2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Sample-Efficient Learning with Online Expert Correction for Autonomous Catheter Steering in Endovascular Bifurcation Navigation

Hao Wang[1], Tianliang Yao[2], Bo Lu[3], Zhiqiang Pei[4], Liu Dong[5], Lei Ma[1], Peng Qi[1,6,*]

***Abstract*— Robot-assisted endovascular intervention offers a safe and effective solution for remote catheter manipulation, reducing radiation exposure while enabling precise navigation. Reinforcement learning (RL) has recently emerged as a promising approach for autonomous catheter steering; however, conventional methods suffer from sparse reward design and reliance on static vascular models, limiting their sample efficiency and generalization to intraoperative variations. To overcome these challenges, this paper introduces a sample-efficient RL framework with online expert correction for autonomous catheter steering in endovascular bifurcation navigation. The proposed framework integrates three key components: (1) A segmentation-based pose estimation module for accurate real-time state feedback, (2) A fuzzy controller for bifurcation-aware orientation adjustment, and (3) A structured reward generator incorporating expert priors to guide policy learning. By leveraging online expert correction, the framework reduces exploration inefficiency and enhances policy robustness in complex vascular structures. Experimental validation on a robotic platform using a transparent vascular phantom demonstrates that the proposed approach achieves convergence in 123 training episodes—a 25.9% reduction compared to the baseline Soft Actor-Critic (SAC) algorithm—while reducing average positional error to 83.8% of the baseline. These results indicate that combining sample-efficient RL with online expert correction enables reliable and accurate catheter steering, particularly in anatomically challenging bifurcation scenarios critical for endovascular navigation.**

I. Introduction

Endovascular interventions are a minimally invasive approach to the diagnosis and treatment of complex cardiovascular diseases, offering shorter recovery times and fewer postoperative complications than open surgery [1]. Effective catheter navigation underlies procedural safety and efficiency, and bifurcation navigation demands accurate branch selection that depends on precise catheter steering in real-time [2]. Robot-assisted systems offer motion scaling and tremor suppression that enable consistent steering and standardized bifurcation navigation, with concurrent reductions in radiation exposure and operator fatigue [3]–[5]. An overview of the expert-augmented navigation paradigm is illustrated in Fig. 1.

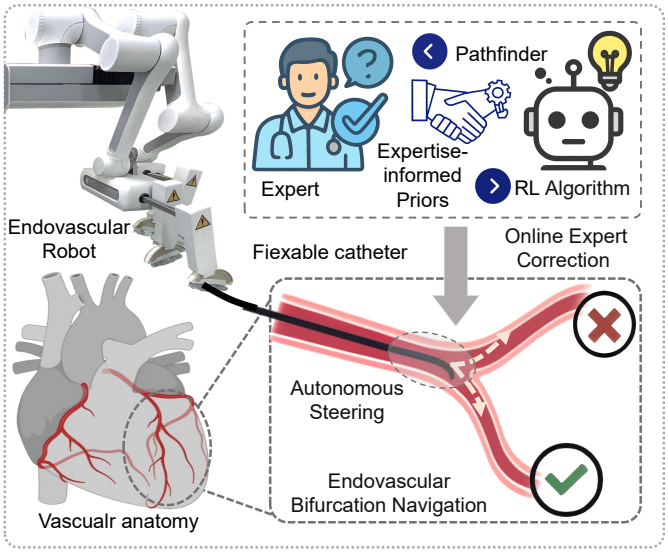

Fig. 1. Illustration of endovascular navigation enhanced by expert knowledge and behavior modeling. The framework utilizes intraoperative imaging to detect the real-time position and orientation of the catheter tip. By combining expert procedural patterns with reinforcement learning strategies, the system dynamically adjusts navigation to ensure safe and efficient traversal through vascular bifurcations. This hybrid approach improves accuracy and reliability in reaching target anatomical sites during complex interventions. The schematic was created using BioRender (`https://biorender.com`).

Advancing toward task-level autonomy in this context holds substantial clinical value by promoting consistent navigation across patient-specific anatomies, timely responses to intraoperative changes, and reduced cognitive burden on clinicians [6]–[8]. Reinforcement learning (RL) has emerged as a promising foundation by enabling agents to learn navigation strategies through environmental interaction [9], [10], with growing evidence of transferability to physical systems [11]. Despite this momentum, current approaches struggle with low sample efficiency, dependence on hand-crafted rewards, and limited adaptability to evolving intraoperative conditions [12]. For example, DDPG in CathSim improved over PPO but relied on uninformed random exploration and remained simulation-only [13], [14]. Inverse RL from offline demonstrations can infer rewards yet may miss real-

This work is supported by the National Key Research and Development Program of China under Grant No. 2023YFB4705200, the National Natural Science Foundation of China under Grant No. 62273257, and the Open Project Fund of State Key Laboratory of Cardiovascular Diseases No.2024SKL-TJ002. *(*Corresponding Author: Peng Qi, email: pqi@tongji.edu.cn).*

[1]Department of Control Science and Engineering, College of Electronics and Information Engineering, and Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092, China;

[2]Department of Electronic Engineering, Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong SAR 999077, China;

[3]Robotics and Microsystems Center, School of Mechanical and Electric Engineering, Soochow University, Suzhou, Jiangsu 215131, China;

[4]School of Oriental Pan-Vascular Devices Innovation College, University of Shanghai for Science and Technology, Shanghai 200093, China;

[5]Shanghai Operation Robot Co., Ltd., Shanghai 201318, China;

[6]State Key Laboratory of Cardiovascular Diseases and Medical Innovation Center, Shanghai East Hospital, School of Medicine, Tongji University 200092, Shanghai, China.

time expert strategies and clinical constraints [15]. Planning-driven systems that pair DQN with fuzzy control assist drilling but depend on static preoperative models, limiting responsiveness to imaging dynamics and precluding online policy refinement [16]. Critically, autonomous catheter steering for bifurcation navigation remains underexplored, with few methods explicitly targeting real-time steering decisions that integrate live imaging, online expert feedback, and robust control on physical platforms [17]–[19]. These gaps motivate sample-efficient learning with online expert correction for autonomous catheter steering in endovascular navigation.

