Extended Abstract

Motivation. Autonomous marine vehicles must track desired trajectories under realistic, non-stationary ocean currents. Model Predictive Control (MPC) can provide expert-like control, but deploying MPC-like strategies at scale may be computationally expensive and sensitive to changing ocean dynamics. This project asks whether imitation learning can make an autonomous kayak follow trajectories like an expert controller. This capability is a foundational step toward a broader long-term vision, *Agentic Blue*: multi-agent autonomous marine systems that can collaborate for coastal monitoring and environmental missions such as plastic debris and sargassum monitoring or capture.

Method. I formulate marine trajectory tracking as scenario-conditioned imitation learning. Each policy receives a kayak state s_t and a scenario descriptor

$$z = [\text{trajectory type},\ \lambda,\ \text{expert mode}],$$

where the trajectory type is circle, lemniscate, or spiral, and $\lambda \in \{0.0, 0.5, 1.0\}$ denotes current intensity. The action is a continuous differential-thrust command $a_t = [a_L, a_R] \in [-1, 1]^2$. I compare four imitation strategies trained from MPC demonstrations: Conditional Behavioral Cloning (BC), BC-GMM, Conditional Flow Matching (FM), and PPO-GAIL / H-GAIL-inspired adversarial refinement. The main metric is closed-loop mean cross-track error (CTE) in GPUOcean simulations.

Implementation. The environment integrates GPUOcean current fields with a simplified Fossen-style kayak dynamics model (3). A current-aware MPC expert tracks three trajectory families across three current regimes, producing state-action demonstrations. Initial flat BC achieved low supervised error but failed catastrophically in closed-loop rollouts, revealing that the core challenge is *regime ambiguity*: similar states can require different actions depending on the trajectory geometry and current intensity. This motivated conditioning policies on trajectory identity, current scale, and expert mode.

Results. The results show that no single imitation method dominates all trajectory-current regimes. In the lemniscate scenario with medium current ($\lambda = 0.5$), early-stopped PPO-GAIL achieved the lowest CTE, outperforming the MPC reference. In the spiral scenario with strong current ($\lambda = 1.0$), Flow Matching with averaged stochastic sampling achieved the best performance. In the circle scenario with medium current ($\lambda = 0.5$), deterministic Flow Matching produced the lowest CTE. Across selected regimes, the best learned policies matched or improved MPC, but the winning method changed by scenario.

Discussion. The results support the hypothesis that scenario conditioning is critical for marine imitation learning under stochastic currents. Generative models such as Flow Matching and BC-GMM can remain near MPC performance in several regimes, while adversarial refinement through PPO-GAIL can be powerful but unstable. Direct BC-GMM-to-PPO distillation achieved low offline action error but did not preserve closed-loop robustness, suggesting covariate shift and motivating DAgger-style correction (8).

Conclusion. Scenario-conditioned imitation learning can produce MPC-competitive tracking policies in realistic ocean-current simulations. However, performance is strongly scenario-dependent. This motivates a future hierarchical imitation framework in which a high-level selector chooses among generative and adversarial low-level policies according to trajectory-current regime. Future work includes stabilizing PPO-GAIL, extending toward full H-GAIL (5), adding DAgger-style BC-GMM distillation, and validating learned controllers on an autonomous kayak platform with long-term extensions toward LoRa-based multi-kayak coordination and solar autonomy in Yucatan coastal waters.

Generative and Hierarchical Imitation Learning for Marine Trajectory Control in Stochastic Ocean Currents

Omar Eduardo Jimenez Lopez
Stanford University / Universidad Panamericana
ejimenel@stanford.edu ojimenel@up.edu.mx

Abstract

This project studies whether imitation learning can make an autonomous kayak track trajectories like an expert controller under realistic, non-stationary ocean currents. A GPUOcean-based marine simulation environment is combined with a simplified kayak dynamics model and a current-aware MPC expert. The expert generates demonstrations for circle, lemniscate, and spiral trajectories across calm, moderate-current, and full-current regimes. Initial flat Behavioral Cloning (BC) achieves low supervised loss but fails catastrophically in closed-loop rollouts, revealing regime ambiguity across trajectory geometries and current intensities. To address this, I compare scenario-conditioned imitation strategies: Conditional BC, BC-GMM, Conditional Flow Matching, and PPO-GAIL / H-GAIL-inspired adversarial refinement. The results show that learned policies can match or improve MPC in selected regimes, but no single method dominates all conditions. Early-stopped PPO-GAIL performs best in the lemniscate medium-current case, while Flow Matching dominates spiral high-current and circle medium-current scenarios. These findings motivate a hierarchical imitation framework in which a high-level module selects the most appropriate low-level policy for each trajectory-current regime.

