Current-Inference Transformer Policies for Robust AUV Navigation from World Feedback

Seihun Kim [1] **Joongheon Kim** [2]

Abstract

Reinforcement learning (RL) from feedback is not limited to human preferences in physical systems, i.e., the environment itself provides feedback through progress signals, safety events, and action-outcome mismatches. We study this idea in autonomous underwater vehicle (AUV) navigation, where strong ocean currents create hidden disturbances that are not directly observed by the policy. Accordingly, we propose a current-inference transformer RL framework that infers latent current effects from observation-action histories and uses this inferred world feedback for residual control. A causal transformer encodes recent interaction history, an auxiliary head predicts the current vector using privileged simulator labels during training, and the actor conditions residual actions on the inferred disturbance representation. Our training results show decreasing current-estimation loss and mean absolute error, suggesting that the transformer learns current-related representations from observation-action histories. In an approximately $1.5\times$ intensified-current stress test, the proposed policy achieves 65% success compared with 37% for a guidance-only controller, suggesting improved robustness when current disturbances exceed the nominal regime.

1. Introduction

Autonomous underwater vehicles (AUVs) must operate under nonlinear 6 degree of freedom (6-DOF) dynamics, actuator limits, limited sensing, moving obstacles, and environmental disturbances such as ocean currents (Fossen, 2011; Prestero, 2001). In particular, strong and time-varying currents create mismatches between commanded actions and the resulting motions, and when the current is not directly observed, this can be considered as a partially observable control problem. Classical AUV controllers, such as line-of-sight guidance, sliding-mode control, disturbance-observer-based control, and model predictive control, provide strong structure and interpretability (Fossen et al., 2003; Fossen, 2011). However, their performances depend on accurate models or explicit disturbance estimation. Reinforcement learning (RL) offers an alternative by learning control policies from interaction (Hadi et al., 2022; Schulman et al., 2017), whereas memoryless policies may struggle when hidden disturbances can only be inferred over time.

Recent feedback-based RL methods learn primarily from *human* preferences (Christiano et al., 2017), but in physical control an alternative signal is already implicit in every interaction, i.e., the environment itself reveals action-outcome mismatches whenever the realized motion deviates from the commanded action. We treat this *world feedback* as the basis for residual policy learning under hidden disturbances, and present a current-inference transformer RL framework in which a causal transformer encodes recent observation-action histories, an auxiliary head predicts the latent current vector from privileged simulator labels during training (Kumar et al., 2021; Lee et al., 2020), and the actor outputs a residual correction (Johannink et al., 2019; Silver et al., 2018) on top of a line-of-sight guidance prior in a remus-style 6-DOF simulation environment.

Contributions. The main contributions of this paper are summarized as follows. First, we formulate strong-current AUV navigation as a partially observable RL problem where hidden current effects are inferred from observation-action history. Second, we propose a transformer-based PPO residual controller with an auxiliary current-estimation head and current-conditioned action policy. Lastly, we provide a preliminary evaluation in a REMUS-style 6-DOF environment and compare against a guidance-only baseline under an intensified-current stress test.

Organization. The rest of this paper is organized as follows. Sec. 2 summarizes related work. Sec. 3 presents environmental setting and Sec. 4 introduced our proposed method. Sec. 5 evaluates the performance of the method and Sec. 6

[1]Department of Defense Convergence Technology, Korea University, Seoul, Korea [2]School of Electrical Engineering, Korea University, Seoul, Korea. Correspondence to: Joongheon Kim <joongheon@korea.ac.kr>.

ICML 2026 Workshop (RLxF: Reinforcement Learning from World Feedback), Seoul, Korea (July 10, 2026)

concludes this paper.

2. Related Work

AUV and RL-based Robotic Control. AUV navigation has traditionally relied on model-based guidance and control methods, including line-of-sight guidance, sliding-mode control, adaptive control, disturbance observers, and model predictive control (Fossen et al., 2003; Fossen, 2011). These methods provide strong structure and interpretability, but their performance can depend on accurate vehicle models or explicit disturbance estimation, both of which are difficult to obtain when the ocean current is hidden and time-varying. Reinforcement learning provides an alternative by learning control policies from interaction, and has been applied to robotic continuous control, path planning, target reaching, and obstacle avoidance for underwater vehicles (Hadi et al., 2022; Carlucho et al., 2018). Proximal policy optimization (PPO) is widely used in continuous-action settings because its clipped policy update improves training stability (Schulman et al., 2017), and is typically combined with generalized advantage estimation (Schulman et al., 2016). Our method combines these directions: we retain a classical guidance prior for basic target tracking, while learning transformer-based residual corrections for hidden current compensation, trained under a curriculum (Bengio et al., 2009) that gradually increases task difficulty.