In this context, catheter steering for bifurcation navigation is an online perception–control problem [20]. The system must detect approaching bifurcations from live imaging, estimate centerlines and relative tip-to-branch pose, and select the appropriate daughter vessel. It then has to orient and advance the catheter and guidewire through the junction while regulating torque, axial push, and rotation to maintain luminal traversal and avoid excessive wall [21], [22]. This task is challenging due to 2D imaging ambiguity and foreshortening, deformable tool–vessel interactions with friction and backlash, and patient-specific anatomical variability [15]. Sparse terminal rewards at branch entry and delayed credit assignment further reduce sample efficiency, while fixed heuristic rewards limit adaptability to intraoperative changes [23]. These factors motivate learning strategies that inject expert priors, enable online correction to keep exploration safe and informative, and incorporate execution-time compensation for modeling errors during branch selection.

This paper presents a sample-efficient learning framework for autonomous catheter steering at vascular bifurcations with online expert correction. The SAC-EIL-GAIL framework integrates Soft Actor-Critic (SAC), Generative Adversarial Imitation Learning (GAIL), and expert-in-the-loop supervision to accelerate learning, improve policy fidelity, and preserve adaptability under live imaging. Catheter kinematics are modeled with a constant-curvature formulation and aligned with image-derived centerlines using sub-pixel skeleton fitting. Expert maneuvers are converted into control commands that inform policy updates and define target poses at bifurcations. A fuzzy logic controller compensates for image-driven modeling errors and stabilizes branch selection.

The primary contributions of this study are as follows:

- A unified SAC-EIL-GAIL framework that combines maximum-entropy RL, adversarial imitation, and online expert correction. The design improves sample efficiency and training stability, enhances policy plausibility through expert guidance, and reduces reliance on hand-crafted rewards.
- A constant-curvature kinematic model is coupled with skeleton-based centerline extraction for sub-pixel trajectory fitting. Expert demonstrations are mapped to control inputs, enabling precise steering and reliable localization at challenging bifurcations.
- Target poses derived from expert-in-the-loop interaction support bifurcation navigation. A fuzzy logic controller compensates for image-derived modeling errors in real time, improving navigation success rates and robustness in complex vascular environments.

II. Methodology

A. System Overview

The experimental system consists of a robotic catheterization platform and an autonomous navigation framework. As shown in Fig. 2(a), two paired robotic arms collaborate in operation, where one executes axial insertion together with rotational control of the catheter, and the other provides stabilization to suppress unintended motion and ensure operational safety. Each arm integrates rotary joints and linear stages with grippers that securely manipulate the catheter.

The software architecture integrates reinforcement learning, imitation learning, and fuzzy control under expert supervision. A YOLO-based vision module continuously detects the catheter tip and vascular bifurcations from intraoperative images. When the tip enters a bifurcation, the intervention module retrieves an appropriate trajectory and posture from a predefined library and refines the motion with fuzzy control. In the remaining segments, actions are autonomously generated by the reinforcement learning agent. Upon reaching the terminal region, the task is rewarded and the catheter is reset to the initial position for the next trial.

B. Problem Formulation

The task of autonomous catheter navigation is formulated as a constrained sequential decision-making problem under the reinforcement learning paradigm. The environment is defined by a state space $\mathcal{S}$, an action space $\mathcal{A}$, and the transition distribution $p(s_{t+1}|s_t, a_t)$. At each time step t, the catheter system is described by a state $s_t \in \mathcal{S}$, which includes information derived from image segmentation such as the skeletonized centerline, the distal tip position, and its orientation. The agent selects an action $a_t \in \mathcal{A}$, corresponding to base manipulations such as translational advancement or axial rotation, and receives a reward R_t that reflects navigation safety and efficiency. The optimal navigation policy is obtained by maximizing the expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T} R_t \right]. \tag{1}$$

Physical constraints are imposed based on catheter mechanics, which are represented using a constant curvature approximation with a fixed small-angle distal bend. Under this model, curvature along the catheter centerline $\gamma(s)$ is characterized by

$$\kappa(s) = \frac{|\gamma'(s) \times \gamma''(s)|}{|\gamma'(s)|^3}, \quad s \in [0, L], \tag{2}$$

ensuring geometric stability during navigation while approximating the inherent bending property of the distal segment. During navigation, when the catheter tip reaches a vascular bifurcation, an expert provides corrective guidance by designating a target pose $(D_{\text{target}}, P_{\text{target}})$. The navigation

system then minimizes both translational and rotational errors, expressed as $e_{\text{trans}} = \|P_{\text{target}} - P_{\text{current}}\|_2$ and $e_{\text{rot}} = D_{\text{target}} - D_{\text{current}}$, respectively. The learning objective integrates environment-driven reward signals with expert priors through a hybrid reinforcement learning strategy. The overall reward is formulated as

$$R_t = w_{\text{SAC}}(t)\, r_{\text{SAC}} + w_{\text{GAIL}}(t)\, r_{\text{GAIL}} + \epsilon_t, \tag{3}$$

where $\epsilon_t \sim \mathcal{U}(-\delta, \delta)$ introduces stochasticity, and the time-dependent weights $w_{\text{SAC}}(t)$ and $w_{\text{GAIL}}(t)$ provide a smooth transition from exploration dominated by reinforcement signals to demonstration-driven optimization.

C. Catheter Modeling with Online Expert Correction Pose Mapping for Robotic Control

The catheter navigation process is modeled using an Expert-in-the-Loop (EIL) strategy, where the robotic system continually adjusts its posture by evaluating discrepancies between the current catheter configuration and expert-selected correction poses. These corrections are translated into fuzzy control rules that guide the underlying actuation signals.

1) Skeleton Extraction: The catheter skeleton is derived from the segmented image by applying a thinning algorithm, where pixels are represented as foreground with value 1 and background with value 0 [24]. Each candidate foreground pixel p_1 and its eight-neighborhood $\{p_2, \ldots, p_9\}$ are retained as skeletal points only if a set of logical conditions is satisfied: the pixel itself must belong to the foreground, the total number of neighboring pixels must lie between two and six, the connectivity number $S(p_1)$ must be equal to one, and simultaneous occupancy of certain triplets of neighbors must be avoided to prevent spurious branches. The connectivity number is formally defined as

$$S(p_1) = \sum_{i=2}^{9} s(p_i, p_{i+1}), \tag{4}$$

where $s(p_i, p_{i+1})$ encodes a 0-to-1 transition along the neighborhood in a clockwise order.

