scientific reports

Check for updates

OPEN

A hybrid APF-DQN framework with transformer-based current prediction for USV path planning in dynamic ocean environments

Nanjie Zhang[1], Yuquan Chen[1]✉, Yunshan Wu[2], Maoqin Ji[1] & Bing Wang[1]

Path planning for Unmanned Surface Vehicles (USVs) amid ocean currents and obstacles remains a challenging problem that has attracted considerable attention. However, existing methods fail to provide robust temporal prediction capabilities and do not effectively leverage the synergy between physics-based and learning-based approaches. To address these limitations, this paper presents a novel artificial-potential-field-guided deep Q-network (APF-DQN) with Transformer-based ocean current prediction for USV path planning in complex marine environments. First, a multi-scale Transformer architecture is employed for high-precision ocean current field prediction. Subsequently, an enhanced adaptive APF is proposed, incorporating a dynamic current-induced force field and an entropy-driven local minima escape mechanism. Furthermore, a median-Q-value-based exploration mechanism is introduced to improve the exploration efficiency of the standard ϵ-greedy strategy. Finally, through state-space augmentation, an APF-informed loss function, and policy fusion, a multi-level integration framework combining APF and DQN is established. Comparative simulation results confirm that the proposed framework achieves a 100% path success rate, 14.7% shorter trajectories, and 37.7% lower energy consumption compared to baseline methods.

Keywords Unmanned surface vehicles, Path planning, Artificial potential field, Deep reinforcement learning, Ocean current prediction

Unmanned Surface Vehicles (USVs) have emerged as versatile and cost-effective platforms for a wide range of marine applications, including oceanographic monitoring, environmental surveying, and maritime logistics[1–3]. However, the complex and dynamic nature of marine environments, characterized by time-varying ocean currents and densely distributed obstacles, poses significant challenges for autonomous USV navigation.

Intelligent path planning, as a critical enabling technology for USV operations, fundamentally determines mission efficiency, energy consumption, and operational safety. In practice, USVs encounter unique challenges: complex and time-varying ocean current dynamics, stringent obstacle avoidance requirements, and the need to balance multiple competing objectives such as energy efficiency and operational safety[4–7]. These challenges surpass the capabilities of conventional planning methods, necessitating the development of advanced algorithmic approaches.

Conventional path planning methods are generally classified into global and local approaches. Global algorithms, such as the A* algorithm[8], Dijkstra's algorithm[9], and Rapidly-exploring Random Trees (RRT)[10,11], generate optimal or near-optimal paths in static environments. However, these algorithms rely heavily on predefined environmental models, which limits their effectiveness in complex ocean current conditions. Furthermore, their computational complexity scales poorly with environment size, hindering real-time implementation. These methods also lack adaptability to ocean current variations, leading to significant deviations between planned and actual trajectories. Despite extensive research efforts to incorporate constraints and multi-objective optimization frameworks[12,13], fundamental limitations persist when operating in highly dynamic marine environments.

In contrast to global approaches, local planning methods, such as the Artificial Potential Field (APF) method[14], enable real-time navigation by utilizing virtual force fields that integrate target attraction with obstacle repulsion. Despite being computationally efficient and responsive, APF methods face two major limitations: susceptibility

[1]College of Artificial Intelligence and Automation, Hohai University, Changzhou 213000, China. [2]College of Information Engineering, Minzu University of China, Beijing 100000, China. ✉email: cyq@mail.ustc.edu.cn

to local minima in obstacle-dense environments[15] and limited adaptability to dynamic conditions[16]. Various enhancements have been proposed to address these challenges. For instance, Ge et al.[17] introduced virtual obstacle configurations to escape local minima, while Li et al.[18] integrated APF with the A* algorithm to improve global optimality under relatively simple environmental conditions. Nevertheless, these modifications have not adequately addressed the inherent limitations in complex dynamic environments[19].

Given the inherent limitations of traditional approaches, Deep Reinforcement Learning (DRL) methods have garnered considerable research attention. DRL offers a promising alternative by learning control policies through trial-and-error interactions without requiring precise environmental models. This characteristic makes DRL particularly suitable for addressing the uncertainties inherent in marine environments, such as time-varying currents and unpredictable obstacles. Among various DRL techniques, the Deep Q-Network (DQN)[20,21] has attracted significant attention due to its capacity for nonlinear function approximation in high-dimensional state spaces. DQN has demonstrated effectiveness in underwater vehicle path planning under current interference[22], suggesting its potential for USV applications. However, applying DQN to USV path planning presents several challenges, including sample inefficiency, high computational complexity, and constrained action space discretization[23].

In response to these challenges, hybrid frameworks that integrate DQN with traditional methodologies have been developed to leverage the strengths of both approaches while mitigating their respective limitations. For instance, Shen et al.[22] demonstrated the potential of such integration by combining DQN with APF through reward function optimization. However, this approach focuses primarily on reward-level integration and does not effectively harness deeper synergies between the two methods, such as state-space augmentation or policy-level fusion.

Despite these advancements, four fundamental limitations persist in current USV navigation approaches:

- Existing environmental models typically rely solely on real-time sensor data, lacking predictive capabilities for future ocean current states[24]. This limitation constrains proactive path planning for evolving current patterns, compromising both efficiency and safety of USV navigation in dynamic tidal zones.
- Traditional APF methods, while computationally efficient, suffer from susceptibility to local minima and limited adaptability to dynamic ocean currents. Existing enhancements have not adequately addressed these issues in complex marine environments where current patterns vary significantly over time[19].
- Conventional DRL methods, particularly those employing ϵ-greedy exploration strategies, exhibit inefficient exploration-exploitation trade-offs that lead to slow convergence and suboptimal policy learning[25]. These inefficiencies are exacerbated in high-dimensional marine navigation scenarios.
- Current hybrid frameworks (e.g.,[22]) achieve only reward-level integration of APF and DQN, failing to exploit deeper synergies such as state-space augmentation and policy fusion. This shallow integration results in underutilization of physics-based prior knowledge and reduced system stability in dynamic marine environments.

To systematically address these limitations, this study proposes a hybrid path planning framework that integrates an enhanced APF with DQN. The framework introduces four key innovations:

- A multi-scale Transformer architecture is proposed for ocean current prediction, incorporating an adaptive spatio-temporal attention mechanism and physics-informed constraints to achieve high-precision current field forecasting.
- An enhanced adaptive APF (E-APF) is developed by introducing a dynamic current-induced force field and an entropy-driven local minima escape mechanism, thereby improving adaptability and robustness in dynamic marine environments.
- A median-Q-value-based exploration mechanism is introduced to address the inefficient exploration-exploitation trade-off of conventional ϵ-greedy strategies, accelerating convergence during the learning process.
- A multi-level integration framework combining APF and DQN is established through state-space augmentation, an APF-informed loss function, and policy fusion, enhancing both path planning efficiency and environmental adaptability.

Comparison to related work

To position the proposed APF-DQN framework within the existing literature, this section presents a systematic comparison with related studies across four key dimensions: reinforcement learning for path planning, Transformer-based prediction methods, physics-data hybrid approaches, and navigation in dynamic flow fields. Table 1 summarizes the key methodological differences.