1 Introduction

Autonomous marine vehicles operating in coastal waters must track trajectories while adapting to non-stationary currents. This capability is important for environmental monitoring, inspection, and future multi-agent missions such as plastic debris and sargassum monitoring or capture. Classical controllers such as Model Predictive Control (MPC) can produce high-quality behavior, but can be computationally expensive and sensitive to modeling assumptions when deployed across diverse ocean conditions.

This project is motivated by a broader vision called *Agentic Blue*: autonomous marine agents that can eventually operate as a collaborative ecosystem in realistic ocean environments. Before multiple kayaks can coordinate as a collective intelligence, each individual kayak must first learn a fundamental skill: tracking desired trajectories reliably under ocean currents. This report therefore focuses on whether imitation learning can make a single autonomous kayak follow trajectories like an expert MPC controller.

An initial hypothesis was that a learned policy could imitate expert actions from demonstrations and then be refined using adversarial imitation learning (4). However, early experiments revealed a more fundamental issue. A flat BC policy achieved very low validation mean-squared error, but diverged by kilometers in closed-loop evaluation under nonzero currents. This indicates that supervised action prediction alone is insufficient; the policy must resolve *regime ambiguity* across trajectory geometries and current intensities.

The contributions of this project are:

- A GPUOcean-based marine imitation learning benchmark using a kayak dynamics model and current-aware MPC expert demonstrations.
- A scenario-conditioned formulation that augments the kayak state with trajectory type, current scale, and expert mode.
- A comparison of Conditional BC, BC-GMM, Conditional Flow Matching, and PPO-GAIL / H-GAIL-inspired adversarial refinement.
- An empirical finding that the best imitation strategy is scenario-dependent: Flow Matching dominates selected circle and spiral regimes, while early-stopped PPO-GAIL performs best in a lemniscate medium-current regime.

2 Related Work

Imitation learning. Behavioral Cloning learns a policy by supervised regression from expert states to expert actions. It is simple and effective on the expert state distribution, but can suffer from compounding error under closed-loop rollouts due to covariate shift. DAgger addresses this issue by aggregating expert-labeled states induced by the learner's policy (8). Mandlekar et al. provide a systematic comparison of offline imitation methods for robot manipulation (7).

Adversarial imitation. Generative Adversarial Imitation Learning (GAIL) frames imitation as matching expert and learner occupancy measures using a discriminator and policy optimization (4), building on prior work in inverse reinforcement learning (1; 11). PPO is a common policy optimization backbone due to its clipped surrogate objective and relative stability (10). In this project, PPO-GAIL is used as an H-GAIL-inspired adversarial refinement stage, where the policy and discriminator are both conditioned on the trajectory-current regime.

Generative action models. Recent generative approaches model multimodal action distributions more flexibly than deterministic regression. Flow Matching learns a vector field transporting a simple base distribution into a target distribution (6). Related action-generation approaches such as Diffusion Policy demonstrate that generative models can capture complex action distributions for control (2). BC-GMM similarly represents conditional action distributions as mixtures of Gaussians, reducing mode averaging.

Marine simulation and control. GPUOcean provides realistic ocean-current fields for simulating marine dynamics (9). The kayak model is based on simplified Fossen-style marine vehicle dynamics (3). The learned policies are evaluated in closed-loop simulation using mean cross-track error.

3 Method

3.1 Conditional Policy Formulation

The task is to control an autonomous kayak so that it tracks a desired trajectory under ocean-current disturbances. The policy observes the kayak state and local flow information, augmented by a scenario descriptor

$$z = [\text{trajectory type},\ \lambda,\ \text{expert mode}], \tag{1}$$

where the trajectory type belongs to {circle, lemniscate, spiral}, and $\lambda \in \{0.0, 0.5, 1.0\}$ denotes current intensity. The action is continuous differential thrust:

$$a_t = [a_L, a_R] \in [-1, 1]^2, \tag{2}$$

where a_L and a_R control the left and right kayak thrusters. The learned policy is

$$\pi_\theta(a_t \mid s_t, z), \tag{3}$$

where s_t is the kayak state and local flow observation.

The main closed-loop evaluation metric is mean cross-track error:

$$\mathrm{CTE}_{\mathrm{mean}} = \frac{1}{T}\sum_{t=1}^{T} d(p_t, \tau), \tag{4}$$

where $d(p_t, \tau)$ denotes the distance between the kayak position p_t and the reference trajectory τ.