Once a coherent skeletonized structure is obtained, the trajectory extraction is guided by geometric significance. The Euclidean distance transform is employed to identify prominent endpoints, where the pixel associated with the largest distance value is labeled as Q_1 and the farthest endpoint relative to Q_1 is designated as Q_2. The longest path connecting these two points is computed using a breadth-first search procedure, ensuring maximal coverage of the skeletal structure. To improve numerical stability and mitigate discretization artifacts, the extracted path is further smoothed with a Savitzky–Golay filter, yielding a continuous sub-pixel centerline representation suitable for subsequent modeling and control, as illustrated in Fig. 2(b).

2) Control Input Mapping: The rotation angle to reproduce the expert pose is inferred from the 2D segmented skeleton. A catheter with a fixed distal bend is modeled as a polyline rotating about its central axis. By rotating the base, the manipulator controls the tip orientation and overall pose.

Algorithm 1: Online Expert-Corrected Reinforcement Learning with Fuzzy Control

Input: Expert-selected pose image I_{expert}, initial state-action pair $(s_{\text{init}}, a_{\text{init}})$, expert dataset $(s_{\text{expert}}, a_{\text{expert}}, r_{\text{expert}})$
Output: Optimal policy π^*

if $episodes < 50$ **then**
  $\pi \leftarrow$ SoftActorCriticExplore$(s_{\text{SAC}}, a_{\text{SAC}})$; record reward r_{SAC};
end
if $episodes \geq 50$ **then**
  $R_t \leftarrow w_{\text{SAC}}(t)\, r_{\text{SAC}} + w_{\text{GAIL}}(t)\, r_{\text{GAIL}} + \epsilon_t$;
  $\pi^* \leftarrow$ TrainPolicy(s_t, a_t, R_t);
  if *bifurcation detected* **then**
    $(D_{\text{target}}, P_{\text{target}}) \leftarrow$ ExtractPose(I_{expert});
    $(D_{\text{current}}, P_{\text{current}}) \leftarrow$ ExtractPose(I_{current});
    while $(|e_{rot}| > \epsilon_{rot}) \vee (|e_{trans}| > \epsilon_{trans})$ **do**
      $e_{\text{rot}} \leftarrow D_{\text{target}} - D_{\text{current}}$;
      $e_{\text{trans}} \leftarrow \|P_{\text{target}} - P_{\text{current}}\|_2$;
      $a_{\text{rule}} \leftarrow$ FuzzyControl$(e_{\text{rot}}, e_{\text{trans}})$;
      Update$(D_{\text{current}}, P_{\text{current}})$;
    end
  end
end

As shown in Fig. 2(b), the catheter tip endpoint $P(x, y, z)$ is rotated about the x-axis by an angle θ, resulting in a new position $P'(x', y', z')$. The projection of P' onto the x–y plane gives a vertical distance $d' = |y\cos\theta - z\sin\theta|$ from the x-axis. When $\theta_{\text{pitch}} \in \left(0, \frac{\pi}{2}\right)$, d' decreases monotonically, and when $\theta_{\text{pitch}} \in \left(\frac{\pi}{2}, \pi\right)$, it increases. This distance d' represents the distance from P' to the fitted line of the proximal shaft. The pitch angle θ_{pitch} is mapped to the actuator input u_r via: $u_r = \mathcal{F}(\theta_{\text{pitch}}) = \mathcal{F}(G(D))$.

In order to distinguish between the possible positive and negative values of y' after rotation in the $(0, \pi)$ range, we select two points $A(0, b)$ and $B(1, k + b)$ on the line $kx - y + b = 0$, as shown in the following formula:

$$D = \begin{cases} d' & \text{if } \overrightarrow{AB} \times \overrightarrow{AP} > 0, \\ -d' & \text{if } \overrightarrow{AB} \times \overrightarrow{AP} < 0. \end{cases} \tag{5}$$

Upon the completion of the catheter modeling, a fuzzy control approach is employed to map the robot's control inputs to the actions selected by the expert. The control algorithm computes the translation and rotation errors from image modeling, adjusting the position and orientation to reach the target. The translation error $e_{\text{trans}} = \|(P_{\text{target}}, P_{\text{current}})\|_2$ is the Euclidean distance between the target point and the current catheter tip , and the rotation error is $e_{\text{rot}} = D_{\text{target}} - D_{\text{current}}$.

3) Fuzzy Control based on Expert Correction: The input error e is mapped to predefined fuzzy categories (NL, NS, Z, PS, PL) with centers m. A triangular membership function, parameterized by a constant half-width w, calculates the degree of membership for each fuzzy set, effectively transforming the crisp input into linguistic variables with