Sequence-Model Policies for Partial Observability. When the underlying state is not fully observable, memoryless policies can struggle to recover the missing information. Recurrent value-based and actor-critic methods address this by maintaining hidden state across time (Hausknecht & Stone, 2015), and recent work shows that recurrent model-free RL is a strong baseline across many partially observable Markov decision processes (POMDPs) (Ni et al., 2022). Transformers offer an alternative that has been used in RL to model trajectories, observation histories, and long-horizon dependencies (Vaswani et al., 2017; Parisotto et al., 2020; Chen et al., 2021). In our setting, the ocean current is not directly given to the policy, but its effect appears over time through vehicle motion. We therefore use a causal transformer to encode observation-action history and infer a latent current representation for robust residual control.

Residual Policy Learning and Privileged Adaptation. Residual policy learning combines a fixed prior controller with a learned residual policy, retaining the safety and interpretability of classical control while letting RL handle compensation that is hard to model (Johannink et al., 2019; Silver et al., 2018). A complementary line of work tackles partial observability of physical parameters by training with privileged information that is unavailable at deployment. Rapid Motor Adaptation (RMA) trains an environment encoder from privileged context (e.g., terrain and dynamics parameters) and then distills it into a history-conditioned online estimator (Kumar et al., 2021); similar teacher-student schemes have enabled robust quadrupedal locomotion (Lee et al., 2020). Domain randomization further improves transfer by training over a wide distribution of dynamics (Tobin et al., 2017). We unify these ideas in a single transformer policy that jointly produces history encoding, current estimation, and residual action.

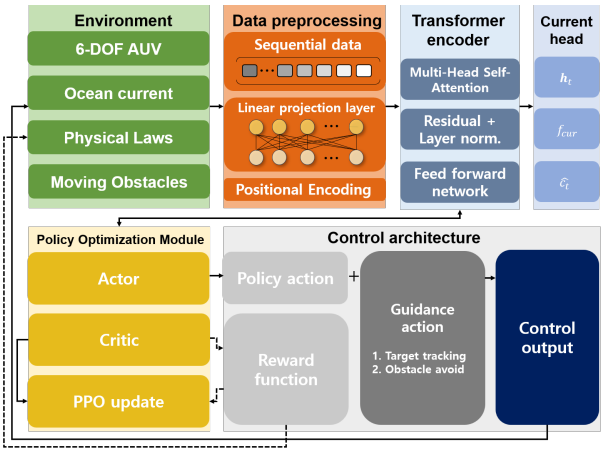

Figure 1. Overall architecture of the proposed framework.

3. Environment and State Definition

We consider a REMUS-style AUV navigating in a bounded three-dimensional workspace with randomized goals, moving obstacles, and time-varying ocean currents. The simulator follows a standard 6-DOF marine craft model with rigid-body and added-mass dynamics, Coriolis effects, hydrodynamic damping, hull lift, cross-flow drag, actuator saturation, and first-order actuator lag. The full dynamics are provided in Appendix A.

The key source of partial observability is the hidden ocean current. The simulator uses the true current internally to compute hydrodynamic forces from the relative velocity between the AUV and the surrounding fluid, but the current vector is not included in the policy observation. Thus, the policy must infer current-induced drift from recent interaction history rather than from direct current measurements.

The continuous action consists of propeller, rudder, and stern-plane commands. The observation contains the vehicle pose and velocity, relative goal position, normalized actuator state, and nearby obstacle information. The reward combines goal progress, heading alignment, action regularization, attitude stability, obstacle clearance, and terminal success or failure terms. Full observation, action, and reward definitions are provided in Appendix B.

4. Proposed Framework

Figure 1 shows the overall architecture of the proposed current-inference transformer PPO framework. The environment provides physics-based AUV transitions, ocean-current disturbances, moving obstacles, and reward feedback. In this setting, world feedback appears as the discrepancy between the commanded action and the realized vehicle motion under hidden current disturbances. The agent encodes recent interaction data with a causal transformer and outputs a residual policy action that is combined with a guidance action for final control.