3.2 MPC Expert

The expert controller is a current-aware MPC policy. The MPC tracks the reference trajectory while compensating for lateral ocean-current disturbances using the torque correction

$$\tau_{r,\mathrm{des}} = |N_r|\, r_{\mathrm{des}} - K_{\mathrm{sway}}\, c_{\mathrm{sway}}, \tag{5}$$

where N_r is the yaw damping coefficient, r_{des} is the desired yaw rate, c_{sway} is the lateral current in the body frame, and K_{sway} is a feed-forward compensation gain. Demonstrations are collected across three trajectories and three current scales to define the expert dataset

$$\mathcal{D}_E = \{(s_t, a_t, z)\}_{t=1}^{N}. \tag{6}$$

3.3 Conditional Behavioral Cloning

Conditional BC trains a deterministic policy by minimizing supervised action prediction error:

$$\min_{\theta} \mathbb{E}_{(s,a,z)\sim\mathcal{D}_E}\left[\|\pi_\theta(s,z) - a\|_2^2\right]. \tag{7}$$

This model tests whether explicit scenario conditioning can reduce the regime ambiguity observed in flat BC.

3.4 BC-GMM

BC-GMM models the conditional expert action distribution as a Gaussian mixture:

$$p_\theta(a \mid s, z) = \sum_{k=1}^{K} w_k(s,z)\,\mathcal{N}\left(a; \mu_k(s,z), \Sigma_k(s,z)\right). \tag{8}$$

At evaluation time, the policy can use a weighted-mean action or select the mean of the most likely component. This model tests whether multimodal supervised imitation is sufficient for robust closed-loop control.

3.5 Conditional Flow Matching

Flow Matching defines a generative action model by learning a conditional vector field (6):

$$\frac{dx_t}{dt} = v_\theta(x_t, t, s, z), \tag{9}$$

with $x_0 \sim \mathcal{N}(0, I)$ and $x_1 \approx a_E$. The model generates MPC-like actions conditioned on the kayak state and scenario descriptor. I evaluate both deterministic prediction and averaged stochastic sampling over multiple generated actions.

3.6 PPO-GAIL / H-GAIL-Inspired Refinement

The adversarial imitation stage uses a discriminator conditioned on state, action, and scenario (4):

$$\max_{\phi} \mathbb{E}_{(s,a,z)\sim\mathcal{D}_E}[\log D_\phi(s,a,z)] + \mathbb{E}_{(s,a,z)\sim\pi_\theta}[\log(1 - D_\phi(s,a,z))]. \tag{10}$$

The policy receives the learned imitation reward

$$r_{\mathrm{GAIL}}(s,a,z) = -\log(1 - D_\phi(s,a,z)), \tag{11}$$

and is updated using PPO (10):

$$\theta \leftarrow \mathrm{PPO}(\pi_\theta, r_{\mathrm{GAIL}}). \tag{12}$$

This is H-GAIL-inspired (5) because the policy and discriminator are scenario-conditioned, and the long-term goal is to learn a high-level selector over trajectory-current regimes.

Table 1: Milestone-stage mean CTE (m) for MPC, flat BC, and Conditional BC. Flat BC often diverges catastrophically despite low supervised loss, while Conditional BC reduces regime ambiguity.

Trajectory	Current	MPC	Flat BC	Cond. BC	Cond./MPC
Circle	$\lambda = 0.0$	0.3	2.0	3.5	11.9×
Circle	$\lambda = 0.5$	9.0	2965.4	13.8	1.5×
Circle	$\lambda = 1.0$	10.4	3502.6	11.9	1.1×
Lemniscate	$\lambda = 0.0$	5.9	130.8	5.9	1.0×
Lemniscate	$\lambda = 0.5$	8.7	8.4	19.4	2.2×
Lemniscate	$\lambda = 1.0$	10.3	3512.4	10.7	1.0×
Spiral	$\lambda = 0.0$	4.0	2520.5	12.4	3.1×
Spiral	$\lambda = 0.5$	9.1	8.8	9.8	1.1×
Spiral	$\lambda = 1.0$	9.8	3361.0	10.9	1.1×

4 Experimental Setup

Simulator. Experiments use a GPUOcean shallow-water simulation environment (9) with realistic, non-stationary current fields from the NorKyst-800 dataset (Lofoten region). The kayak is perturbed by local flow during closed-loop tracking.

Vehicle model. The vehicle is an autonomous kayak with differential left/right thruster commands based on simplified Fossen-style marine vehicle dynamics (3). The kayak includes a skeg for directional stability and a thruster arm of $B_{\text{thr}} = 0.36\,\text{m}$. Actions are clipped to $[-1, 1]^2$.

Trajectories and current regimes. The reference trajectories are circle ($R = 2000\,\text{m}$), lemniscate ($a = 2000\,\text{m}$), and spiral ($R_{\text{out}} = 2000\,\text{m}$, $R_{\text{in}} = 600\,\text{m}$). Each trajectory is evaluated under current scales $\lambda \in \{0.0, 0.5, 1.0\}$, corresponding to calm, moderate-current, and full NorKyst-800 current regimes.