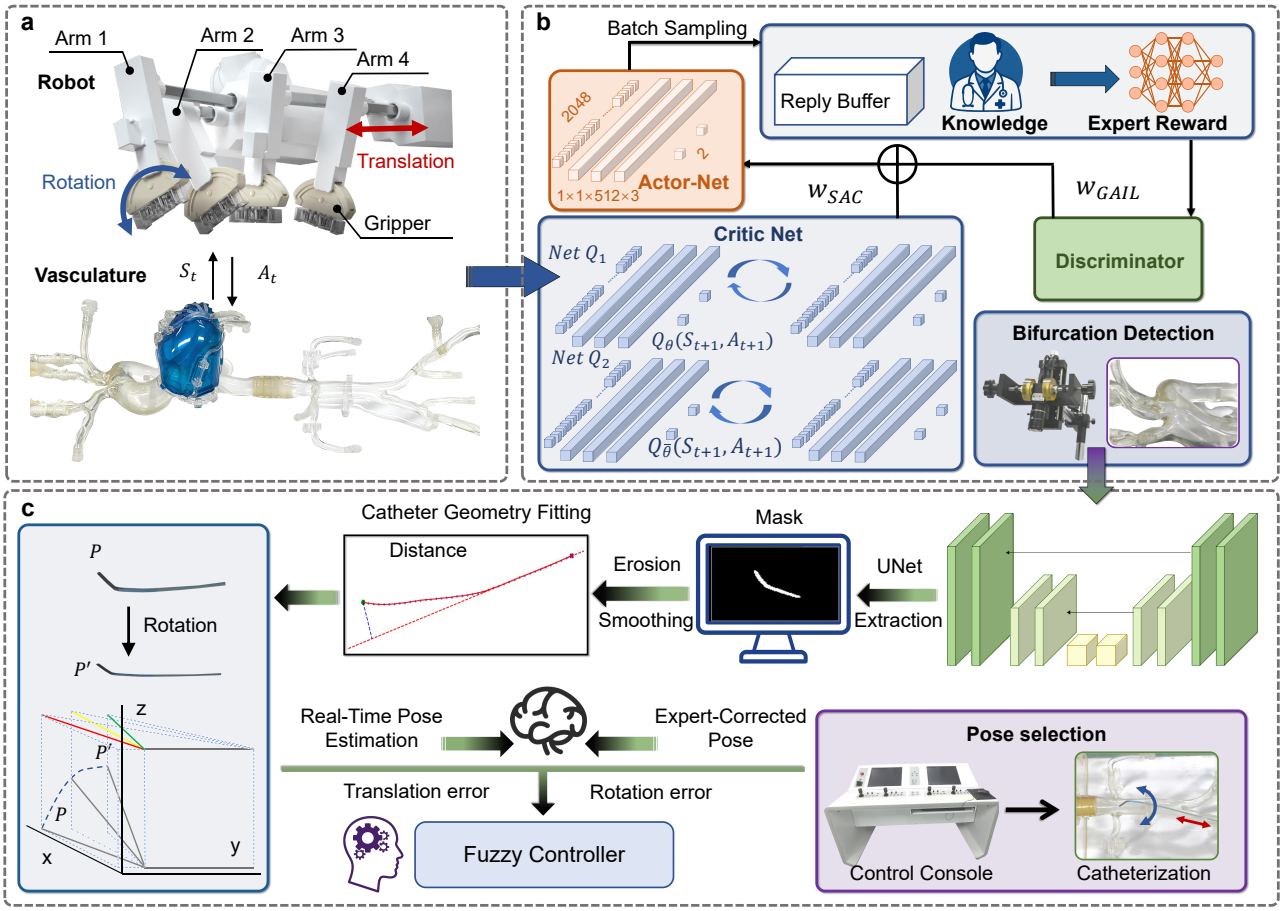

Fig. 2. Expert-in-the-loop catheter navigation framework integrating reinforcement learning and fuzzy control. (a) Robotic catheterization setup and agent–environment interaction: the agent observes the vascular state s_t and outputs action a_t to command catheter translation, rotation, and gripping. (b) Policy learning that combines Soft Actor–Critic (SAC) with Generative Adversarial Imitation Learning (GAIL): mini-batches from a replay buffer train an actor–critic with twin critics (Q_1, Q_2). A discriminator supplies an expert reward shaped by prior knowledge, and the policy is optimized using a weighted sum of SAC and GAIL rewards (w_{SAC}, w_{GAIL}). Bifurcation detection triggers the expert-in-the-loop module. (c) Online fuzzy pose correction at bifurcations: an expert selects a target pose; a U-Net-based segmentation extracts the catheter mask, followed by smoothing/erosion and geometry fitting for real-time pose estimation. Translation and rotation errors feed a fuzzy controller that adjusts robot commands to reach the expert-corrected pose.

associated membership values.

$$\mu(e) = \begin{cases} 0, & e \le m - w \text{ or } e \ge m + w \\ \frac{e-(m-w)}{w}, & m - w < e \le m \\ \frac{(m+w)-e}{w}, & m < e < m + w \end{cases} \tag{6}$$

Fuzzy inference is performed using a rule base defined over all combinations of the fuzzified inputs. Each rule's activation strength is calculated using the minimum operator (representing fuzzy AND):

$$\mu_{\text{rule}} = \min\left(\mu_t(e_{\text{trans}}), \mu_r(e_{\text{rot}})\right). \tag{7}$$

The control outputs are aggregated using the maximum operator (fuzzy OR) across all activated rules:

$$\mu_C(u) = \max\left(\mu_C(u), \mu_{\text{rule}}\right). \tag{8}$$

The final crisp control outputs are obtained using the centroid method:

$$u^* = \frac{\sum_i \mu_i \cdot m_i}{\sum_i \mu_i} \tag{9}$$

where μ_i and $m_i$denote the membership degree and center of the i-th output fuzzy set.

D. Expert Experience Guided Trajectory Optimization

We propose a hybrid reinforcement learning framework that integrates SAC with Generative Adversarial Imitation Learning (GAIL). The method gradually shifts the reward contribution from environment-driven signals to expert imitation, enabling early exploration and later incorporation of expert priors for balanced learning.

1) Exploration-Oriented Reinforcement Learning with Soft Actor–Critic Algorithm: SAC is an off-policy deep reinforcement learning method based on maximum entropy policy optimization [25]. It aims to maximize the expected cumulative reward while encouraging higher entropy in the policy to promote exploration. Its objective function is formulated as:

$$J_\pi = \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\left[r(s_t, a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\right] \tag{10}$$

where $r(s_t, a_t)$ denotes the instantaneous environmental reward, $\mathcal{H}(\pi(\cdot|s_t))$ is the policy entropy, and α is the temperature coefficient.

2) Expert-Guided Policy Shaping via Generative Adversarial Imitation Learning: GAIL aims to minimize the divergence between the agent's trajectory distribution and that of an expert via adversarial training [26]. A discriminator D_ψ is trained to differentiate between expert trajectories and agent-generated ones, while the policy network π_θ attempts to fool the discriminator:

$$\min_{\pi_\theta} \max_{D_\psi} \mathbb{E}_{\pi_\theta} \left[\log\left(1 - D_\psi(s,a)\right)\right] + \mathbb{E}_{\pi_E} \left[\log D_\psi(s,a)\right] \tag{11}$$

The imitation reward from the discriminator is given by:

$$r_{\mathrm{GAIL}}(s,a) = -\log\left(1 - D_\psi(s,a)\right) \tag{12}$$

This transforms expert prior knowledge into a reward signal that guides the agent toward expert-like behavior.