Reinforcement learning for path planning

Deep reinforcement learning has emerged as a promising paradigm for sequential decision-making in dynamic environments. Cao et al.[26] applied DQN to dynamic job shop scheduling with Automated Guided Vehicles, demonstrating DQN's effectiveness in handling discrete state spaces and operational constraints. Xu et al.[27] proposed a spatial memory-augmented visual navigation framework using hierarchical deep RL, where explicit memory structures record visited locations to enhance exploration efficiency. Shen et al.[22] pioneered the integration of APF with DQN for AUV path planning, incorporating APF-derived terms into the reward signal to guide learning. Chen et al.[28] addressed the entrapment problem for planetary rovers using Bayesian optimization to find optimal escape action sequences through black-box parameter search. In contrast, APF-DQN achieves multi-level integration through three mechanisms: state-space augmentation with potential field features, an APF-informed loss function for gradient-level guidance, and adaptive policy fusion. The entropy-

Method	Application	Core technique	Dyn.	Key difference from APF-DQN
DQN-AGV[26]	Scheduling	DQN	○	No physics guidance; discrete states
Memory-RL[27]	Navigation	Hierarchical RL	●	Memory-based vs. field-based guidance
APF-DQN-AUV[22]	AUV	DQN + APF reward	●	Reward-level vs. multi-level integration
Bayesian[28]	Rover escape	Bayesian opt.	○	Black-box opt. vs. entropy-based escape
Transformer[29]	Driving	Transformer	●	Agent trajectory vs. flow field prediction
DST2former[30]	Traffic	Spatio-temporal attn.	●	Graph structure vs. continuous field
Attn-RUL[31]	Health	Channel-temporal attn.	○	Health vs. environmental prediction
Physics-NN[32]	Dynamics	Physics + NN	○	Modeling vs. planning application
Twisted-G[33]	Planning	Risk field	●	Risk quantification vs. force guidance
GNSS-Fus[34]	Urban nav.	Adaptive fusion	●	Sensor fusion vs. policy fusion
Airship[37]	Aerial	MPC	●	Model-based vs. learning-based
UAV-Mar[38]	Maritime	Trajectory opt.	●	Comm. objective vs. path objective
Submeso[39]	Ocean	Physical analysis	●	Process study vs. planning application
APF-DQN	USV	Transformer+E-APF+DQN	●	Multi-level physics-RL integration

Table 1. Comparison with related path planning and prediction methods. Dyn. = Dynamic environment support (● = Yes, ○ = No/Partial)

based escape mechanism in E-APF provides a principled approach to local minima detection, differing from the black-box optimization paradigm.

Transformer-based prediction for navigation

The application of Transformer architectures to sequential prediction has demonstrated remarkable success across domains. Chen et al.[29] developed an interaction-aware trajectory prediction framework using Transformer with transfer learning for autonomous driving, predicting future trajectories of surrounding vehicles to enable safe motion planning. Jiang et al.[30] proposed DST2former for traffic flow prediction, utilizing dynamic spatio-temporal attention mechanisms to capture complex dependencies in graph-structured transportation networks. Li et al.[31] applied channel and temporal attention networks for remaining useful life prediction of stratospheric airships, demonstrating the versatility of attention mechanisms in capturing temporal dependencies from multi-sensor data. Unlike these approaches that target discrete agents or health states, APF-DQN predicts continuous ocean current velocity fields. The multi-scale Transformer architecture incorporates physics-informed constraints (mass and vorticity conservation) specific to fluid dynamics, ensuring physically plausible predictions that respect fundamental conservation laws.

Physics-data hybrid methods

The integration of physics-based models with data-driven learning has gained significant attention for improving model interpretability and generalization. Zhang et al.[32] proposed a hybrid framework combining neural networks with physics-based estimators for vehicle longitudinal dynamics modeling, where physics models provide structural priors while neural networks learn residual errors. Wang et al.[33] introduced a twisted Gaussian risk model that encodes vehicle motion states into risk fields for trajectory planning, representing a field-based approach conceptually similar to APF. Liu et al.[34] addressed robust navigation in urban environments through multipath inflation factors for GNSS/IMU/VO fusion, dynamically adjusting measurement confidence based on environmental uncertainty. APF-DQN extends this physics-data philosophy to both prediction and planning: the Transformer incorporates physical conservation laws as soft constraints, while E-APF provides physics-based navigation guidance with adaptive weighting that dynamically adjusts field contributions based on local environmental conditions.

Navigation in dynamic flow fields

Path planning in time-varying flow fields presents unique challenges that distinguish marine and aerial navigation from ground-based robotics. Lolla et al.[35] and Subramani et al.[36] established foundational work on energy-optimal path planning in ocean currents, employing level-set methods and stochastic optimization, respectively. These approaches provide globally optimal solutions but require complete knowledge of the flow field and substantial computational resources. Liu et al.[37] addressed dynamic control of stratospheric airships in time-varying wind fields for communication coverage missions using model predictive control with explicit wind field models. Wang et al.[38] studied UAV trajectory optimization in maritime networks coexisting with satellite systems, addressing trajectory optimization for communication objectives rather than navigation efficiency. Tang et al.[39] investigated submesoscale kinetic energy patterns induced by tropical cyclones, revealing the multi-scale and non-stationary nature of real ocean dynamics that motivates the use of multi-scale prediction architectures. APF-DQN trades global optimality for real-time adaptability, learning policies that generalize across different flow configurations without requiring complete environmental knowledge.

The comparison reveals that while individual components of the proposed framework (Transformer prediction, APF guidance, DQN learning) have been explored separately in various domains, the systematic

multi-level integration—combining physics-constrained prediction, enhanced potential field guidance, and adaptive policy fusion—represents a novel contribution to USV path planning in dynamic ocean environments.

Problem statement

This section presents the theoretical foundation for the proposed path planning framework. The framework requires accurate modeling of both the vehicle and its operating environment, as well as a suitable interface for learning-based control. First, a three-degree-of-freedom (3-DOF) USV dynamics model is introduced to characterize vessel motion under propulsion and external disturbances. Subsequently, a time-varying ocean current model based on potential flow theory is formulated to capture the spatiotemporal dynamics of the marine environment. Finally, to bridge continuous physical dynamics with discrete decision-making, a structured action space is designed for deep reinforcement learning implementation.

USV dynamics model

The planar motion of the USV is modeled using a 3-DOF framework that captures surge, sway, and yaw dynamics[40–42]. Although the motion is constrained to the horizontal plane, the equations are formulated in both the Earth-fixed frame $\{O\text{-}x_o y_o\}$ and the body-fixed frame $\{B\text{-}x_b y_b\}$, as illustrated in Fig. 1. This simplified rigid-body model retains essential motion characteristics while reducing computational complexity by neglecting higher-order hydrodynamic effects—an assumption valid for low-speed operations[43,44].

The kinematic and dynamic equations governing USV motion are derived using the Newton-Euler formulation:

$$\begin{aligned} \dot{\eta} &= R(\psi)\nu, \\ M\dot{\nu} + C(\nu)\nu + D(\nu)\nu &= \tau + \tau_c, \end{aligned} \tag{1}$$

where $\eta = [x, y, \psi]^\top$ denotes the position and heading vector in the Earth-fixed frame, $\nu = [u, v, r]^\top$ represents the velocity vector in the body-fixed frame, M is the mass-inertia matrix, $C(\nu)$ is the Coriolis-centripetal matrix, $D(\nu)$ is the hydrodynamic damping matrix, τ is the control force vector from the propulsion system, and τ_c is the disturbance force vector induced by ocean currents.