Dataset. The current-aware MPC expert generated 12,000 state-action pairs across trajectory-current-expert conditions, with 12 accepted episodes and maximum CTE below 23 m. Each training example contains a 12-dimensional state observation, a 2-dimensional action, and a scenario descriptor z.

Implementation details. PyTorch models were run on CPU to avoid CUDA-context conflicts with GPUOcean. PPO-GAIL policies were initialized from supervised BC weights. For adversarial training, early stopping was important because later PPO updates sometimes degraded closed-loop tracking.

5 Results

5.1 Flat BC Failure and Regime Ambiguity

Flat BC achieved low supervised validation error (val MSE = 0.00008) but failed in closed-loop rollouts under nonzero currents, with CTE exceeding 2,500 m in several regimes. This revealed that predicting the correct action on the expert state distribution is not sufficient. The policy must know which trajectory-current regime it is operating in. Figure 1 visualizes this closed-loop failure mode, and Table 1 summarizes the milestone-stage comparison between MPC, flat BC, and Conditional BC.

5.2 Best Learned Policy by Scenario

Table 2 shows the best learned method in three representative trajectory-current regimes. Values below $1.0\times$ indicate lower CTE than the MPC reference, i.e., the learned policy outperforms the expert in that validation window.

The central result is that learned policies can match or improve MPC in selected regimes, but the best strategy is not universal. PPO-GAIL performs best in the lemniscate medium-current case, while Flow Matching dominates the selected spiral and circle regimes.

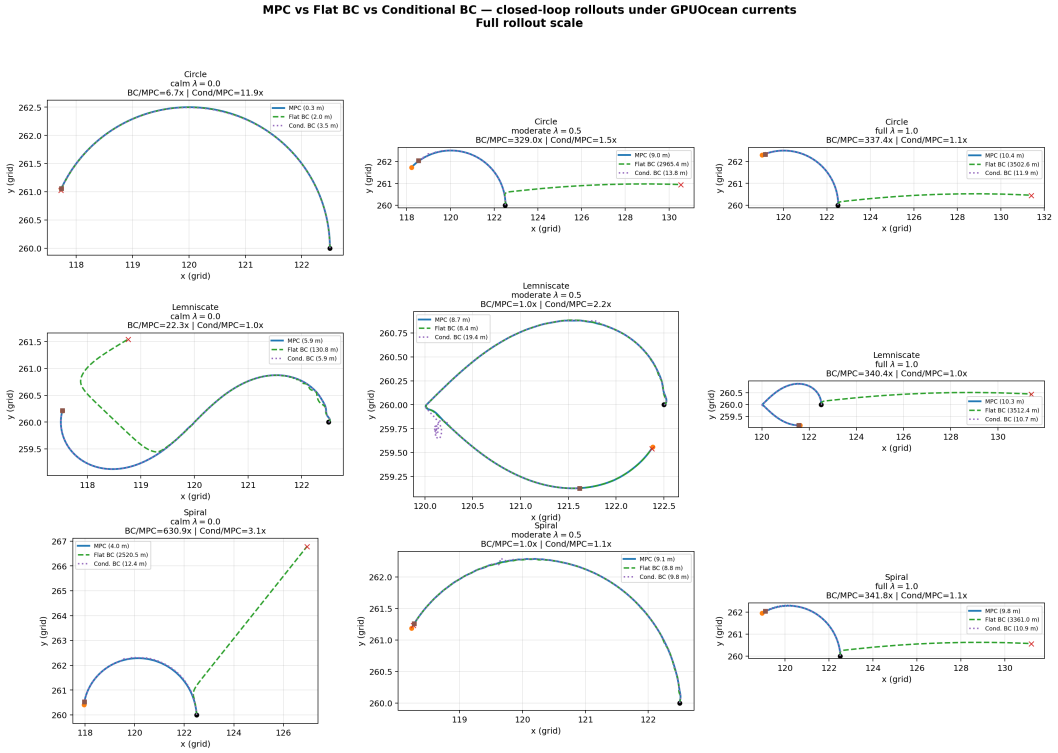

Figure 1: Closed-loop rollouts comparing the MPC expert, Flat BC, and Conditional BC under GPUOcean current fields. Flat BC achieves low supervised loss but diverges in closed-loop control when trajectory geometry and current intensity change. Conditional BC substantially reduces this failure by conditioning the policy on trajectory type, current scale, and expert mode.

Table 2: Best learned policy by selected scenario. The winning method changes across regimes, showing scenario-dependent specialization. CTE values correspond to closed-loop GPUOcean rollouts.