3) Hybrid Reward Scheduling for Balancing Exploration and Imitation: To achieve a smooth transition from reinforcement learning to imitation learning, we propose a sigmoid-based dynamic reward scheduling mechanism. During training, rewards from the SAC and GAIL branches are combined with time-dependent weights, enabling a gradual shift from exploration to demonstration-guided behavior.Let T be the total number of training episodes and t the current episode. The weighting factor is defined as:

$$\alpha(t) = \frac{1}{1 + \exp\left(-k \cdot \left(t - \frac{T}{2}\right)\right)} \tag{13}$$

$$w_{\mathrm{SAC}}(t) = 1 - 0.5 \cdot \alpha(t), \quad w_{\mathrm{GAIL}}(t) = 0.5 \cdot \alpha(t) \tag{14}$$

where k is a temperature parameter that controls the steepness of the transition. As training progresses, $w_{\mathrm{SAC}}(t)$ gradually decreases from 1 to 0.5 while $w_{\mathrm{GAIL}}(t)$ increases from 0 to 0.5, ensuring a smooth shift from exploration to exploitation.

The total reward at episode t is then defined as:

$$R_t = w_{\mathrm{SAC}}(t) \cdot r_{\mathrm{SAC}} + w_{\mathrm{GAIL}}(t) \cdot r_{\mathrm{GAIL}} + \epsilon_t \tag{15}$$

where $\epsilon_t \sim \mathcal{U}(-\delta, \delta)$ is a uniform random perturbation that encourages stochasticity and robustness during learning.

By replacing the original reward in SAC's Bellman objective with R_t, the optimization becomes:

$$J_\pi = \sum_{t=0}^{T} \mathbb{E}\left[R_t + \alpha \mathcal{H}(\pi(\cdot|s_t))\right]. \tag{16}$$

III. Experiments and Results

A. Implementation Details

Experiments were conducted on a 3D silicone renal artery phantom filled with saline, which provided a controlled vascular environment. Catheter manipulation was performed using a pair of custom-built 2-DoF robotic arms: one executed translational and rotational motions to mimic physician operation, while the other provided system stabilization. The entire setup was monitored by a Hikvision MV-CS050-10GM industrial camera at 10 frames per second and computations were accelerated on an NVIDIA RTX 3090 GPU.

Catheter tip detection relied on a YOLOv5 model trained on 7,293 brightness-augmented images [27]. After each action, the centroid position $P(x, y)$ was extracted, with detection failures ($x < 0$) prompting a short pause and reacquisition. When consecutive failures occurred, the last valid frame was reused to maintain temporal consistency. Catheter segmentation was performed using a U-Net model trained on 1,000 augmented images derived from 400 originals. The tip was identified by fitting a line to the first 30% (empirical) from the endpoint farther from the bifurcation, and taking the opposite end as the tip.

Fuzzy control employed five linguistic sets {NL, NS, Z, PS, PL} for both translation and rotation. Translation membership centers were {-2.0, -1.0, 0, 1.0, 2.0} cm with half-width $\Delta t = 1.0$ cm, derived from the calibrated mapping $D \in [0, 215]$ pixel $\rightarrow d \in [0, 2.5]$ cm. Rotation membership centers were $-60°, -30°, 0°, 30°, 60°$ with half-width $\Delta r = 30°$, based on the mapping $E \in [0, 80] \rightarrow e \in [0, 90°]$, where $e = (80 - E) \times 1.125°/\text{pixel}$.

The reward function combined sparse terminal signals with distance-based shaping. Episodes terminated with a penalty of $r_t = -150$ when push distance exceeded 500 mm, step count exceeded 20, or the catheter moved out of bounds ($x_h < 10$ or $x_h > 900$). In all other cases, r_t was shaped by the distance to the target, with additional penalties for rotational deviation and misalignment.

B. Model Performance Evaluation

The navigation task is meticulously designed to evaluate both the accuracy and efficiency of autonomous guidance within a 3D anatomical model. As illustrated in Fig.3(a), navigation initiates at the femoral artery entry point in the lower right and proceeds autonomously toward the target region in the upper left, guided by real-time coordinates of the catheter tip. To assess navigation performance, five reinforcement learning (RL) frameworks were trained and evaluated across 300 episodes summarized in Table I. The baseline algorithm was implemented via PyTorch and Stable-Baselines3 [28].

The proposed method exhibits clear over the baseline algorithms TD3 and SAC across multiple evaluation criteria. With respect to convergence, within 300 episodes, the proposed approach attains stable convergence after only 123 episodes, whereas TD3 and SAC require 175 and 166 episodes, respectively. Since fewer episodes indicate faster convergence, this result substantiates the enhanced convergence efficiency of the proposed method. Moreover, the convergence time is shortened to 59.41 seconds, further demonstrating superior execution performance.

The ablation study demonstrates that integrating expert imitation learning with adversarial learning significantly improves policy quality. The full model achieves a success rate of 59.00%, which is 5.33 percentage points higher than SAC-GAIL and 7.67 points higher than SAC-EIL,

TABLE I

COMPARISON OF AUTONOMOUS NAVIGATION ACROSS RL ALGORITHMS FOR BIFURCATION STEERING

Algorithm	Episodes	Avg. Steps	Success Rate	Avg. Time (s)	Avg. Error (px)
TD3	300	9.21	41.67% (125/300)	78.09	208.76
SAC	300	8.32	44.67% (134/300)	60.39	98.55
SAC-GAIL	300	7.57	53.67% (161/300)	59.04	149.53
SAC-EIL	300	7.66	51.33% (154/300)	61.05	86.90
SAC-EIL-GAIL	300	7.61	59.00% (177/300)	59.41	82.58

Note: An episode is counted as successful if it reaches the target and is flanked by five consecutive successful episodes both before and after. Avg. Time denotes the average duration per episode until convergence. Avg. Error is the mean deviation from the expert pose (translation + rotation) at bifurcations, computed per successful episode.

