of RCM dynamic calibration with ACT is vital for achieving smooth, precise, and reliable robotic motion. By combining these two components, we ensure that the robotic system can perform tasks effectively, even in environments where the motion base point may vary or change over time.

D. Inferring phase

In this section, we describe the inference process utilized in the robot-assisted task for foreign body ring placement, where the robotic manipulator is responsible for automatically picking up a black ring and placing it inside a larger orange ring within a scaled eye model. This process involves acquiring data from the robot's sensors, performing dynamic calibration, and predicting actions that guide the robot in completing the task.

Before the inference process begins, the robot's position $\mathbf{p}_t = (x_t, y_t, z_t)$ and orientation $\mathbf{R}_t$ are updated based on the calibration data from the RCM sensors. This ensures that the robot operates within a well-calibrated coordinate system, maintaining precision during the task. The rotation matrix $\mathbf{R}_t$ is then used to realign the data to the global coordinate system $\mathcal{C}_0$, ensuring that all subsequent data is accurately positioned within the global frame of reference.

Once calibration is completed, the policy inference process begins. At each timestep t, the robot's observation o_t is obtained, which includes both sensor data and camera images. The trained policy π_θ is used to predict the next action $\hat{a}_{t:t+k}$, which spans k time steps. The predicted actions are stored in a First-In, First-Out (FIFO) buffer for future predictions. Then, the actions at each timestep are determined by applying a weighted average over the stored actions. The weight w_i decays exponentially as time moves backward, ensuring that more recent actions have a stronger influence on the current decision.

During inference, the robot's data (such as proprioception and actions) is normalized using pre-defined statistics, ensuring that the data fed into the model aligns with the expected range for accurate predictions. After the model predicts the actions, the data is denormalized back to its original scale to ensure that the robot's actions correspond accurately to the real-world environment in which it operates.

Algorithm 2: RCM-ACT Inference

Input: Trained policy π_θ, episode length T, weight m

Output: Actions for the robot, based on dynamic calibration and inference

Episode-Start Calibration:
Compute the rigid body transformation, consisting of a rotation matrix R and a translation vector d, for the current episode;

Initialize FIFO buffers:
Initialize FIFO buffers $\mathcal{B}[0:T]$, where $\mathcal{B}[t]$ stores actions predicted for timestep t;

for *each timestep* $t = 1, 2, \ldots, T$ **do**
  Obtain observation o_t from the camera and proprioception data at timestep t;
  Realign the current observation o_t to the global coordinate system $\mathcal{C}_0$ using the pre-computed transformation (R, t);
  Predict action $\hat{a}_{t:t+k}$ using the policy $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$, where $z = 0$;
  Add predicted action $\hat{a}_{t:t+k}$ to buffers $\mathcal{B}[t]$;
end

Inference:
Obtain current step actions $A_t = \mathcal{B}[t]$;
Apply action $a_t = \sum_i w_i A_t[i] / \sum_i w_i$, with $w_i = \exp(-m \cdot i)$;

Return action to robot via rostopic:
Publish the computed action through the model output rostopic;

Robot Execution:
The robot uses the received action to perform the corresponding operation;

Finally, the computed action is communicated back to the robot through a rostopic labeled "model output," allowing the robot to execute the task of picking up the black ring and placing it inside the orange ring, as required. This process integrates real-time calibration with action prediction to ensure that the robot successfully performs the task with precision and efficiency.

E. Experiment Setup

1) Task: The robotic system was tasked with autonomously manipulating foreign body rings in a simulated intraocular environment, requiring precise grasping of a 1.22 mm black ring (R_{black}) and threading it onto a 4.02 mm orange ring (R_{orange}). The task specifications included positioning accuracy < 0.15 mm and an operation time limit of 20 ± 5 s. The eye phantom was mounted on a 7-DoF motion platform to simulate anatomical variations during ophthalmic surgery.

2) Data Collection: We collected 30 demonstration episodes (112 ± 18 s each) with randomized ring positions, generating 112 GB of training data. To facilitate the high-precision manipulations required for the task, a motion scaling factor of 1/30 was applied during the expert demonstrations, mapping the operator's hand movements from the controller to the instrument's motion. The data acquisition utilized ROS-based pipelines with asynchronous recording:

- Proprioceptive data (5D end-effector state $\mathbf{q}_t = [x, y, z, \theta, \dot{g}]^\top \in R^5$) sampled at 60 Hz via rostopic
- Stereo vision streams (1920×2160@60 Hz) captured using stereo microscope cameras
- Raw sensor data stored in HDF5 format without temporal alignment

3) Training of Models: The RCM-ACT model was trained for 2000 epochs on an NVIDIA RTX 4090 GPU using AdamW optimizer ($\alpha = 3\times10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.95$) with cosine learning rate decay. Training configurations included: 32-sequence batches (8 timesteps each), dropout rate 0.1, and KL-divergence regularization ($\beta = 0.5$). Model selection used five independent test sets (30 episodes each).