Scenario	Best Method	CTE (m)	Ratio vs. MPC
Lemniscate, $\lambda = 0.5$	PPO-GAIL (early stop)	6.95	0.79×
Spiral, $\lambda = 1.0$	FM average8	9.53	0.97×
Circle, $\lambda = 0.5$	FM deterministic	7.39	0.82×

A single iteration of DAgger-style data aggregation on the lemniscate $\lambda = 0.5$ scenario achieved CTE = 10.07 m (1.15× MPC), outperforming Conditional BC (2.22×) and both Flow Matching variants, but not reaching BC-GMM weighted (1.07×). This smoke test suggests that even one round of expert relabeling on learner-induced states meaningfully reduces covariate shift (8). Full DAgger convergence with multiple iterations and additional scenarios remains future work.

5.3 Full Method Comparison

Table 3 summarizes the full comparison for the three selected regimes. PPO-GAIL achieves a strong early-stopped result in the lemniscate case, but diverges in circle and spiral. Flow Matching and BC-GMM remain relatively close to MPC in several scenarios.

5.4 Effect of Stochastic Averaging in Flow Matching

Flow Matching was evaluated using both deterministic action prediction and averaged stochastic sampling over 8 rollouts. The stochastic average reduces closed-loop error in regimes where the

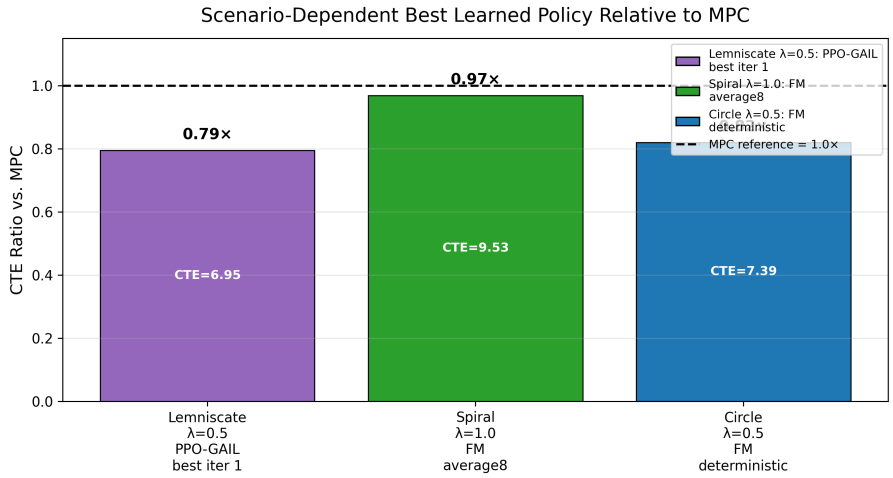

Figure 2: Best learned policy relative to MPC across selected trajectory-current regimes. Values below $1.0\times$ indicate lower mean CTE than MPC. The winning method varies by scenario, indicating that the learned control strategy is scenario-dependent.

Table 3: Closed-loop mean CTE (m) across selected scenarios. Lower is better. Large PPO-GAIL values indicate adversarial instability outside the lemniscate regime. BC-pretrained actor refers to the PPO initialization before adversarial updates. DAgger-style BC was evaluated only on lemniscate $\lambda = 0.5$ (smoke test, 1 iteration); dashes indicate not evaluated.

Method	Lem. λ=0.5	Spiral λ=1.0	Circle λ=0.5
MPC	8.75	9.83	9.01
Conditional BC	19.44	10.89	13.75
BC-GMM weighted	9.40	10.23	9.03
FM deterministic	56.13	15.47	7.39
FM average8	14.37	9.53	8.93
PPO-GAIL (best iter)	6.95	2232.71	1804.01
PPO-GAIL (final)	121.65	2232.81	1950.67
BC-pretrained actor	126.09	133.52	2317.74
DAgger-style BC (1 iter)	10.07	—	—

generative model captures useful action variability, although it is not uniformly beneficial across all current scales.

6 Discussion

Scenario conditioning is necessary but not sufficient. Flat BC fails because it averages across incompatible regimes. Conditional BC substantially reduces this ambiguity, but still underperforms in some scenarios. Thus, conditioning is necessary, but additional modeling capacity is needed for robust closed-loop behavior.

Generative models provide stable baselines. BC-GMM and Flow Matching can capture richer action distributions than deterministic BC. Flow Matching was especially strong in the selected circle and spiral regimes. The contrast between deterministic and averaged stochastic FM suggests that different trajectories benefit from different action-generation strategies, consistent with the multimodal character of the expert dataset.

Adversarial imitation is powerful but unstable. PPO-GAIL achieved the best result in the lemniscate medium-current case with early stopping, outperforming MPC in that validation window. However, it diverged in circle and spiral. This suggests that adversarial refinement can improve tracking, but requires stabilization, regularization, and careful stopping criteria (4).