The mass-inertia matrix M comprises the rigid-body inertia M_{RB} and the added mass M_A:

$$M = M_{RB} + M_A = \begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & I_z \end{bmatrix} + \begin{bmatrix} -X_{\dot{u}} & 0 & 0 \\ 0 & -Y_{\dot{v}} & -Y_{\dot{r}} \\ 0 & -N_{\dot{v}} & -N_{\dot{r}} \end{bmatrix}, \tag{2}$$

where m is the USV mass, I_z is the moment of inertia about the vertical axis, and $X_{\dot{u}}, Y_{\dot{v}}, Y_{\dot{r}}, N_{\dot{v}}, N_{\dot{r}}$ are added mass coefficients representing the additional inertia due to the surrounding fluid.

The Coriolis-centripetal matrix $C(\nu)$ and the rotation matrix $R(\psi)$ are given by:

$$C(\nu) = \begin{bmatrix} 0 & 0 & -mv \\ 0 & 0 & mu \\ mv & -mu & 0 \end{bmatrix}, \quad R(\psi) = \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

The hydrodynamic damping matrix $D(\nu)$ consists of linear and nonlinear components:

$$D(\nu) = D_l + D_n(\nu). \tag{4}$$

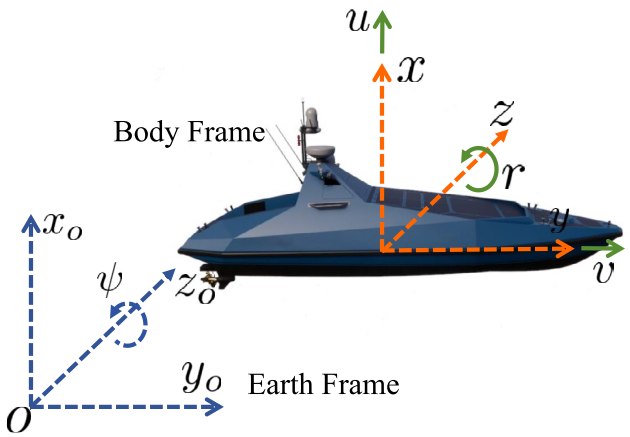

Fig. 1. Coordinate systems for USV modeling: the earth-fixed frame $\{O\text{-}x_o y_o\}$ and the body-fixed frame $\{B\text{-}x_b y_b\}$.

The linear damping D_l dominates at low speeds, while the nonlinear damping $D_n(\nu)$ becomes significant during high-speed maneuvers:

$$D_l = \begin{bmatrix} -X_u & 0 & 0 \\ 0 & -Y_v & 0 \\ 0 & 0 & -N_r \end{bmatrix}, \quad D_n(\nu) = \begin{bmatrix} -X_{|u|u}|u| & 0 & 0 \\ 0 & -Y_{|v|v}|v| & 0 \\ 0 & 0 & -N_{|r|r}|r| \end{bmatrix}. \tag{5}$$

The current-induced disturbance force τ_c is modeled based on the relative velocity between the USV and the ambient flow:

$$\tau_c = -D(\nu_r)\nu_c, \quad \nu_r = \nu + \nu_c, \tag{6}$$

where ν_c denotes the ocean current velocity vector, which is characterized in the following subsection.

Ocean current model

Having established the USV dynamics, this subsection characterizes the ocean current field that constitutes the primary environmental disturbance. A two-dimensional, depth-averaged, time-varying current field model based on potential flow theory is adopted, following the formulations of Lolla et al.[35] and Subramani et al.[36]. This model is particularly suitable for simulating nearshore current dynamics, playing a critical role in USV navigation.

The current field is described by a stream function defined in the Earth-fixed frame:

$$\psi(x, y, t) = A \sin[\pi f(x, t)] \sin(\pi y), \tag{7}$$

where A denotes the peak current velocity amplitude. The spatiotemporal modulation function $f(x, t)$ is given by:

$$f(x, t) = a(t)x^2 + b(t)x, \tag{8}$$

with time-varying coefficients:

$$a(t) = \mu \sin(\omega t), \quad b(t) = 1 - 2\mu \sin(\omega t). \tag{9}$$

Here, ω is the angular frequency corresponding to the tidal period $T = 2\pi/\omega$, and μ governs the vortex spatial distribution. The constraint $|\mu \sin(\omega t)| < 1/2$ is imposed to ensure flow regularity.

Remark 1 The constraint $|\mu \sin(\omega t)| < 1/2$ guarantees that the function $f(x, t)$ remains monotonic in x for $x \in [0, 1]$, thereby preventing the formation of stagnation points and ensuring a well-defined velocity field throughout the domain. This condition is essential for maintaining numerical stability in path planning simulations.

By varying the parameter μ, the model generates different flow configurations, ranging from single-vortex to dual-vortex systems, as illustrated in Fig. 2.

The velocity field is obtained from the stream function through the standard relationship for incompressible two-dimensional flow:

$$\nu_c = \begin{bmatrix} u_c \\ v_c \end{bmatrix} = \begin{bmatrix} -\dfrac{\partial \psi}{\partial y} \\ \dfrac{\partial \psi}{\partial x} \end{bmatrix} = \begin{bmatrix} -A\pi \sin[\pi f(x, t)] \cos(\pi y) \\ A\pi \cos[\pi f(x, t)](2a(t)x + b(t)) \sin(\pi y) \end{bmatrix}, \tag{10}$$

where $\nu_c = [u_c, v_c]^\top$ represents the current velocity vector in the Earth-fixed frame. This formulation has been validated against experimental observations[35] and effectively captures the essential features of nearshore tidal currents for USV path planning applications.

USV action space

Unlike traditional USV control systems that employ continuous input signals, a discrete action space is adopted in this study to facilitate deep reinforcement learning and to reduce computational complexity[24]. As illustrated in Fig. 3, the action space $\mathcal{A}$ comprises eight permissible motion directions—four cardinal (N, E, S, W) and four intercardinal (NE, SE, SW, NW)—uniformly distributed at 45° intervals:

$$\mathcal{A} = \{a_i : \theta_i = i \times 45^\circ,\ i = 0, 1, \ldots, 7\}, \tag{11}$$

where θ_i denotes the heading angle of action a_i measured from the positive x-axis in the Earth-fixed frame. The propulsion velocity magnitude is assumed constant across all actions.