Curriculum Learning. Training directly in the strongest current and obstacle setting is unstable. We therefore use a four-stage curriculum that gradually increases task difficulty from short-range target reaching to strong-current navigation with moving obstacles. The final stage includes randomized 3D targets, 12 moving obstacles, and strong time-varying ocean currents. In this stage, the horizontal current speed is sampled from a strong-current range of 0.55–0.95 m/s. Detailed curriculum parameters are provided in Appendix E.

Transformer-Based Current Inference. Because the ocean current is hidden from the policy, a single observation is insufficient to identify the disturbance. The policy therefore consumes a finite observation-action history:

$$\mathcal{H}_t = \{(\mathbf{o}_{t-K+1}, \mathbf{a}_{t-K}), \ldots, (\mathbf{o}_t, \mathbf{a}_{t-1})\}. \tag{1}$$

The history is encoded by a causal transformer, and the resulting feature is shared by the actor, critic, and auxiliary current-estimation head. Following the privileged-adaptation principle of RMA-style methods (Kumar et al., 2021; Lee et al., 2020), the current-estimation head predicts a latent current vector and is supervised with privileged simulator labels that are available in simulation but not at deployment:

$$\hat{\mathbf{c}}_t = f_{cur}(\mathbf{h}_t), \tag{2}$$

$$\mathcal{L}_{cur} = \|\hat{\mathbf{c}}_t - \mathbf{c}_t\|_2^2. \tag{3}$$

The true current is not provided to the policy during execution. Additional history encoding and PPO objective details are provided in Appendix C.

Guidance-Residual Control Architecture. Rather than applying the RL action directly to the AUV, we adopt the residual policy learning paradigm (Johannink et al., 2019; Silver et al., 2018) and use the policy output as a residual correction around a line-of-sight guidance prior:

$$\mathbf{a}_t = \text{clip}\left(\mathbf{a}_t^G + \mathbf{S}\mathbf{a}_t^{RL}, -1, 1\right), \tag{4}$$

where $\mathbf{a}_t^G$ is the guidance action and $\mathbf{S}$ scales the residual policy output. The guidance prior provides basic target tracking, while the transformer policy learns residual corrections for current compensation and obstacle-aware navigation.

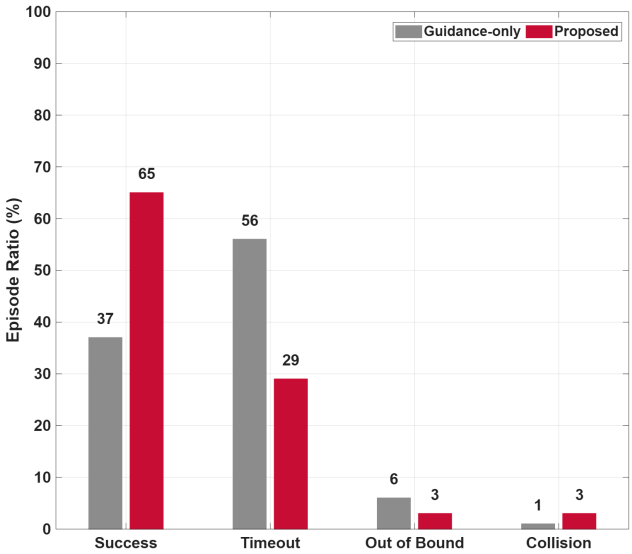

Figure 2. Comparison between guidance-only and proposed method with intensified Stage-4 currents.

5. Performance Evaluation

5.1. Evaluation Protocol

We evaluate the proposed current-inference transformer PPO framework using deterministic evaluations. Each evaluation consists of 100 randomized episodes. We report success rate, collision rate, timeout rate, final distance to the goal, path ratio, obstacle clearance, mean current speed, and current-estimation error. As an initial proof-of-concept study, we first analyze the proposed framework under the nominal curriculum and then compare it with a guidance-only baseline under an intensified-current stress test.

5.2. Evaluation Results

Curriculum Evaluation. Table 1 summarizes representative evaluation results across the curriculum. The selected validation checkpoints show stable task performance across stages, including the final strong-current moving-obstacle stage. In Stage 4, the selected checkpoint is evaluated under strong time-varying currents and 12 moving obstacles. The path ratio increases in the final stage because the AUV must compensate for strong current-induced drift and avoid moving obstacles.