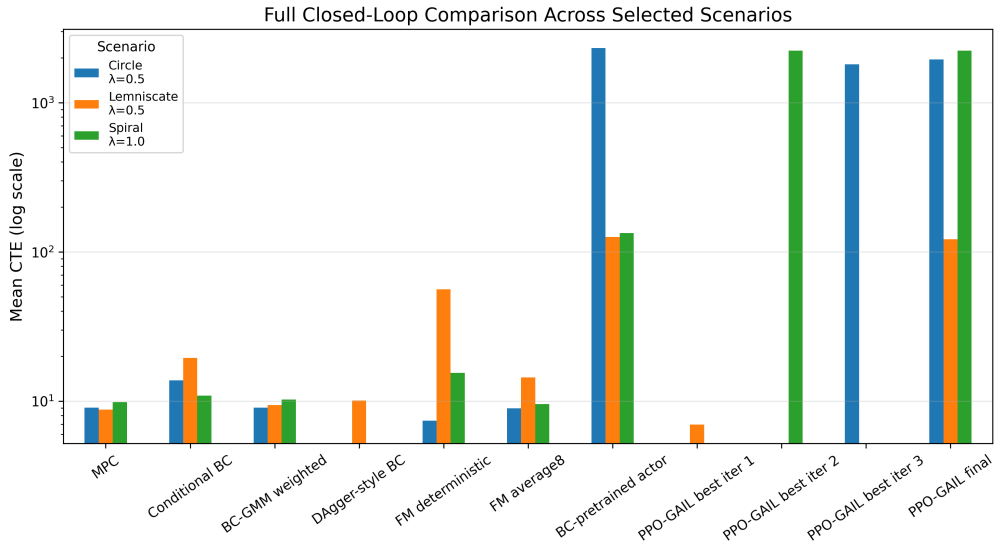

Figure 3: Full method comparison on log-scale CTE. Flow Matching and BC-GMM remain relatively robust near MPC in several regimes, while PPO-GAIL is highly scenario-sensitive: early stopping improves the lemniscate regime, but adversarial updates diverge in the selected circle and spiral scenarios.

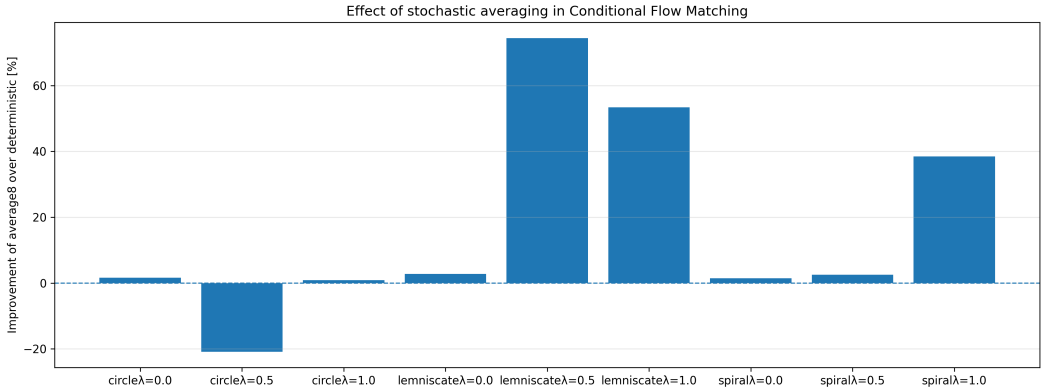

Figure 4: Effect of stochastic averaging in Conditional Flow Matching. Positive values indicate that averaging eight stochastic Flow Matching samples reduces mean CTE relative to deterministic prediction. Averaging provides substantial improvement in some regimes, especially lemniscate and spiral scenarios with stronger currents, suggesting that stochastic generative policies can capture useful action variability.

Toward hierarchical imitation. The most important finding is that the best method depends on the trajectory-current regime. This supports a hierarchical interpretation (5): a high-level module should infer or select the appropriate low-level policy depending on the scenario. Such a high-level selector is the natural next step toward full H-GAIL.

7 Conclusion

This project demonstrates that scenario-conditioned imitation learning can produce MPC-competitive marine tracking policies in realistic GPUOcean simulations. The best learned policies matched or improved MPC in selected regimes, but performance was strongly scenario-dependent. Flow Matching dominated the selected circle and spiral scenarios, while early-stopped PPO-GAIL performed best in

the lemniscate case. These results motivate a hierarchical imitation framework where a high-level selector chooses the most suitable learned policy for each trajectory-current regime.

Future work includes stabilizing PPO-GAIL with early stopping and regularization (10), developing DAgger-style BC-GMM distillation to reduce covariate shift (8), extending toward full H-GAIL with a learned high-level selector (5), and validating the learned controller on an autonomous kayak platform. Long-term deployment will explore LoRa-based multi-kayak coordination and solar autonomy for sargassum and plastic debris monitoring or capture in Yucatan coastal waters.