The resultant velocity of the USV in the Earth-fixed frame is obtained by combining the propulsion velocity with the ambient current:

$$\nu_r^E = R(\psi)\nu_b + \nu_c, \tag{12}$$

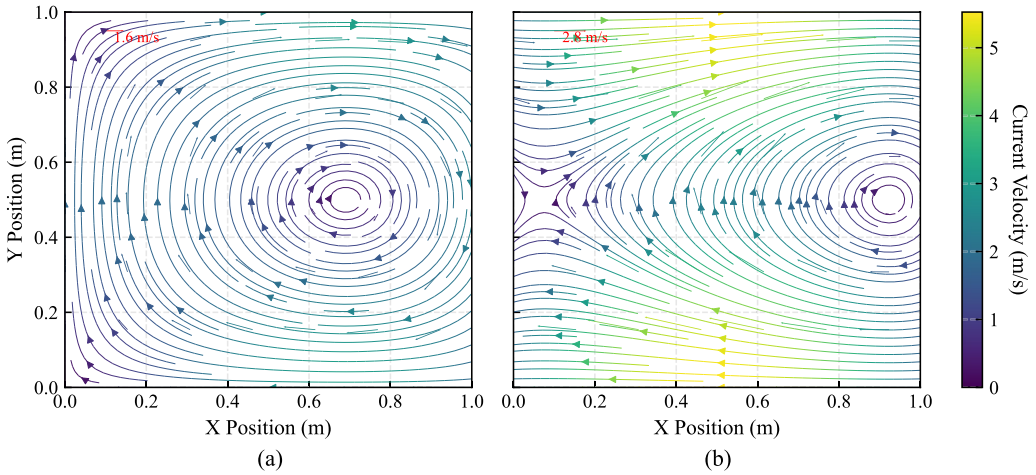

Fig. 2. Ocean current configurations: (**a**) single-vortex system with clockwise rotation; (**b**) dual-vortex system with counter-rotating vortices. Streamlines indicate flow direction, and color intensity corresponds to current speed (m/s).

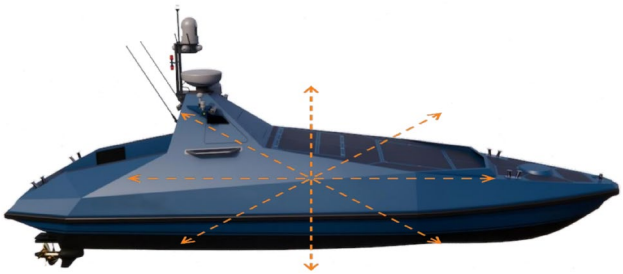

Fig. 3. Schematic illustration of the discrete action space. Arrows indicate the eight permissible motion directions in the horizontal plane, uniformly distributed at 45° intervals.

where ν_b is the propulsion velocity in the body-fixed frame, ν_c is the ocean current velocity defined in Eq. (10), and $R(\psi)$ is the rotation matrix introduced in Equation (1).

The discrete action space formulation offers two key advantages for reinforcement learning. First, it reduces the action space dimensionality from continuous to finite, enabling efficient exploration and accelerating policy convergence. Second, it aligns naturally with value-based methods such as DQN, where Q-values are computed for each discrete action. This formulation is particularly well-suited for dynamic marine environments that demand rapid decision-making.

Method

Building upon the theoretical foundation established in the previous section, this section details the proposed APF-DQN framework for USV path planning in dynamic ocean environments. The framework integrates physics-based guidance from an Enhanced Artificial Potential Field (E-APF) with the learning capability of Deep Q-Networks (DQN), achieving synergy between domain knowledge and data-driven optimization.

The methodology is organized into three main modules encompassing six key components: (1) a multi-dimensional state space representation encoding both local and global environmental information; (2) a multi-scale Transformer architecture for ocean current prediction; (3) an Enhanced APF method incorporating current-induced forces and entropy-based escape mechanisms; (4) a hierarchical reward function balancing task completion, path efficiency, and environmental adaptation; (5) a median Q-value based exploration strategy to improve learning efficiency; and (6) a multi-level integration framework that fuses APF guidance with DQN through state augmentation, loss function design, and policy fusion. These components work synergistically: the Transformer provides predictive environmental awareness, the E-APF offers physics-based navigation guidance, and the DQN learns adaptive policies that leverage both sources of information. The overall architecture is illustrated in Fig. 4.

State space representation

To enable effective path planning in complex and dynamic marine environments, a multi-dimensional state space framework is proposed that integrates spatial, environmental, and temporal information. The state space $\mathcal{S}$ is defined as:

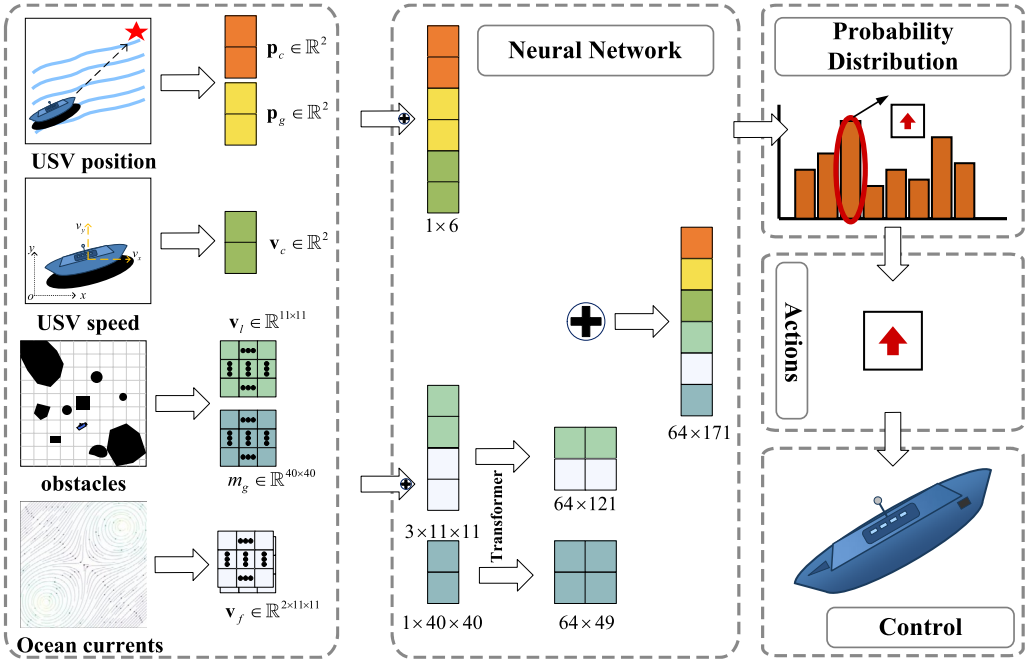

Fig. 4. Overall architecture of the proposed APF-DQN framework. The system comprises three main modules: (1) the Transformer-based ocean current predictor that forecasts local flow fields (Section "Multi-Scale Transformer Model for Ocean Current Prediction"); (2) the Enhanced APF module that computes physics-based navigation guidance (Section "Enhanced Artificial Potential Field"); and (3) the DQN module that learns optimal policies through state representation, reward shaping, exploration optimization, and policy fusion (Sections "State Space Representation", "Reward Function", "Deep Q-Network with Median-Based Exploration", and "APF-Guided Deep Q-Network Integration"). Arrows indicate information flow between components.

$$\mathcal{S} = \{\mathrm{p}_c, \mathrm{p}_g, \dot{\mathrm{p}}_c, \mathrm{V}_l, \mathrm{M}_g, \mathrm{V}_f, \mathcal{H}_t\}, \tag{13}$$

where the components are organized into three functional categories as detailed below.

Navigation Reference: This category encodes the USV's kinematic state relative to the mission objective:

- $\mathrm{p}_c, \mathrm{p}_g \in [-1, 1]^2$: the normalized current and goal positions in the Earth-fixed frame.
- $\dot{\mathrm{p}}_c \in \mathbb{R}^2$: the instantaneous velocity of the USV in Cartesian coordinates.

Remark 2 Position coordinates are normalized to the range $[-1, 1]$ using:

$$f_{\mathrm{norm}}(x, y) = 2 \cdot \left(\frac{x}{W}, \frac{y}{H}\right) - 1,$$

where W and H denote the global map dimensions. This normalization ensures numerical stability and consistent scaling across different operational scenarios.