Intensified-Current Stress Test. The nominal Stage-4 setting already includes strong time-varying currents, but the guidance prior remains a strong controller in that regime. To further probe current robustness, we evaluate both the guidance-only controller and the proposed current-inference transformer policy under an intensified-current setting, where the Stage-4 current magnitude is increased by approximately $1.5\times$. This corresponds to a horizontal current range of approximately 0.825–1.425 m/s. Figure 2

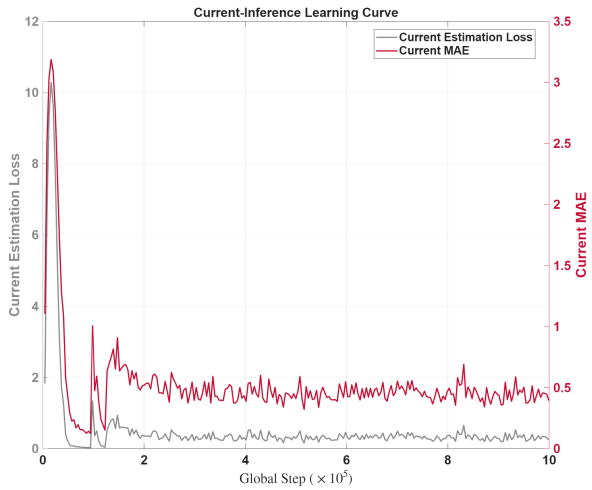

Figure 3. Current-estimation loss and MAE decrease during transformer PPO training.

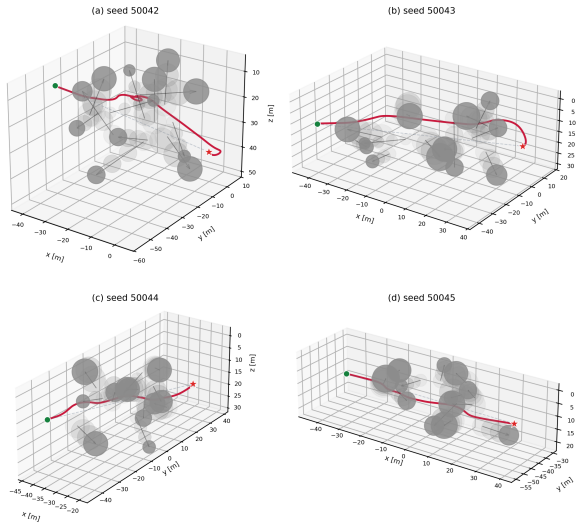

Figure 4. Successful Stage-4 trajectories under strong currents and moving obstacles.

Table 1. Representative evaluation performance across curriculum stages.

Stage	Setting	Step	Succ. [%] ↑	Coll. [%] ↓	Final dist. [m] ↓	Path ratio ↓
1	2.5D no current	20,480	100.0	0.0	1.23	0.97
2	3D current	81,920	100.0	0.0	1.24	0.99
3	3D obstacles	122,880	100.0	0.0	1.24	1.02
4	Strong current + moving obstacles	962,560	100.0	0.0	1.24	1.14

shows that the proposed policy achieves higher success than guidance-only control under intensified currents. The main failure mode for guidance-only control is timeout: its timeout rate is 56%, whereas the proposed policy reduces it to 29%. Out-of-bounds failures also decrease from 6% to 3%. However, the proposed policy does not improve every metric: its collision rate is slightly higher than the guidance-only controller. We interpret this as preliminary evidence that history-based current inference improves robustness under stronger disturbances rather than dominating the guidance prior, and the collision-rate trade-off motivates future work on residual-scale calibration and safety-aware constraints.

Current-Inference Learning. Since task success saturates near 99–100% in the nominal Stage-4 setting, we treat success and collision rates as task-level validation metrics rather than the main learning signal. Our main diagnostic is current inference. Figure 3 shows the auxiliary current-estimation loss and current MAE of the proposed model over training. Both metrics rapidly decrease in the early phase and then stabilize, suggesting that the transformer learns current-related information from observation-action histories. Figure 4 visualizes Stage-4 trajectories under strong currents and moving obstacles.

6. Concluding Remarks

We presented a current-inference transformer PPO framework that treats action-outcome mismatches as world feedback, encodes recent observation-action histories with a causal transformer, and uses an auxiliary current-estimation head to learn disturbance representations for residual control. The current-estimation loss and MAE decrease during training, and under an intensified-current stress test the method achieves 65% success versus 37% for guidance-only control while reducing timeout rate. These results provide initial evidence that history-based current inference improves robustness when disturbances exceed the nominal regime. Future work includes recurrent and MLP baselines (Ni et al., 2022), ablations on the current-estimation head and history length, and robustness evaluation under domain randomization (Tobin et al., 2017).