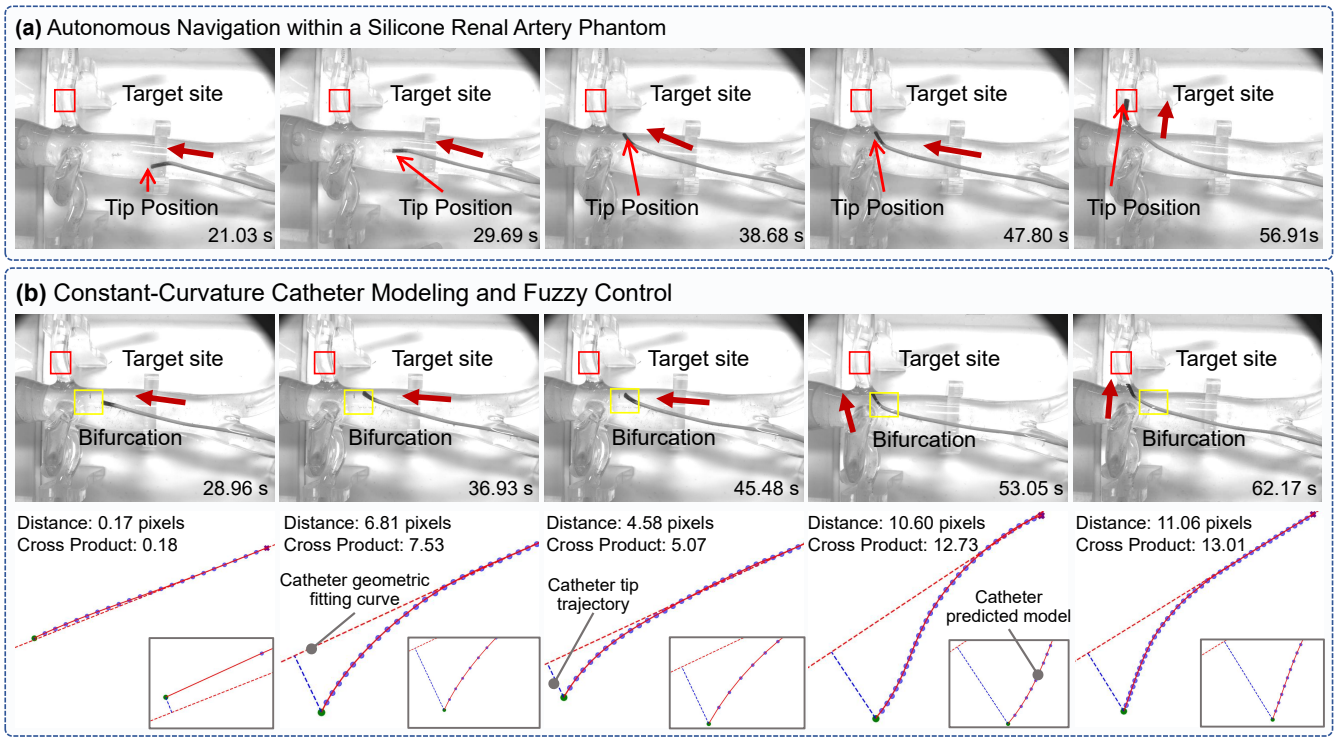

Fig. 3. Autonomous catheter navigation and constant-curvature modeling. (a) After training, the learned policy autonomously steers the catheter to the predefined target region in a silicone renal artery phantom by continuously localizing the tip in fluoroscopic frames. (b) Upon bifurcation detection, the catheter centerline (skeleton) is extracted and fitted with a constant-curvature model. Directional consistency is assessed by the sign/magnitude of the cross product between the fitted tangent and the skeleton tangent; the resulting geometric errors drive an expert-in-the-loop fuzzy controller that updates robot commands to achieve the expert-specified pose. Insets report the tip–target distance (pixels) and cross-product values over time.

confirming its superior task reliability. The updated SAC-EIL-GAIL algorithm has an average step count of 7.61, slightly higher than SAC-GAIL's 7.57 with an increase of approximately 0.53%, but lower than SAC-EIL's 7.66 with a reduction of about 0.65%. The slight increase in steps may result from the adaptation process after incorporating expert data, as the policy requires some extra steps to adjust. This is a common phenomenon in imitation-reinforcement learning integration. Although the average episode duration shows marginal differences, 1.64 seconds faster than SAC-EIL and 0.37 seconds slower than SAC-GAIL, the full model demonstrates an overall improvement in efficiency.

These results validate the proposed method's superiority in both baseline and ablation evaluations, particularly in enhancing policy reliability and efficiency through the combination of expert imitation and adversarial learning. The consistent improvements across multiple metrics suggest that the integration of these two components not only accelerates learning but also fosters more stable and adaptable strategies. By leveraging expert demonstrations to guide early exploration and incorporating adversarial mechanisms to refine long-term decision-making, the method establishes a complementary progression that enhances both learning stability and adaptability.

C. Error-Based Analysis and Confidence Evaluation

The normalized comparison of algorithmic performance across key metrics, as illustrated in Fig.4(a), highlights the

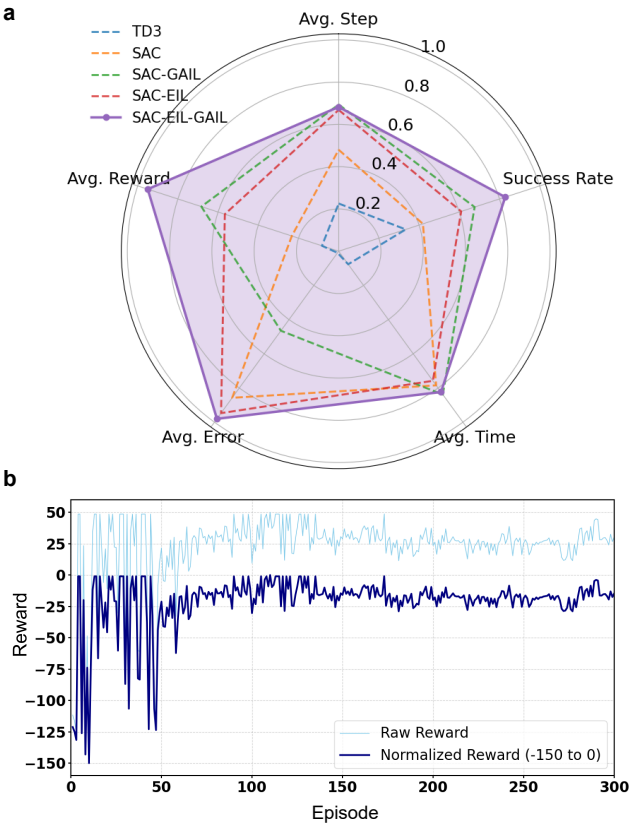

Fig. 4. Comparison of algorithm performance metrics. (a) Normalized Comparison of Key Performance Indicators Across Different RL Algorithms. (b) Average reward per episode during SAC-EIL-GAIL training. As the number of episodes increases, the reward gradually converges. This indicates stable and effective learning.