4) Deployment and Inference of Model: Figure 2 illustrates the actual deployment environment and workflow diagram. Real-time inference was deployed on a mobile workstation (Intel Core® i9-13900HX, NVIDIA® RTX 4090 Laptop GPU) running ROS Noetic, where the trained policy generated action sequences (x, y, z, θ, $\dot{g}^\top$) at 5 Hz through the `/model_output` rostopic. A dual-thread architecture achieved parallel execution: 1) *Perception Thread* asynchronously processed stereo images (OpenCV 4.5.5) and proprioceptive states at 30 Hz, 2) *Control Thread* performed RCM dynamic calibration with a 15 Hz generation speed for action prediction. Temporal misalignment from unsynchronized training data was mitigated by a weighted action buffer employing exponential decay:

$$\mathbf{a}_t = \sum_{k=0}^{N} m^k \mathbf{a}_{t-k} \quad (m = 0.8,\ N = 3), \tag{4}$$

F. Experiment Results

To quantitatively evaluate the policy's performance, we use the following metrics, reported in Table I.

- **Mean Squared Error (MSE):** This metric measures the fidelity of the policy's predicted action trajectory to the expert's ground-truth trajectory on a held-out test set. It is calculated as:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T} \|a_{\text{predicted},t} - a_{\text{expert},t}\|_2^2, \tag{5}$$

  where a_t is the 5D action vector $[x, y, z, \theta, \dot{g}]^T$ at timestep t, and T is the total number of timesteps in the episode. The expert trajectory is used strictly for evaluation and is not available to the policy during inference.
- **Grasp Deviation:** This measures the positional accuracy at the critical moment of grasping the ring. It is defined as the Euclidean distance between the 3D position of the autonomous agent's gripper and the expert's gripper at the frame of grasping:

$$\text{Grasp Deviation} = \|p_{\text{robot, grasp}} - p_{\text{expert, grasp}}\|_2, \tag{6}$$

  where p_{grasp} represents the 3D coordinate vector $[x, y, z]$ of the end-effector when the grasp is initiated.
- **Grasping Latency:** This metric evaluates how well the policy replicates the timing of the expert's actions. It is the absolute difference in frame number (timestep) between when the autonomous agent initiates the grasp versus when the expert did in the demonstration:

$$\text{Grasping Latency} = |t_{\text{robot, grasp}} - t_{\text{expert, grasp}}|. \tag{7}$$

  where t_{grasp} is the timestep at which the grasping action occurs.

The results in Table I reveal significant performance variations across methods in both deployment and evaluation metrics, emphasizing the critical role of dynamic RCM calibration and resampling mechanisms. In deployment tests, the baseline ACT method achieved partial success, grasping the black ring twice and placing the orange ring three times across five trials, but failed to complete the full task sequence due to insufficient coordination between subtasks. The RCM-ACT variant without resampling performed marginally better, with one successful grasp and four placements, yet also failed to achieve full-task completion. In contrast, the full RCM-ACT model demonstrated robust reliability, successfully grasping the black ring three times, placing the orange ring five times, and completing the entire task sequence in three out of five trials. The RCM-ACT with encoder variant could not be deployed effectively, as encoder-induced noise caused excessive mechanical jitter, rendering it inoperable for practical execution.

For evaluation metrics, the full RCM-ACT model outperformed all variants, achieving the lowest mean squared error of 0.014988, indicating precise alignment with expert trajectories. This result significantly surpassed the baseline ACT and the non-resampling RCM-ACT variant, which exhibited higher errors of 0.019876 and 0.022512, respectively. Grasp deviation, a critical measure of positional accuracy during ring interaction, further validated the superiority of RCM-ACT over 40% lower than other methods. Latency analysis revealed that RCM-ACT operated with near-real-time responsiveness at 17.50 frames, while the baseline ACT and the encoder-based variant suffered substantial delays of 48.50 frames and 102.00 frames, respectively. The extreme latency in the encoder variant underscores the destabilizing effects of noise on real-time performance.