Impact Statement

This work studies reinforcement learning for AUV navigation under simulated ocean currents and could support safer underwater inspection and monitoring. Because real-world AUV deployment is safety-critical, the method requires further validation, fail-safe mechanisms, and conservative testing before deployment.

Acknowledgments

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2026-RS-2024-00436887) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation); and also by IITP grant funded by MSIT (RS-2024-00439803, SW Star Lab).

References

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *International Conference on Machine Learning (ICML)*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380.

Carlucho, I., De Paula, M., Wang, S., Petillot, Y., and Acosta, G. G. Adaptive low-level control of autonomous underwater vehicles using deep reinforcement learning. *Robotics and Autonomous Systems*, 107:71–86, 2018. doi: 10.1016/j.robot.2018.05.016.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097, 2021.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Fossen, T. I. *Handbook of Marine Craft Hydrodynamics and Motion Control*. John Wiley & Sons, 2011. doi: 10.1002/9781119994138.

Fossen, T. I., Breivik, M., and Skjetne, R. Line-of-sight path following of underactuated marine craft. *IFAC Proceedings Volumes*, 36(21):211–216, 2003. doi: 10.1016/S1474-6670(17)37809-6.

Hadi, B., Khosravi, A., and Sarhadi, P. Deep reinforcement learning for adaptive path planning and control of an autonomous underwater vehicle. *Applied Ocean Research*, 129:103326, 2022. doi: 10.1016/j.apor.2022.103326.

Hausknecht, M. and Stone, P. Deep recurrent Q-learning for partially observable MDPs. In *AAAI Fall Symposium on Sequential Decision Making for Intelligent Agents*, 2015.

Johannink, T., Bahl, S., Nair, A., Luo, J., Kumar, A., Loskyll, M., Ojea, J. A., Solowjow, E., and Levine, S. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6023–6029, 2019. doi: 10.1109/ICRA.2019.8794127.

Kumar, A., Fu, Z., Pathak, D., and Malik, J. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems (RSS)*, 2021. doi: 10.15607/RSS.2021.XVII.011.

Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., and Hutter, M. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020. doi: 10.1126/scirobotics.abc5986.

Ni, T., Eysenbach, B., and Salakhutdinov, R. Recurrent model-free RL can be a strong baseline for many POMDPs. In *International Conference on Machine Learning (ICML)*, pp. 16691–16723, 2022.

Parisotto, E., Song, H. F., Rae, J. W., Pascanu, R., Gulcehre, C., Jayakumar, S. M., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., Botvinick, M. M., Heess, N., and Hadsell, R. Stabilizing transformers for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pp. 7487–7498, 2020.

Prestero, T. T. J. Verification of a six-degree of freedom simulation model for the remus autonomous underwater vehicle. Master's thesis, Massachusetts Institute of Technology and Woods Hole Oceanographic Institution, 2001.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017.

Silver, T., Allen, K., Tenenbaum, J., and Kaelbling, L. Residual policy learning, 2018.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Environment Details

A.1. 6-DOF Marine Craft Dynamics

The AUV pose and body-frame velocity are defined as

$$\boldsymbol{\eta} = [x, y, z, \phi, \theta, \psi]^\top, \qquad \boldsymbol{\nu} = [u, v, w, p, q, r]^\top. \tag{5}$$

The pose evolves according to

$$\dot{\boldsymbol{\eta}} = \mathbf{J}(\boldsymbol{\eta})\boldsymbol{\nu}, \qquad \mathbf{J}(\boldsymbol{\eta}) = \begin{bmatrix} \mathbf{R}(\phi, \theta, \psi) & \mathbf{0} \\ \mathbf{0} & \mathbf{T}(\phi, \theta) \end{bmatrix}, \tag{6}$$

where $\mathbf{R}$ is the body-to-inertial rotation matrix and $\mathbf{T}$ maps body angular velocity to Euler-angle rates.

The dynamics follow the standard 6-DOF marine craft form (Fossen, 2011):

$$\mathbf{M}\dot{\boldsymbol{\nu}} + \mathbf{C}(\boldsymbol{\nu})\boldsymbol{\nu} = \boldsymbol{\tau}_{act} + \boldsymbol{\tau}_{hydro} + \boldsymbol{\tau}_{rest}. \tag{7}$$

The mass matrix is

$$\mathbf{M} = \mathbf{M}_{RB} + \mathbf{M}_A, \tag{8}$$

where $\mathbf{M}_{RB}$ is the rigid-body inertia matrix and $\mathbf{M}_A$ is the added-mass matrix. The added-mass model includes sway–yaw and heave–pitch coupling terms.