Team Contributions

All work was conducted by Omar Eduardo Jimenez Lopez. This includes the GPUOcean simulation integration, kayak dynamics setup, MPC expert development, dataset generation, BC and Conditional BC experiments, BC-GMM and Flow Matching baselines, PPO-GAIL / H-GAIL-inspired refinement experiments, analysis, poster preparation, and final report writing.

Changes from Proposal

The original proposal emphasized GAIL as a way to initialize reinforcement learning from expert demonstrations. After closed-loop experiments, the hypothesis changed: the main difficulty was not only exploration efficiency, but regime ambiguity across trajectory geometry and ocean-current intensity. This motivated a hierarchical and scenario-conditioned reformulation, with generative models and adversarial imitation evaluated as complementary strategies rather than sequential stages.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

[2] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[3] Thor I. Fossen. *Handbook of Marine Craft Hydrodynamics and Motion Control*. John Wiley & Sons, 1st edition, 2011.

[4] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[5] Hoang M. Le, Nan Jiang, Alekh Agarwal, Miroslav Dudík, Yisong Yue, and Hal Daumé III. Hierarchical imitation and reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 2917–2926, 2018.

[6] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[7] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021.

[8] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pages 627–635. PMLR, 2011.

[9] Martin L. Sætra, Trond Mannseth, Sølve Eidnes, and Annette Samuelsen. GPUOcean: GPU-accelerated shallow-water simulations for ocean modelling. *Geoscientific Model Development*, 16:4821–4845, 2023.

[10] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[11] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.

A Additional Implementation Notes

PyTorch models were run on CPU while GPUOcean used GPU resources, avoiding CUDA-context conflicts. Actions were clipped to $[-1, 1]^2$ for left and right thruster commands. For PPO-GAIL, the best checkpoint was selected using short closed-loop validation because final adversarial updates sometimes degraded tracking performance. The observation vector has 12 dimensions: $[x, y, \theta, u, v, r, u_c, v_c, c_{\text{surge}}, c_{\text{sway}}, e_{\text{cte}}/d_{\text{scale}}, e_\psi]$, where c_{surge} and c_{sway} are the ocean-current components in the kayak body frame.

B Additional Results

Direct BC-GMM-to-PPO actor distillation achieved low offline imitation error on expert states but failed in closed-loop validation. This suggests that offline distillation alone does not preserve robustness under learner-induced state distributions. A natural next step is DAgger-style BC-GMM distillation (8), where the BC-GMM or MPC labels states visited by the learner policy.

C Additional Diagnostic Figures

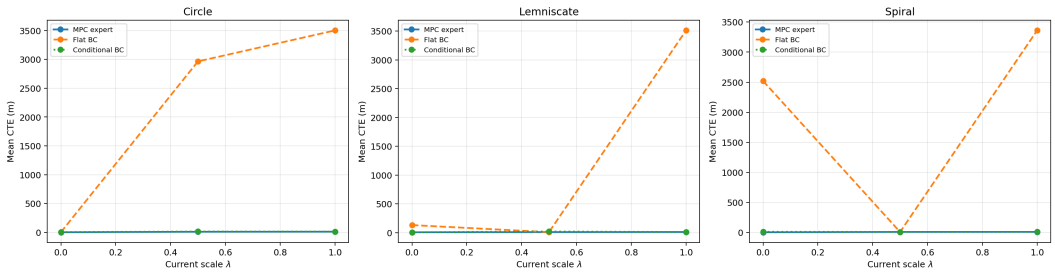

Figure 5: Robustness curves for MPC, Flat BC, and Conditional BC across current scales. Flat BC degrades catastrophically in several nonzero-current regimes, while Conditional BC remains much closer to MPC by using scenario information.