Environmental Perception: This category provides spatial awareness of obstacles and ocean currents:

- $\mathrm{M}_g \in \{0, 1\}^{W \times H}$: the global obstacle map, where 1 indicates an obstacle cell and 0 indicates free space.
- $\mathrm{V}_l \in \{0, 1\}^{n \times n}$: the local obstacle observation extracted from M_g as an $n \times n$ window centered at p_c:

$$\mathrm{V}_l = \mathrm{M}_g \left[x_c - \frac{n-1}{2} : x_c + \frac{n-1}{2},\ y_c - \frac{n-1}{2} : y_c + \frac{n-1}{2}\right].$$

- $\mathrm{V}_f \in \mathbb{R}^{2 \times n \times n}$: the predicted ocean current velocity field within the local observation window, with two channels corresponding to the x- and y-velocity components. This field is obtained through the ocean current prediction model described in the following subsection.**Temporal Context:** A historical buffer $\mathcal{H}_t$ maintains temporal information essential for ocean current prediction:

$$\mathcal{H}_t = \{\mathrm{p}_{t-T:t},\ \mathrm{a}_{t-T:t},\ \nu_{c,t-T:t}\},$$

where:

- $\mathrm{p}_{t-T:t}$: the position trajectory over the past T time steps.

- $\mathrm{a}_{t-T:t}$: the action sequence over the past T time steps.
- $\nu_{c,t-T:t}$: the observed ocean current velocity history, enabling temporal pattern recognition for flow prediction.

Table 2 summarizes the state space parameters, determined through preliminary experiments to balance computational efficiency and perceptual coverage.

Multi-scale transformer model for ocean current prediction

The ocean current field exhibits significant spatiotemporal variability, posing challenges for USV path planning. Traditional sensing equipment, such as Acoustic Doppler Current Profilers, provides only single-point instantaneous velocity measurements[45], limiting the USV's ability to perceive the surrounding flow distribution. To address this limitation, a multi-scale Transformer architecture that incorporates physical constraints is proposed for ocean current prediction. This model integrates hierarchical attention mechanisms with fundamental conservation principles, enabling high-resolution local flow field forecasting.

Architecture overview

The prediction model follows an encoder-decoder structure. The global encoder processes the complete ocean current field to extract macroscopic flow patterns, while the local decoder generates high-resolution predictions within a neighborhood of the USV's current position. The global ocean current field $\mathrm{V}_{\mathrm{global}} \in \mathbb{R}^{2\times W\times H}$, generated by the potential flow model defined in the Problem Statement section, serves as the primary input.

The global Transformer encoder extracts flow features as:

$$\mathrm{F}_{\mathrm{global}} = \mathrm{Transformer}_{\mathrm{global}}\left(\mathrm{V}_{\mathrm{global}}, \mathcal{H}_t\right), \tag{14}$$

where $\mathcal{H}_t$ denotes the historical observation buffer and $\mathrm{F}_{\mathrm{global}} \in \mathbb{R}^d$ is the resulting global feature vector of dimension d.

The local decoder predicts the ocean current velocity field V_f conditioned on the global features and the USV's current position:

$$\mathrm{V}_f = \mathrm{Transformer}_{\mathrm{local}}\left(\mathrm{F}_{\mathrm{global}}, \mathrm{p}_c, n\right), \tag{15}$$

where n specifies the local prediction window size.

Feature representation

Multi-scale features are constructed through concatenation:

$$\begin{aligned} \mathrm{S}_t &= \left[\nu_c(t);\ \mathrm{V}_{\mathrm{local},t};\ \mathrm{F}_{\mathrm{global}};\ \sigma(\nu_c(t))\right], \\ \mathrm{X}_t &= \left[\mathrm{p}_t;\ \mathrm{a}_t;\ \mathrm{S}_t;\ \mathrm{PE}_{\mathrm{spatial}}(\mathrm{p}_t);\ \mathrm{TE}(t)\right], \end{aligned} \tag{16}$$

where $\nu_c(t)$ denotes the ocean current velocity at time t, $\mathrm{V}_{\mathrm{local},t}$ represents historical local observations, $\sigma(\cdot)$ computes velocity statistics (mean and variance), and $\mathrm{PE}_{\mathrm{spatial}}$ and TE are spatial and temporal positional encodings, respectively.

Spatiotemporal attention mechanism

A spatiotemporally aware attention mechanism integrates content similarity with spatial-temporal proximity:

$$\alpha_{t,s}^{(i)} = \frac{\exp\left(\dfrac{\mathrm{Q}_t^{(i)}(\mathrm{K}_s^{(i)})^\top}{\sqrt{d_k}} - \lambda_1^{(i)} d_{ts} - \lambda_2^{(i)}\|\mathrm{p}_t - \mathrm{p}_s\|\right)}{\sum_{s'}\exp\left(\dfrac{\mathrm{Q}_t^{(i)}(\mathrm{K}_{s'}^{(i)})^\top}{\sqrt{d_k}} - \lambda_1^{(i)} d_{ts'} - \lambda_2^{(i)}\|\mathrm{p}_t - \mathrm{p}_{s'}\|\right)}, \tag{17}$$

Parameter	Description	Dimension/Value
n	Local observation window size	11
T	Temporal history length	10
W, H	Global map dimensions	40×40
$\mathrm{p}_c, \mathrm{p}_g$	Normalized positions	$[-1, 1]^2$
$\dot{\mathrm{p}}_c$	USV velocity	$\mathbb{R}^2$
V_l	Local obstacle map	$\{0, 1\}^{11\times 11}$
M_g	Global obstacle map	$\{0, 1\}^{40\times 40}$
V_f	Predicted current field	$\mathbb{R}^{2\times 11\times 11}$

Table 2. State space parameter configurations.

where $d_{ts} = |t - s|$ denotes the temporal distance, $\|\mathrm{p}_t - \mathrm{p}_s\|$ denotes the spatial distance, and $\lambda_1^{(i)}$, $\lambda_2^{(i)}$ are learnable parameters controlling the influence of temporal and spatial proximity in attention head *i*.

Physics-informed loss function

The loss function incorporates four physically motivated constraints to ensure prediction validity:

$$\begin{aligned}\mathcal{L} = & \underbrace{\|\hat{\mathrm{V}}_f - \mathrm{V}_{\mathrm{local}}\|_2^2}_{\text{Prediction error}} + \lambda_1 \underbrace{\|\nabla \cdot \hat{\mathrm{V}}_f\|_2^2}_{\text{Mass conservation}} \\ & + \lambda_2 \underbrace{\|\nabla \times \hat{\mathrm{V}}_f - \boldsymbol{\omega}\|_2^2}_{\text{Vorticity conservation}} + \lambda_3 \underbrace{\|\hat{\mathrm{V}}_f(t) - \hat{\mathrm{V}}_f(t-1)\|_1}_{\text{Temporal continuity}} \\ & + \lambda_4 \underbrace{\mathcal{L}_{\mathrm{KL}}(\hat{\mathrm{V}}_f \| \mathrm{V}_{\mathrm{global}})}_{\text{Statistical consistency}},\end{aligned} \tag{18}$$

where $\hat{\mathrm{V}}_f$ denotes the predicted velocity field, $\mathrm{V}_{\mathrm{local}}$ denotes the ground truth, $\nabla\cdot$ and $\nabla\times$ represent the divergence and curl operators, $\boldsymbol{\omega}$ is the observed vorticity, and $\mathcal{L}_{\mathrm{KL}}$ is the Kullback-Leibler divergence.