G. Parameter Sensitivity

We systematically evaluate the impact of the chunk size parameter on ring task performance in Figure 3. To facilitate cross-metric comparisons, all measured values are normalized to the [0,1] range. When increasing the chunk size from smaller to medium values, latency decreases significantly,

TABLE I: Performance Comparison Between Deployment Results and Evaluation Tests

Method	Deployment			Model Evaluation		
	R_{black} Grasp	R_{orange} Place	Task Success Rate	MSE	Grasp Deviation (mm)	Grasping Latency (frame)
ACT	2	3	0/5	0.019876	1.182	48.50
RCM-ACT w.Encoder	-	-	-	0.017307	-	102.00
RCM-ACT w/o Resample	1	4	0/5	0.022512	1.386	94.25
RCM-ACT	3	5	5/10	0.014988	0.686	17.50

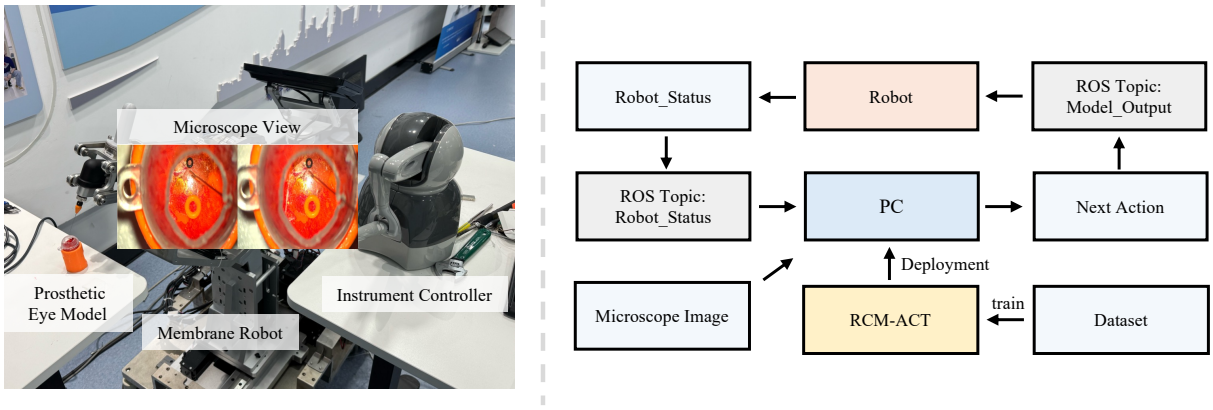


Fig. 2: Diagram of the Deployment Process for the Eye Surgery Robot

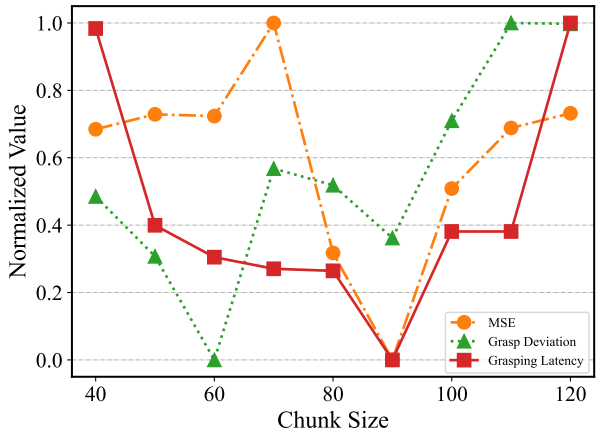


Fig. 3: Impact of the Chunk Size

achieving near-real-time responsiveness at intermediate configurations. Notably, while smaller chunk sizes (<90) yield the lowest normalized grasping deviation—critical for precise ring positioning in our ring task—their latency remains suboptimal for dynamic robotic operations. Our chosen configuration (chunk size 90) demonstrates an optimal balance, achieving the highest normalized prediction accuracy and near-minimal latency while maintaining grasping deviation within task-specific precision thresholds. Larger chunk sizes (>90) exhibit performance degradation across all metrics, with latency increasing by 15-30% and prediction accuracy decreasing by 12-18% compared to size 90, indicating overfitting to extended temporal contexts. This pattern confirms that while moderately large chunk sizes enhance motion continuity modeling, excessive values introduce redundant temporal dependencies that hinder real-time adaptation.

IV. CONCLUSIONS

In this work, we presented a framework for autonomous intraocular foreign body removal manipulation using imitation learning with an eye surgical robot. Our approach achieves submillimeter precision in instrument-ring interactions. This is achieved by leveraging RCM dynamic calibration and motion scaling-invariant state representations, enabling robust visuomotor policy training solely on surgical-grade kinematics and stereo vision data. However, there are still some limitations. The policy operates reactively based on real-time observations and lacks the ability to adapt behavior based on human input, which is essential for collaborative surgical scenarios. Future work will focus on addressing these challenges by integrating miniaturized endoscopic cameras with optimized mounting systems to enhance clinical relevance. Furthermore, to address the critical aspect of operational safety as we move towards clinical application, future work will focus on developing a robust safety framework. This includes implementing ”virtual fixtures” to create software-defined no-go zones around delicate anatomical structures like the retina, preventing any unintended contact regardless of the policy’s output.