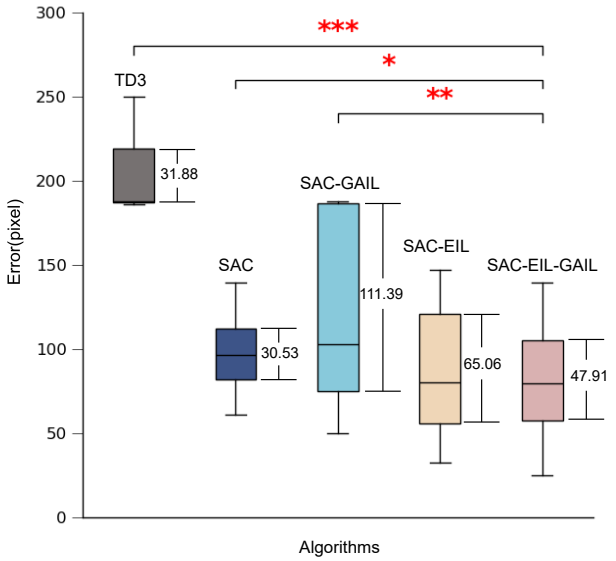

Fig. 5. This figure illustrates the error distributions of TD3, SAC, SAC-GAIL, SAC-EIL, and SAC-EIL-GAIL. The boxplots highlight the central tendency and variability of each method, while the red asterisks indicate statistically significant differences with SAC-EIL-GAIL ($*p < 0.05, **p < 0.01, ***p < 0.001$).

overall advantage of the proposed SAC-EIL-GAIL framework. SAC-EIL-GAIL achieves the lowest average error of 82.58 pixels, significantly smaller than TD3 (208.76), SAC (98.55), SAC-GAIL (149.53), and SAC-EIL (86.90). This reduction in deviation reflects the improved precision of the learned navigation policy, particularly in handling complex bifurcation scenarios.

The error distribution achieved by SAC-EIL-GAIL exhibits a clear advantage over those of TD3, SAC, and SAC-GAIL, with lower variance and tighter concentration around the minimal error range, as shown in Fig. 5. Statistical comparisons, with p-values falling below conventional significance thresholds, support the robustness of this improvement. The boxplot underscores SAC-EIL-GAIL's consistent and precise performance, particularly under more demanding conditions. These results suggest that integrating expert imitation learning with adversarial regularization not only enhances convergence efficiency but also improves trajectory accuracy, yielding a more reliable framework for autonomous navigation.

IV. Discussion

The SAC-EIL-GAIL framework consistently outperforms both baseline and ablation methods, achieving faster convergence, higher success rates, and the lowest navigation error. Statistical analysis confirms that combining online expert correction with adversarial imitation enhances policy quality and yields stable improvements across all key metrics. This hybrid design enables the agent to adapt rapidly while leveraging expert knowledge for high-precision navigation, particularly in complex bifurcation scenarios.

Comparison with ablation settings shows that, although both SAC-EIL-GAIL and SAC-EIL employ online correction, the relative advantage in step efficiency over SAC-GAIL is less pronounced. This arises from the additional adjustments required during the early training phase, when adaptation to expert guidance introduces more refinements. As training progresses, the initial alignment improves and fewer corrections are needed, indicating that efficiency gains become most apparent near convergence, while early-stage benefits lie primarily in stability.

In terms of accuracy, SAC-EIL-GAIL achieves the lowest error and mitigates the limitations of SAC in reward stability, ensuring faster convergence and more reliable performance. Nonetheless, runtime improvements remain modest and the advantage over SAC-EIL is incremental. Occasional fluctuations in error further suggest that robustness under varied conditions is not yet fully achieved. These results highlight that SAC-EIL-GAIL represents a meaningful advance in autonomous catheter navigation but also indicate the necessity for further refinement to enhance robustness, efficiency, and generalization under diverse clinical environments.

V. Conclusion and Future Work

This paper proposed SAC-EIL-GAIL, an imitation learning framework that combines expert demonstrations and prior knowledge with reinforcement learning to enhance robot autonomy in navigating vascular bifurcations. By leveraging

expert data and training on a silicone renal artery phantom, the framework improves pose adjustment when entering branch vessels, shortens training time, and achieves a 17.33% higher success rate compared to the TD3 baseline.

Future directions will focus on multi-instrument collaborative navigation of guidewires and catheters, given their high flexibility and complex nonlinear behavior. Another direction will focus on adaptive expert feedback mechanisms that can operate online, providing corrective inputs during training and execution without compromising safety. Additional emphasis will be placed on enhancing generalization through simulation-to-real transfer and patient-specific vascular modeling, enabling the framework to accommodate diverse anatomical geometries and clinical conditions. In the longer term, integration into the clinical workflow has the potential to support patient-specific rehearsal, intraoperative guidance, and postoperative assessment, thereby contributing to safer, more precise, and more efficient robot-assisted endovascular interventions.

REFERENCES

[1] B. Li, B. E. Warren, N. Eisenberg, D. Beaton, D. S. Lee, B. Aljabri, R. Verma, D. N. Wijeysundera, O. D. Rotstein, C. de Mestral, *et al.*, "Machine learning to predict outcomes of endovascular intervention for patients with pad," *JAMA Network Open*, vol. 7, no. 3, pp. e242350–e242350, 2024.