For a general mass matrix partitioned as

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix}, \qquad \boldsymbol{\nu} = [\boldsymbol{\nu}_1^\top, \boldsymbol{\nu}_2^\top]^\top, \tag{9}$$

the Coriolis and centripetal matrix is computed using

$$\mathbf{a} = \mathbf{M}_{11}\boldsymbol{\nu}_1 + \mathbf{M}_{12}\boldsymbol{\nu}_2, \qquad \mathbf{b} = \mathbf{M}_{21}\boldsymbol{\nu}_1 + \mathbf{M}_{22}\boldsymbol{\nu}_2, \tag{10}$$

and

$$\mathbf{C}(\boldsymbol{\nu}) = \begin{bmatrix} \mathbf{0} & -\mathbf{S}(\mathbf{a}) \\ -\mathbf{S}(\mathbf{a}) & -\mathbf{S}(\mathbf{b}) \end{bmatrix}, \tag{11}$$

where $\mathbf{S}(\cdot)$ is the skew-symmetric matrix operator.

A.2. Ocean Current and Relative Velocity

The ocean current is modeled as a slowly time-varying inertial-frame velocity:

$$\mathbf{v}_c^n(t) = \mathbf{v}_{c,0}^n + \mathbf{a}_c \odot \sin(\boldsymbol{\omega}_c t + \boldsymbol{\varphi}_c), \tag{12}$$

where $\mathbf{v}_{c,0}^n$ is the base current, $\mathbf{a}_c$ is the oscillation amplitude, $\boldsymbol{\omega}_c$ is the oscillation frequency, and $\boldsymbol{\varphi}_c$ is the phase.

The current is transformed into the body frame as

$$\mathbf{v}_c^b = \mathbf{R}(\phi, \theta, \psi)^\top \mathbf{v}_c^n. \tag{13}$$

Hydrodynamic forces are computed using the relative velocity between the vehicle and the surrounding fluid:

$$\boldsymbol{\nu}_r = \boldsymbol{\nu} - \boldsymbol{\nu}_c, \qquad \boldsymbol{\nu}_c = [(\mathbf{v}_c^b)^\top, 0, 0, 0]^\top. \tag{14}$$

The true current is used internally by the simulator but is not included in the policy observation.

A.3. Hydrodynamic Forces

The hydrodynamic force and moment are decomposed as

$$\boldsymbol{\tau}_{hydro} = \boldsymbol{\tau}_D + \boldsymbol{\tau}_H + \boldsymbol{\tau}_{CF}, \tag{15}$$

where $\boldsymbol{\tau}_D$ is residual damping, $\boldsymbol{\tau}_H$ is hull lift and moment, and $\boldsymbol{\tau}_{CF}$ is cross-flow drag.

The residual damping is

$$\boldsymbol{\tau}_D = -\left(\mathbf{D}_l \boldsymbol{\nu}_r + \mathbf{D}_q |\boldsymbol{\nu}_r| \odot \boldsymbol{\nu}_r\right). \tag{16}$$

The flow speed, angle of attack, and sideslip angle are

$$U = \|\boldsymbol{\nu}_{r,1:3}\|_2, \qquad \alpha = \tan^{-1}\left(\frac{w_r}{|u_r| + \epsilon}\right), \qquad \beta = \tan^{-1}\left(\frac{v_r}{|u_r| + \epsilon}\right). \tag{17}$$

Using dynamic pressure

$$q_d = \frac{1}{2}\rho U^2, \tag{18}$$

the hull force and moment terms are modeled as

$$X_H = -q_d S\left(C_{D0} + C_{D\alpha}\alpha^2 + C_{D\beta}\beta^2\right), \tag{19}$$
$$Y_H = -q_d S\left(C_{Y\beta}\beta + C_{Yr}\hat{r}\right), \tag{20}$$
$$Z_H = -q_d S\left(C_{Z\alpha}\alpha + C_{Zq}\hat{q}\right), \tag{21}$$
$$M_H = -q_d SL\left(C_{M\alpha}\alpha + C_{Mq}\hat{q}\right), \tag{22}$$
$$N_H = -q_d SL\left(C_{N\beta}\beta + C_{Nr}\hat{r}\right). \tag{23}$$