Physical Constraint Implementation. The physical constraints are enforced through soft penalty terms rather than hard architectural constraints, offering three advantages: (1) flexibility to accommodate measurement noise and model uncertainties; (2) end-to-end differentiability for gradient-based optimization; and (3) tunable trade-offs between prediction accuracy and physical consistency.

The hyperparameters are determined through grid search on a validation set:

- **Mass conservation** ($\lambda_1 = 0.1$): Enforces $\nabla \cdot \mathrm{V} \approx 0$ for incompressible flow.
- **Vorticity conservation** ($\lambda_2 = 0.1$): Maintains consistency with observed rotational flow patterns.
- **Temporal continuity** ($\lambda_3 = 0.05$): Ensures smooth temporal evolution via L1 regularization.
- **Statistical consistency** ($\lambda_4 = 0.01$): Aligns local predictions with global flow statistics.

The effectiveness of these constraints is validated through ablation experiments in Section "Ablation Study on Transformer Physical Constraints".

The complete prediction procedure is summarized in Algorithm 1.

Require: Global field $\mathbf{V}_{\text{global}}$, history buffer $\mathscr{H}_t$, position $\mathbf{p}_c$, window size n
Ensure: Predicted local field $\mathbf{V}_f$

/* *Phase 1: Global Encoding* */
$\mathbf{X}_{\text{global}} \leftarrow \text{Concat}(\mathbf{V}_{\text{global}}, \mathscr{H}_t, \mathbf{PE}_{\text{spatial}}, \mathbf{TE})$
for $l = 1$ **to** L_{global} **do**
  $\boldsymbol{\alpha}^{(i)} \leftarrow \text{SpatioTemporalAttention}(\mathbf{X}_{\text{global}})$ ▷ Eq. (4)
  $\mathbf{X}_{\text{global}} \leftarrow \text{FFN}(\text{MultiHead}(\mathbf{X}_{\text{global}}, \boldsymbol{\alpha}^{(i)}))$
end for
$\mathbf{F}_{\text{global}} \leftarrow \mathbf{X}_{\text{global}}$

/* *Phase 2: Local Decoding* */
$\mathbf{R}_c \leftarrow \text{ExtractRegion}(\mathbf{p}_c, n)$ ▷ Local region centered at $\mathbf{p}_c$
$\mathbf{S}_t, \mathbf{X}_t \leftarrow \text{ConstructFeatures}(\mathbf{R}_c, \mathbf{F}_{\text{global}})$ ▷ Eq. (3)
$\mathbf{X}_{\text{local}} \leftarrow \mathbf{X}_t$
for $l = 1$ **to** L_{local} **do**
  $\mathbf{X}_{\text{local}} \leftarrow \text{TransformerLayer}(\mathbf{X}_{\text{local}}, \mathbf{F}_{\text{global}})$
end for
$\hat{\mathbf{V}}_f \leftarrow \text{Decoder}(\mathbf{X}_{\text{local}})$

/* *Training Phase* */
Minimize $\mathscr{L}$ as defined in Eq. (18)

return $\hat{\mathbf{V}}_f$

Algorithm 1. Multi-scale transformer-based ocean current prediction.

The predicted ocean current field V_f serves as a critical input for both the state representation and the force field computation. The following subsection describes how this information is incorporated into an Enhanced Artificial Potential Field method.

Enhanced artificial potential field

Traditional Artificial Potential Field (APF) methods are susceptible to local minima and exhibit limited adaptability to dynamic environments. To address these limitations, an Enhanced APF (E-APF) is proposed that incorporates three key mechanisms: adaptive force weighting, ocean current integration, and entropy-based escape from local minima.

Composite force field

The total force acting on the USV is computed as a weighted combination of three components:

$$\mathrm{F}_{\mathrm{total}} = w_{\mathrm{att}}\mathrm{F}_{\mathrm{att}} + w_{\mathrm{rep}}\mathrm{F}_{\mathrm{rep}} + w_{\mathrm{flow}}\mathrm{F}_{\mathrm{flow}}, \tag{19}$$

where $\mathrm{F}_{\mathrm{att}}$, $\mathrm{F}_{\mathrm{rep}}$, and $\mathrm{F}_{\mathrm{flow}}$ denote the attractive, repulsive, and current-induced force components, respectively.

The attractive force directs the USV toward the goal:

$$\mathrm{F}_{\mathrm{att}} = k_{\mathrm{att}}(\mathrm{p}_g - \mathrm{p}_c), \tag{20}$$

where k_{att} is the attraction gain coefficient, p_g is the goal position, and p_c is the current USV position.

The repulsive force prevents collisions with obstacles:

$$\mathrm{F}_{\mathrm{rep}} = \sum_{i=1}^{N_{\mathrm{obs}}} k_{\mathrm{rep}} \left(\frac{1}{\|\mathrm{p}_c - \mathrm{p}_{\mathrm{obs}}^i\|} - \frac{1}{d_{\mathrm{eff}}} \right) \frac{\hat{\mathrm{n}}_i}{\|\mathrm{p}_c - \mathrm{p}_{\mathrm{obs}}^i\|^2}, \tag{21}$$

where $\mathrm{p}_{\mathrm{obs}}^i$ denotes the position of the i-th obstacle, k_{rep} is the repulsion gain, and $\hat{\mathrm{n}}_i = (\mathrm{p}_c - \mathrm{p}_{\mathrm{obs}}^i)/\|\mathrm{p}_c - \mathrm{p}_{\mathrm{obs}}^i\|$ is the unit vector pointing from the obstacle to the USV. The effective repulsion range d_{eff} is adaptively reduced as the USV approaches the goal:

$$d_{\mathrm{eff}} = \min\left(d_0,\ \lambda\|\mathrm{p}_g - \mathrm{p}_c\|\right), \tag{22}$$

where d_0 is the maximum repulsion range and λ is a scaling factor. This adaptation prevents excessive repulsion from interfering with the approach to the goal.

The current-induced force incorporates ocean current effects:

$$\mathrm{F}_{\mathrm{flow}} = k_f \nu_c(\mathrm{p}_c), \tag{23}$$

where k_f is the flow gain coefficient and $\nu_c(\mathrm{p}_c)$ denotes the ocean current velocity at the USV's position.

Adaptive weight adjustment

The force weights are dynamically adjusted based on environmental conditions:

$$w_{\mathrm{att}} = \frac{e^{-\alpha d_{\min}}}{Z}, \quad w_{\mathrm{rep}} = \frac{1 - e^{-\beta N_{\mathrm{near}}}}{Z}, \quad w_{\mathrm{flow}} = \frac{\|\nu_c\|}{Z}, \tag{24}$$

where $d_{\min}$ denotes the distance to the nearest obstacle, N_{near} denotes the number of obstacles within a sensing radius, α and β are sensitivity parameters, and Z is a normalization factor ensuring $w_{\mathrm{att}} + w_{\mathrm{rep}} + w_{\mathrm{flow}} = 1$.