[2] R. Konda, T. A. Brumfiel, Z. L. Bercu, J. A. Grossberg, and J. P. Desai, "Robotically steerable guidewires—current trends and future directions," *Science Robotics*, vol. 10, no. 105, p. eadt7461, 2025.

[3] T. Yao, B. Lu, M. Kowarschik, Y. Yuan, H. Zhao, S. Ourselin, K. Althoefer, J. Ge, and P. Qi, "Advancing embodied intelligence in robotic-assisted endovascular procedures: A systematic review of ai solutions," *IEEE Reviews in Biomedical Engineering*, vol. 19, pp. 248–266, 2026.

[4] A. Stevenson, A. Kirresh, M. Ahmad, and L. Candilio, "Robotic-assisted pci: the future of coronary intervention?," *Cardiovascular Revascularization Medicine*, vol. 35, pp. 161–168, 2022.

[5] E. T. Fry, J. Kuvin, and J. Sibley, "Maintenance of competence in cardiovascular practice: it's time for more learning, less testing," 2023.

[6] Y. Li, R. Ge, A. Zhu, J. Zhao, D. Shi, Y. Sun, Y. Li, and L. Yang, "A hierarchical framework for real-time path planning of microswarm in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 11, no. 3, pp. 3891–3898, 2026.

[7] A. Pore, Z. Li, D. Dall'Alba, A. Hernansanz, E. De Momi, A. Menciassi, A. C. Gelpi, J. Dankelman, P. Fiorini, and E. Vander Poorten, "Autonomous navigation for robot-assisted intraluminal and endovascular procedures: A systematic review," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2529–2548, 2023.

[8] T. Yao, H. Wang, B. Lu, J. Ge, Z. Pei, M. Kowarschik, L. Sun, L. Seneviratne, and P. Qi, "Sim2real learning with domain randomization for autonomous guidewire navigation in robotic-assisted endovascular procedures," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 13842–13854, 2025.

[9] P. E. Dupont and A. Degirmenci, "The grand challenges of learning medical robot autonomy," *Science Robotics*, vol. 10, no. 104, p. eadz8279, 2025.

[10] Z. Yan, C. Liu, Y. Jiang, W. Zheng, X. Chen, and A. Krieger, "A mobile magnetic manipulation platform for gastrointestinal navigation with deep reinforcement learning control," *arXiv preprint arXiv:2601.15545*, 2026.

[11] T. Yao, M. Ban, B. Lu, Z. Pei, and P. Qi, "Sim4endor: A reinforcement learning centered simulation platform for task automation of endovascular robotics," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 824–830, 2025.

[12] Y. Cho, J.-H. Park, J. Choi, and D. E. Chang, "Sim-to-real transfer of image-based autonomous guidewire navigation trained by deep deterministic policy gradient with behavior cloning for fast learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3468–3475, IEEE, 2022.

[13] W. Tian, J. Guo, S. Guo, and Q. Fu, "A ddpg-based method of autonomous catheter navigation in virtual environment," in *2023 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 889–893, IEEE, 2023.

[14] T. Jianu, B. Huang, M. N. Vu, M. E. Abdelaziz, S. Fichera, C.-Y. Lee, P. Berthet-Rayne, F. R. y Baena, and A. Nguyen, "Cathsim: an open-source simulator for endovascular intervention," *IEEE Transactions on Medical Robotics and Bionics*, vol. 6, no. 3, pp. 971–979, 2024.

[15] T. Yao, Y. Xu, H. Wang, X. Qiu, K. Althoefer, and P. Qi, "Multi-agent fuzzy reinforcement learning with llm for cooperative navigation of endovascular robotics," *IEEE Transactions on Fuzzy Systems*, pp. 1–11, 2025.

[16] G. Ji, Q. Gao, T. Zhang, L. Cao, and Z. Sun, "A heuristically accelerated reinforcement learning-based neurosurgical path planner," *Cyborg and Bionic Systems*, vol. 4, p. 0026, 2023.

[17] A. G. Truesdell, M. A. Alasnag, P. Kaul, S. T. Rab, R. F. Riley, M. N. Young, W. B. Batchelor, A. Maehara, F. G. Welt, A. J. Kirtane, *et al.*, "Intravascular imaging during percutaneous coronary intervention: Jacc state-of-the-art review," *Journal of the American College of Cardiology*, vol. 81, no. 6, pp. 590–605, 2023.

[18] T. Yao, C. Wang, X. Wang, X. Li, Z. Jiang, and P. Qi, "Enhancing percutaneous coronary intervention with heuristic path planning and deep-learning-based vascular segmentation," *Computers in Biology and Medicine*, vol. 166, p. 107540, 2023.

[19] J. Luo, C. Xu, J. Wu, and S. Levine, "Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning," *Science Robotics*, vol. 10, no. 105, p. eads5033, 2025.

[20] T. Yao, Z. Pei, Y. Li, Y. Yuan, and P. Qi, "Real-time guidewire tip tracking using a siamese network for image-guided endovascular procedures," *Advanced Intelligent Systems*, p. 2500425, 2025.

[21] T. Yao, B. Li, B. Lu, Z. Pei, Y. Yuan, and P. Qi, "Real-time 3d guidewire reconstruction from intraoperative dsa images for robot-assisted endovascular interventions," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17344–17351, IEEE, 2025.

[22] M. Lunardi, Y. Louvard, T. Lefèvre, G. Stankovic, F. Burzotta, G. S. Kassab, J. F. Lassen, O. Darremont, S. Garg, B.-K. Koo, *et al.*, "Definitions and standardized endpoints for treatment of coronary bifurcations," *Journal of the American College of Cardiology*, vol. 80, no. 1, pp. 63–88, 2022.

[23] A. Peloso, R. Damiano, X. Zhang, A. Bicchi, E. Votta, and E. De Momi, "Imitation learning for path planning in cardiac percutaneous interventions," *IEEE Transactions on Biomedical Engineering*, 2025.

[24] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.

[25] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning*, pp. 1861–1870, Pmlr, 2018.

[26] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[27] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, J. Poznanski, L. Yu, P. Rai, R. Ferriday, *et al.*, "ultralytics/yolov5: v3. 0," *Zenodo*, 2020.

[28] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.