Thus,

$$\boldsymbol{\tau}_H = [X_H, Y_H, Z_H, 0, M_H, N_H]^\top. \tag{24}$$

Cross-flow drag is approximated by dividing the hull into N longitudinal sections. For section i at coordinate x_i,

$$v_i = v_r + x_i r, \qquad w_i = w_r - x_i q. \tag{25}$$

The sectional forces are

$$dY_i = -\frac{1}{2}\rho C_{D,cf} d\Delta x |v_i| v_i, \tag{26}$$
$$dZ_i = -\frac{1}{2}\rho C_{D,cf} d\Delta x |w_i| w_i, \tag{27}$$

where d is the hull diameter and Δx is the section length. The total cross-flow term is

$$\boldsymbol{\tau}_{CF} = \sum_{i=1}^{N} [0, dY_i, dZ_i, 0, -x_i dZ_i, x_i dY_i]^\top. \tag{28}$$

A.4. Actuator Dynamics

The action controls the propeller, rudder, and stern plane:

$$\mathbf{a}_t = [a_t^p, a_t^r, a_t^s]^\top \in [-1, 1]^3. \tag{29}$$

The desired actuator states are

$$n_d = f_p(a_t^p), \qquad \delta_{r,d} = a_t^r \delta_{r,\max}, \qquad \delta_{s,d} = a_t^s \delta_{s,\max}. \tag{30}$$

Each actuator follows first-order lag with rate limits:

$$\dot{\zeta} = \text{clip}\left(\frac{\zeta_d - \zeta}{T_\zeta}, -\dot{\zeta}_{\max}, \dot{\zeta}_{\max}\right), \qquad \zeta \in \{n, \delta_r, \delta_s\}. \tag{31}$$

The resulting actuator force and moment are

$$\boldsymbol{\tau}_{act} = \boldsymbol{\tau}_{prop} + \boldsymbol{\tau}_{fin}. \tag{32}$$

B. Observation and Reward Details

B.1. Observation

The policy observation is

$$\mathbf{o}_t = \left[\boldsymbol{\eta}_t,\ \boldsymbol{\nu}_t,\ \mathbf{R}_t^\top(\mathbf{p}_g - \mathbf{p}_t),\ \bar{\boldsymbol{\zeta}}_t,\ \mathbf{o}_t^{obs}\right], \tag{33}$$

where $\mathbf{p}_g$ is the goal position, $\bar{\boldsymbol{\zeta}}_t$ is the normalized actuator state, and $\mathbf{o}_t^{obs}$ contains relative body-frame obstacle positions and radii. The current vector is hidden from the policy input.

B.2. Reward Decomposition

The total reward is

$$r_t = r_t^{prog} + r_t^{close} + r_t^{align} - c_t^{step} - c_t^{act} - c_t^{smooth} - c_t^{att} - c_t^{obs} + r_t^{term}. \tag{34}$$

The progress reward is

$$r_t^{prog} = w_p \frac{d_{t-1}^g - d_t^g}{d_0^g}, \qquad d_t^g = \|\mathbf{p}_t - \mathbf{p}_g\|_2. \tag{35}$$

The closing-speed and heading-alignment rewards are

$$r_t^{close} = w_c \max\left(\frac{d_{t-1}^g - d_t^g}{\Delta t}, 0\right), \qquad r_t^{align} = w_a \max(0, \cos e_t^\psi), \tag{36}$$

where e_t^ψ is the heading error.

The regularization penalties are

$$c_t^{step} = w_s, \qquad c_t^{act} = w_u\|\mathbf{a}_t\|_2^2, \qquad c_t^{smooth} = w_{\Delta u}\|\mathbf{a}_t - \mathbf{a}_{t-1}\|_2^2. \tag{37}$$

The attitude penalty is

$$c_t^{att} = w_\phi|\phi_t| + w_\theta|\theta_t|. \tag{38}$$

Obstacle avoidance is encouraged using the minimum clearance

$$c_t = \min_i \left(\|\mathbf{p}_t - \mathbf{p}_{i,t}^{obs}\|_2 - r_{AUV} - r_i^{obs}\right), \tag{39}$$

with penalty

$$c_t^{obs} = w_o \max(0, m_o - c_t), \tag{40}$$

where m_o is the obstacle safety margin.