Temporal smoothing

To suppress trajectory oscillations caused by rapid force variations, temporal smoothing is applied:

$$\mathrm{F}_t = \eta\mathrm{F}_{\mathrm{total}} + (1 - \eta)\mathrm{F}_{t-1}, \tag{25}$$

where $\eta \in (0, 1)$ is the smoothing coefficient that blends the current force with the previous timestep's force.

Probabilistic action selection

Rather than directly following the force vector, E-APF employs probabilistic action selection using a temperature-scaled softmax function. For each candidate action a_i in the discrete action space, a score is computed:

$$S(\mathrm{a}_i) = \omega_1 \cos\theta_{Fi} + \omega_2/d_{\mathrm{proj},i}, \tag{26}$$

where θ_{Fi} denotes the angle between action a_i and the smoothed force F_t, $d_{\mathrm{proj},i}$ denotes the projected distance to the goal along action a_i, and ω_1, ω_2 are weighting factors.

The action probability is then computed as:

$$P(\mathrm{a}_i) = \frac{\exp(S(\mathrm{a}_i)/T)}{\sum_{j=1}^{N_a} \exp(S(\mathrm{a}_j)/T)}, \tag{27}$$

where the temperature $T = T_0 e^{-t/\tau}$ decreases exponentially over time, with initial temperature T_0 and decay constant τ. This annealing schedule transitions from exploration (high T) to exploitation (low T).

Entropy-based escape mechanism

Local minima occur when attractive and repulsive forces approximately cancel, leaving no clear preferred direction. This situation manifests as high entropy in the action probability distribution. The entropy is computed as:

$$H(\mathcal{A}) = -\sum_{i=1}^{N_a} P(\mathrm{a}_i) \log P(\mathrm{a}_i). \tag{28}$$

When $H(\mathcal{A}) > H_{\mathrm{threshold}}$, indicating potential entrapment, the algorithm temporarily selects random actions from a safe action set $\mathcal{A}_{\mathrm{safe}}$(actions that do not lead to immediate collision) to escape the local minimum.

The complete E-APF procedure is summarized in Algorithm 2.

Require: Start $\mathbf{p}_s$, goal $\mathbf{p}_g$, current field $\nu_c(\cdot)$, obstacles $\mathscr{O}$, step size Δt
Ensure: Path $\mathscr{P}$

```
/* Initialization */
p_c ← p_s;   P ← {p_c};   F_{t-1} ← 0
/* Main Loop */
while ‖p_g − p_c‖ > ε do
    // Force computation
    d_min, N_near ← ObstacleAnalysis(O, p_c)
    w_att, w_rep, w_flow ← AdaptiveWeights(d_min, N_near)
    F_att, F_rep, F_flow ← ComputeForces(p_c, p_g, O, ν_c)
    F_total ← w_att F_att + w_rep F_rep + w_flow F_flow
    F_t ← η F_total + (1 − η) F_{t-1}                          ▷ Temporal smoothing
    // Entropy-based action selection
    S(a_i), P(a_i) ← ActionScores(F_t, A)
    H(A) ← −Σ_i P(a_i) log P(a_i)
    if H(A) > H_threshold then
        a* ← RandomSelect(A_safe)                              ▷ Escape local minima
    else
        a* ← arg max_i P(a_i)
    end if
    // State update
    p_c ← p_c + (v_{a*} + ν_c(p_c)) · Δt
    P ← P ∪ {p_c};   F_{t-1} ← F_t
end while
return P
```

Algorithm 2. Enhanced artificial potential field path planning.

While E-APF provides physics-based navigation guidance, the DQN component requires a carefully designed reward function to learn effective policies. The following subsection presents the hierarchical reward framework.

Reward function

The reward function is a critical component that directly shapes the learned navigation policy. A hierarchical reward framework is proposed, comprising four modules: task completion, path efficiency, environmental adaptation, and exploration. The total reward at each time step is:

$$R_{\mathrm{total}} = \sum_{i=1}^{4} w_i \cdot r_i(s_t, a_t, s_{t+1}), \tag{29}$$

where w_i denotes the weight for the i-th reward component.

Task completion reward
The task-oriented reward provides sparse signals for goal achievement and collision avoidance:

$$r_{\text{task}} = \begin{cases} R_{\text{goal}}, & \text{if } \|\mathrm{p}_c - \mathrm{p}_g\| \le \epsilon \\ -R_{\text{collision}}, & \text{if collision detected} \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

where R_{goal} denotes the goal achievement reward, $R_{\text{collision}}$ denotes the collision penalty, and ϵ is the goal proximity threshold.

Path efficiency reward
Dense rewards encourage progress toward the goal and trajectory smoothness:

$$\begin{aligned} r_{\text{dist}} &= \alpha(d_t - d_{t+1}), \\ r_{\text{smooth}} &= -\beta_1|\theta_t - \theta_{t-1}| - \beta_2\|\dot{\mathrm{p}}_t - \dot{\mathrm{p}}_{t-1}\|, \end{aligned} \tag{31}$$

where $d_t = \|\mathrm{p}_c - \mathrm{p}_g\|$ denotes the distance to the goal at time t, θ_t denotes the heading angle, and $\dot{\mathrm{p}}_t$ denotes the USV velocity. The distance reduction reward r_{dist} encourages goal approach, while r_{smooth} penalizes abrupt heading and velocity changes.

Environmental adaptation reward
This module encourages current utilization and obstacle avoidance:

$$\begin{aligned} r_{\text{curr}} &= \gamma \max(0, \cos\theta_{\text{ac}}) \cdot (1 + \tanh\|\nu_c\|), \\ r_{\text{safe}} &= \delta_1 \frac{DT(\mathrm{p}_c)}{\max(DT)} - \delta_2 \exp\left(-\frac{\min_i \|\mathrm{p}_c - \mathrm{p}_{\text{obs}}^i\|}{d_{\text{safe}}}\right), \end{aligned} \tag{32}$$

where θ_{ac} denotes the angle between the USV heading and the current direction, ν_c denotes the ocean current velocity, $DT(\mathrm{p}_c)$ denotes the distance transform value at the current position (larger values indicate greater distance from obstacles), and d_{safe} is a safety distance parameter. The current alignment reward r_{curr} encourages energy-efficient navigation by utilizing favorable currents, while r_{safe} rewards maintaining safe distances from obstacles.

Exploration and energy reward
This module balances exploration incentives with energy consumption penalties:

$$\begin{aligned} r_{\text{exp}} &= \eta_1 \mathbb{I}(\mathrm{p}_c \notin \mathcal{T}) + \eta_2 \exp(-\|\nabla\Phi_{\text{exp}}\|), \\ r_{\text{energy}} &= -\lambda\|\dot{\mathrm{p}}_t - \nu_c(\mathrm{p}_c)\|, \end{aligned} \tag{33}$$

where $\mathcal{T}$ denotes the set of previously visited positions, $\mathbb{I}(\cdot)$ denotes the indicator function, Φ_{exp} denotes the exploration potential field encoding spatial coverage information, and $\|\dot{\mathrm{p}}_t - \nu_c\|$ represents the relative velocity magnitude (proportional to propulsion effort). The first term of r_{exp} rewards visiting new areas, while the second term encourages movement toward regions with lower exploration potential gradients. The energy penalty r_{energy} discourages excessive propulsion effort.

Table 3 summarizes the reward weights and parameter values.

With the reward function defined, the next subsection presents the enhanced DQN architecture that learns navigation policies through interaction with the environment.