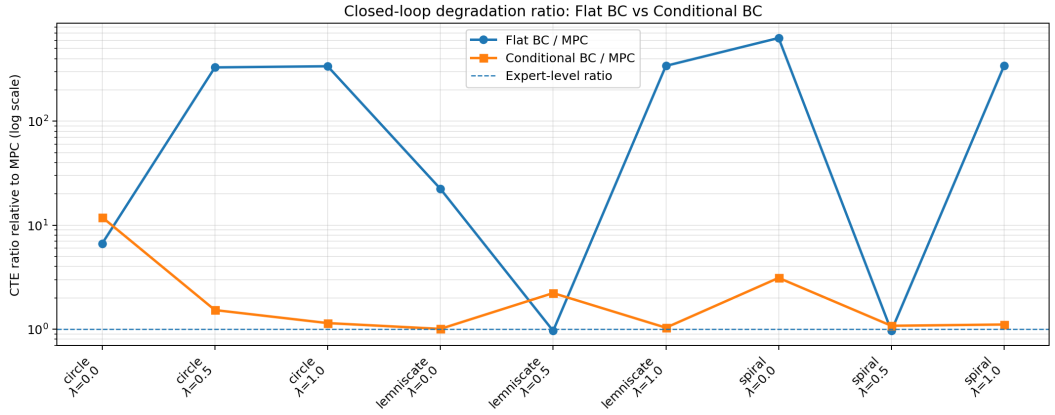

Figure 6: Closed-loop degradation ratio relative to MPC for Flat BC and Conditional BC on a logarithmic scale. Flat BC degrades by orders of magnitude under nonzero current regimes, while Conditional BC remains much closer to the MPC expert.

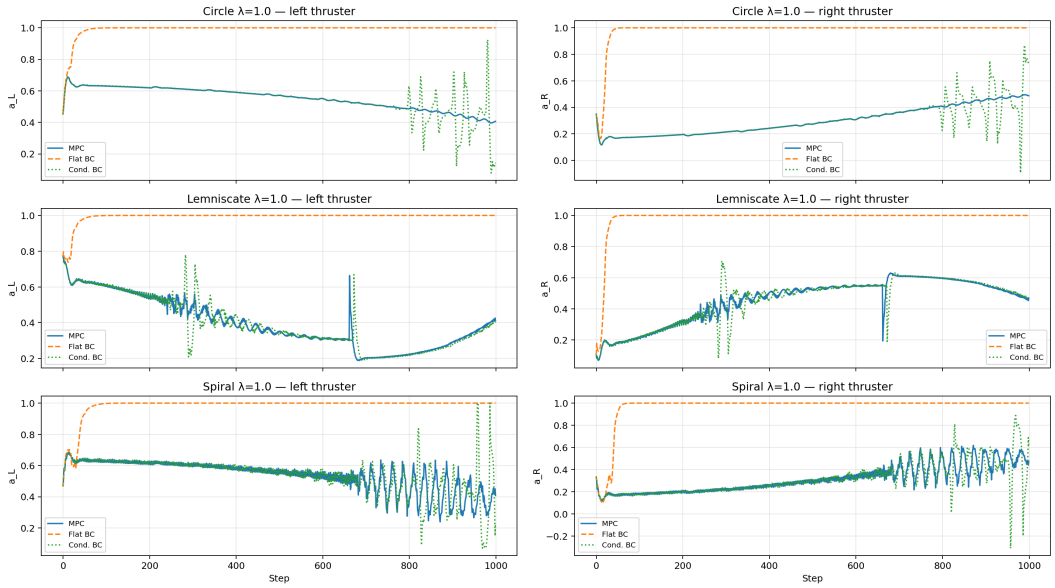

Figure 7: Left and right thruster action time series under full current ($\lambda = 1.0$). Flat BC saturates thruster commands immediately while Conditional BC more closely follows the MPC action structure. This action-level diagnostic explains the closed-loop divergence observed in Flat BC.

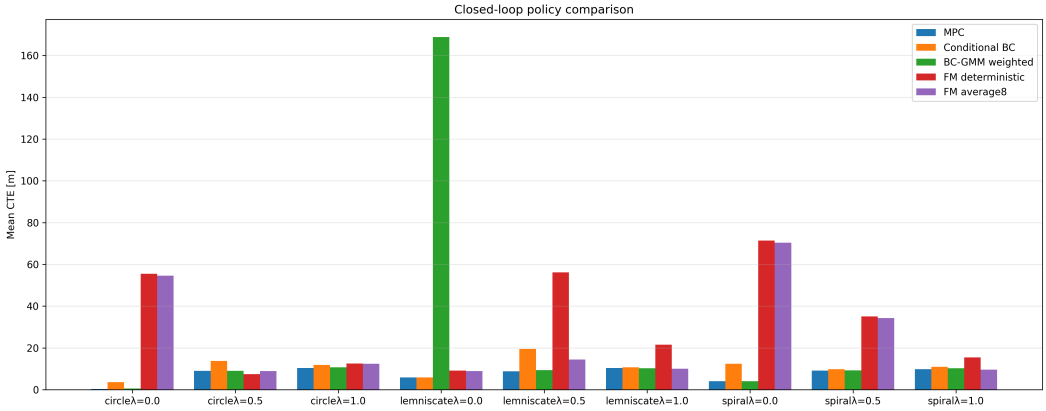

Figure 8: Global closed-loop CTE comparison across MPC, Conditional BC, BC-GMM, and Flow Matching variants. This diagnostic complements the selected-scenario results by showing broader method behavior across trajectory-current regimes.