The terminal reward is

$$r_t^{term} = R_g\mathbb{I}_{goal} - C_{col}\mathbb{I}_{col} - C_{out}\mathbb{I}_{out} - \mathbb{I}_{time}\left(C_{time} + w_d d_t^g\right). \tag{41}$$

C. Transformer Policy and Training Objective

C.1. History Encoding

The policy receives a history of observations and previous actions:

$$\mathcal{H}_t = \{(\mathbf{o}_{t-K+1}, \mathbf{a}_{t-K}), \ldots, (\mathbf{o}_t, \mathbf{a}_{t-1})\}. \tag{42}$$

Each pair is embedded as

$$\mathbf{z}_i = \mathbf{W}_{in}[\mathbf{o}_i, \mathbf{a}_{i-1}] + \mathbf{p}_i, \tag{43}$$

where $\mathbf{p}_i$ is a learned positional embedding. A causal transformer encodes the sequence:

$$\mathbf{h}_t = \mathrm{Transformer}_{causal}(\mathbf{z}_{t-K+1:t})_K. \tag{44}$$

C.2. Current Estimation and Residual Control

The current-estimation head predicts

$$\hat{\mathbf{c}}_t = f_{cur}(\mathbf{h}_t), \tag{45}$$

and is trained using privileged simulator current labels:

$$\mathcal{L}_{cur} = \|\hat{\mathbf{c}}_t - \mathbf{c}_t\|_2^2. \tag{46}$$

The actor conditions on both the transformer feature and the inferred current:

$$\mathbf{a}_t^{RL} \sim \pi_\theta\left(\cdot \mid [\mathbf{h}_t, \hat{\mathbf{c}}_t]\right). \tag{47}$$

The final applied action is a residual correction around the guidance prior:

$$\mathbf{a}_t = \text{clip}\left(\mathbf{a}_t^G + \mathbf{S}\mathbf{a}_t^{RL}, -1, 1\right), \tag{48}$$

where $\mathbf{a}_t^G$ is the guidance action and $\mathbf{S}$ scales the residual action.

C.3. PPO Objective

The policy is optimized with PPO (Schulman et al., 2017). The clipped policy objective is

$$\mathcal{L}_{PPO} = -\mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t, \text{clip}\left(\rho_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right], \tag{49}$$

where

$$\rho_t(\theta) = \frac{\pi_\theta(\mathbf{a}_t|\mathcal{H}_t)}{\pi_{\theta_{old}}(\mathbf{a}_t|\mathcal{H}_t)}. \tag{50}$$

The full training loss is

$$\mathcal{L} = \mathcal{L}_{PPO} + \lambda_V\mathcal{L}_V - \lambda_H\mathcal{H}(\pi_\theta) + \lambda_{cur}\mathcal{L}_{cur}, \tag{51}$$

where $\mathcal{L}_V$ is the value loss, $\mathcal{H}(\pi_\theta)$ is the policy entropy, and λ_{cur} is the current-estimation loss coefficient.

D. Training Pseudocode

At each step, the agent builds an observation-action history, encodes it with a causal transformer, and predicts a latent current estimate. The policy then outputs a residual action, which is combined with a classical guidance prior before being applied to the AUV environment. The policy is updated with the PPO objective together with the auxiliary current-estimation loss.

Algorithm 1 Current-Inference Transformer PPO

```
Initialize transformer actor-critic policy π_θ, current head f_cur, and rollout buffer B
for curriculum stage s = 1, ..., S do
    Configure environment difficulty for stage s
    for each PPO rollout do
        for each environment step t do
            Build history H_t = {(o_{t-K+1}, a_{t-K}), ..., (o_t, a_{t-1})}
            Encode history: h_t = Transformer_causal(H_t)
            Estimate current: ĉ_t = f_cur(h_t)
            Sample residual action: a_t^RL ~ π_θ(·|[h_t, ĉ_t])
            Compute guidance action a_t^G
            Apply action a_t = clip(a_t^G + S a_t^RL, -1, 1)
            Step environment and store transition with current label c_t in B
        end for
        Update π_θ with PPO loss and auxiliary loss L_cur = ||ĉ_t - c_t||_2^2
    end for
end for
```

E. Curriculum Details

In the final stage, the current speed is sampled from a strong-current range and the obstacles are biased toward the mission path. This stage is used as the main stress-test environment for evaluating robust navigation under hidden disturbances.

Table 2. Curriculum stages.

Stage	Description	Obstacles	Current	Target distance
1	Short 2.5D target reaching	0	Disabled	25–45 m
2	Medium 3D target reaching	0	Enabled	40–80 m
3	3D navigation with obstacles	6	Enabled	50–100 m
4	Strong-current moving-obstacle navigation	12	Strong	60–110 m