ACKNOWLEDGMENT

This research was partially supported by the Fundamental Research Funds for the Central Universities under Grant No. 226-2024-00227, the National Natural Science Foundation of China under Grant No. 62476246, and the Opening Foundation of the State Key Laboratory of Transvascular Implantation Devices under Grant No. SKLTID2024003.

REFERENCES

[1] W. J. Wirostko, S. Bhatia, W. F. Mieler, and C. M. McCabe, "Removal of intraocular foreign bodies," in *Vitreoretinal Surgical Techniques, Second Edition*. Routledge, 2019, pp. 492–504.

[2] A. Al-Ani, M. S. Bondok, K. Madjedi, S. Kherani, and A. Kherani, "Clinical outcomes and characterization of intraocular foreign body injuries from a canadian centre: a 20-year retrospective study and literature review," *Canadian journal of ophthalmology*, 2024.

[3] X. Liu, Q. Bai, and X. Song, "Clinical and imaging characteristics, outcomes and prognostic factors of intraocular foreign bodies extracted by vitrectomy," *Scientific Reports*, vol. 13, no. 1, p. 14136, 2023.

[4] H. C. Jung, S. Y. Lee, C. K. Yoon, U. C. Park, J. W. Heo, and E. K. Lee, "Intraocular foreign body: diagnostic protocols and treatment strategies in ocular trauma patients," *Journal of clinical medicine*, vol. 10, no. 9, p. 1861, 2021.

[5] G. Singh, W. W. J. Jie, M. T. Sun, R. Casson, D. Selva, and W. Chan, "Overcoming the impact of physiologic tremors in ophthalmology," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 260, no. 12, pp. 3723–3736, 2022.

[6] R. Retnaningsih, A. Budiyono, R. Ismail, D. Tugasworo, R. Danuaji, S. Syahrul, and H. Gunawan, "Robotics applications in neurology: a review of recent advancements and future directions," *arXiv preprint arXiv:2312.06956*, 2023.

[7] A. J. Thirunavukarasu, M. L. Hu, W. P. Foster, K. Xue, J. Cehajic-Kapetanovic, and R. E. MacLaren, "Robot-assisted eye surgery: a systematic review of effectiveness, safety, and practicality in clinical settings," *Translational Vision Science & Technology*, vol. 13, no. 6, pp. 20–20, 2024.

[8] N. Ingelhag, J. Munkeby, M. C. Welle, M. Moletta, and D. Kragic, "Real-time operator takeover for visuomotor diffusion policy training," *arXiv preprint arXiv:2502.02308*, 2025.

[9] Z. Wang and A. H. Qureshi, "Deri-igp: Learning to manipulate rigid objects using deformable linear objects via iterative grasp-pull," *IEEE Robotics and Automation Letters*, 2025.

[10] Y. Zhang, D. Chen, W. He, A. E. P. Barceló, J. V. S. Luces, and Y. Hirata, "Multi-critic reinforcement learning for garment handling: Addressing unpredictability in temporal-phase continuous contact tasks," *IEEE Transactions on Automation Science and Engineering*, 2025.

[11] K. Imai, Y. Takahashi, S. Tsuichihara, and M. Haruna, "Autonomous teleoperated robotic arm based on imitation learning using instance segmentation and haptics information," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 29, no. 1, pp. 79–94, 2025.

[12] K. Kawaharazuka, K. Okada, and M. Inaba, "Robotic constrained imitation learning for the peg transfer task in fundamentals of laparoscopic surgery," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 606–612.

[13] X. Jian, Y. Song, D. Liu, Y. Wang, X. Guo, B. Wu, and N. Zhang, "Motion planning and control of active robot in orthopedic surgery by cdmp-based imitation learning and constrained optimization," *IEEE Transactions on Automation Science and Engineering*, 2025.

[14] M. M. Marinho, Y. Koyama, Y. Taniguchi, T. Yamanaka, F. Arai, S. Omata, K. Sugimoto, T. Ueta, K. Totsuka, T. Shiraya, *et al.*, "Evaluation of (shared) autonomy in robot-assisted vitreoretinal surgery using a surgical model," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 21, no. 1, p. e70040, 2025.

[15] D. Arikan, P. Zhang, M. Sommersperger, S. Dehghani, M. Esfandiari, R. H. Taylor, M. A. Nasseri, P. Gehlbach, N. Navab, and I. Iordachita, "Towards motion compensation in autonomous robotic subretinal injections," *arXiv preprint arXiv:2411.18521*, 2024.

[16] J. W. Kim, P. Zhang, P. Gehlbach, I. Iordachita, and M. Kobilarov, "Towards autonomous eye surgery by combining deep imitation learning with optimal control," in *Conference on Robot Learning*. PMLR, 2021, pp. 2347–2358.

[17] M. Briel, L. Haide, M. Hess, J. Schimmelpfennig, P. Matten, R. Peter, M. Hillenbrand, E. Tagliabue, and F. Mathis-Ullrich, "Intraoperative adaptive eye model based on instrument-integrated oct for robot-assisted vitreoretinal surgery," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2025.

[18] E. Rahimy, J. Wilson, T. Tsao, S. Schwartz, and J. Hubschman, "Robot-assisted intraocular surgery: development of the iriss and feasibility studies in an animal model," *Eye*, vol. 27, no. 8, pp. 972–978, 2013.

[19] A. Ebrahimi, S. Sefati, P. Gehlbach, R. H. Taylor, and I. I. Iordachita, "Simultaneous online registration-independent stiffness identification and tip localization of surgical instruments in robot-assisted eye surgery," *IEEE Transactions on Robotics*, vol. 39, no. 2, pp. 1373–1387, 2022.

[20] B. Keller, M. Draelos, K. Zhou, R. Qian, A. N. Kuo, G. Konidaris, K. Hauser, and J. A. Izatt, "Optical coherence tomography-guided robotic ophthalmic microsurgery via reinforcement learning from demonstration," *IEEE Transactions on Robotics*, vol. 36, no. 4, pp. 1207–1218, 2020.

[21] F. Turgut, G. M. Somfai, F. M. Heussen, A. Eberle, M. D. de Smet, and M. D. Becker, "Robot-assisted epiretinal membrane peeling: a prospective assessment of pre-and intra-operative times and of surgeons' subjective perceptions," *Journal of clinical medicine*, vol. 12, no. 8, p. 2768, 2023.

[22] M. Zhou, Q. Yu, K. Huang, S. Mahov, A. Eslami, M. Maier, C. P. Lohmann, N. Navab, D. Zapp, A. Knoll, *et al.*, "Towards robotic-assisted subretinal injection: A hybrid parallel–serial robot system design and preliminary evaluation," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 8, pp. 6617–6628, 2019.

[23] M. Zhou, X. Guo, M. Grimm, E. Lochner, Z. Jiang, A. Eslami, J. Ye, N. Navab, A. Knoll, and M. A. Nasseri, "Needle detection and localisation for robot-assisted subretinal injection using deep learning," *CAAI Transactions on Intelligence Technology*, 2023.

[24] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] S. Yildirim, "Design of adaptive robot control system using recurrent neural network," *Journal of Intelligent and Robotic Systems*, vol. 44, pp. 247–261, 2005.

[26] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Transformer-based deep imitation learning for dual-arm robot manipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8965–8972.

[27] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE robotics and automation letters*, vol. 7, no. 3, pp. 8170–8177, 2022.

[28] P. Katara, Z. Xian, and K. Fragkiadaki, "Gen2sim: Scaling up robot learning in simulation with generative models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6672–6679.

[29] B. Pang, T. Zhao, X. Xie, and Y. N. Wu, "Trajectory prediction with latent belief energy-based model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 814–11 824.

[30] Y. Wang, Y. Zhang, M. Huo, R. Tian, X. Zhang, Y. Xie, C. Xu, P. Ji, W. Zhan, M. Ding, *et al.*, "Sparse diffusion policy: A sparse, reusable, and flexible policy for robot learning," *arXiv preprint arXiv:2407.01531*, 2024.

[31] Z. Chen, Z. Mandi, H. Bharadhwaj, M. Sharma, S. Song, A. Gupta, and V. Kumar, "Semantically controllable augmentations for generalizable robot learning," *The International Journal of Robotics Research*, p. 02783649241273686, 2024.

[32] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[33] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (srt): Imitation learning for surgical tasks," *arXiv preprint arXiv:2407.12998*, 2024.

[34] J. Lin and M. T. Chua, "A low cost surrogate eye model for corneal foreign body removal," *BMC ophthalmology*, vol. 20, pp. 1–7, 2020.