Category	Component	Weight	Parameter value
Task	Goal achievement (r_{task})	1.0	$R_{\text{goal}} = 500$
	Collision penalty	–	$R_{\text{collision}} = 10$
Efficiency	Distance reduction (r_{dist})	2.0	–
	Trajectory smoothness (r_{smooth})	3.0	–
Environment	Current alignment (r_{curr})	1.5	–
	Safety margin (r_{safe})	4.0	–
Exploration	Area coverage (r_{exp})	1.0	–
	Energy penalty (r_{energy})	2.0	–

Table 3. Reward function weights and parameters.

Deep Q-network with median-based exploration

This study adopts an enhanced Deep Q-Network (DQN) architecture to improve decision-making capabilities in complex marine environments. The proposed algorithm incorporates four key mechanisms: Dueling network structure, prioritized experience replay, adaptive soft update, and a novel median Q-value based exploration strategy.

Dueling network architecture

The Dueling DQN decomposes the state-action value function $Q(s, a)$ into state value and action advantage components[46]:

$$Q(s,a) = V(s) + \left(A(s,a) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a') \right), \tag{34}$$

where $V(s)$ represents the state value function and $A(s, a)$ denotes the action advantage function. This decomposition enables independent learning of state values and action advantages, resulting in more accurate and robust value estimation.

Prioritized experience replay

To enhance sample efficiency, prioritized experience replay[47] is employed. For each transition (s_t, a_t, r_t, s_{t+1}), the sampling probability is determined by the TD error:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha}, \quad p_i = |\delta_i| + \varepsilon, \tag{35}$$

where the TD error is computed as:

$$\delta_i = r_i + \gamma \max_{a'} Q(s_{i+1}, a'; \theta') - Q(s_i, a_i; \theta), \tag{36}$$

α controls the degree of prioritization, and ε is a small constant ensuring non-zero sampling probabilities.

Importance sampling weights correct for the bias introduced by non-uniform sampling:

$$w_i = \frac{(N \cdot P(i))^{-\beta}}{\max_j w_j}, \tag{37}$$

where N denotes the replay buffer size and β gradually increases from an initial value to 1 during training, progressively correcting the sampling bias.

The network is optimized using a weighted TD loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[w_i \left(r + \gamma \max_{a'} Q(s', a'; \theta') - Q(s, a; \theta) \right)^2 \right]. \tag{38}$$

Adaptive soft update

The target network update rate is dynamically adjusted based on learning progress[48]:

$$\tau = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \text{sigmoid}(\overline{|\delta|}), \tag{39}$$

where $\overline{|\delta|}$ denotes the mean absolute TD error of the current batch. The target network parameters are updated as:

$$\theta' \leftarrow \tau\theta + (1 - \tau)\theta'. \tag{40}$$

This adaptive mechanism increases the update rate when TD errors are large (indicating significant learning potential) and decreases it when errors are small (prioritizing stability).

Median Q-value based exploration

The standard ϵ-greedy strategy selects random actions uniformly from the entire action space:

$$a_t = \begin{cases} \text{random}(\mathcal{A}), & \text{if } \xi < \epsilon \\ \arg\max_a Q(s_t, a), & \text{otherwise} \end{cases} \tag{41}$$

Although simple, this approach wastes exploration resources on clearly suboptimal actions.

To address this limitation, a median Q-value based exploration mechanism is proposed:

$$a_t = \begin{cases} \text{random}(\mathcal{A}_{\text{mid}}), & \text{if } \xi < \epsilon \\ \arg\max_a Q(s_t, a), & \text{otherwise} \end{cases} \tag{42}$$

where $\mathcal{A}_{\text{mid}}$ contains actions with Q-values in the interquartile range:

$$\mathcal{A}_{\text{mid}} = \{a_i \mid Q_{25\%}(s_t) \leq Q(s_t, a_i) \leq Q_{75\%}(s_t)\}. \tag{43}$$

Here, $Q_{25\%}(s_t)$ and $Q_{75\%}(s_t)$ denote the 25th and 75th percentiles of Q-values across all actions in state s_t.

Theoretical Motivation. The median-Q exploration strategy is motivated by three observations specific to USV path planning:

1. **Structured Q-value Distribution**: In the 8-direction discrete action space, Q-values exhibit characteristic patterns. High-Q actions (top 25%) correspond to goal-directed movements aligned with favorable currents, which are already exploited during greedy selection. Low-Q actions (bottom 25%) represent clearly suboptimal choices where exploration provides minimal learning value.
2. **Information Gain Efficiency**: Medium-Q actions (25%–75%) represent the region of highest uncertainty where exploration yields maximum information gain. Standard ϵ-greedy allocates exploration uniformly, including low-Q actions unlikely to improve policy quality. Median-Q exploration redirects samples to actions with higher potential for value refinement.
3. **Accelerated Convergence**: By excluding extreme Q-values from exploration, the strategy maintains meaningful exploration of promising alternatives while avoiding detrimental actions, as demonstrated in Section "Analysis of the Effectiveness of the Median-based Exploration Strategy".**Potential Bias Mitigation.** Three mechanisms address the theoretical concern that median-Q exploration may bias against initially underestimated optimal actions:

- **Q-value Dynamics**: Underestimated optimal actions receive positive TD updates when selected via the greedy component ($1 - \epsilon$ probability), gradually moving into the explorable range.
- **APF Guidance**: The E-APF component provides physics-based action preferences independent of Q-values, offering an alternative pathway for discovering optimal actions through the policy fusion mechanism.
- **Reward Structure**: The hierarchical reward function ensures that beneficial actions (goal-directed, current-aligned, obstacle-avoiding) quickly accumulate positive returns, preventing them from remaining in the bottom 25% for extended periods.

APF-guided deep Q-network integration

Having introduced the E-APF method and the enhanced DQN algorithm separately, this subsection describes how these components are integrated into a unified framework. A multi-level integration approach is established through three complementary mechanisms: state space augmentation, APF-guided loss function, and policy fusion.

State space augmentation

The state representation is augmented with potential field information to enhance environmental awareness:

$$s_t^{\text{aug}} = [s_t;\ \mathrm{F}_{\text{total}};\ \nabla \mathrm{F}_{\text{total}}], \tag{44}$$

where $\mathrm{F}_{\text{total}}$ denotes the total force vector computed by E-APF at the current position, and $\nabla \mathrm{F}_{\text{total}}$ denotes its spatial gradient. This augmentation provides the DQN with explicit information about the force field topology, enabling physics-informed policy learning.

APF-guided loss function

The Q-network training is regularized to align with E-APF guidance through a composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DQN}} + \lambda_{\text{APF}} \left\| \mathrm{q}(s) - \alpha \mathrm{F}_{\text{total}} \right\|_2^2, \tag{45}$$

where $\mathcal{L}_{\text{DQN}}$ denotes the standard TD loss, $\mathrm{q}(s) = [Q(s, a_1), \ldots, Q(s, a_{|\mathcal{A}|})]^\top$ denotes the Q-value vector across all actions. This term encourages the Q-value landscape to reflect the structure of the potential field, with higher Q-values for actions aligned with the E-APF force direction.

Design Rationale. The L_2 norm is chosen for its convexity and differentiability, which provide stable gradient signals that encourage the learned Q-value distribution to align with the physics-based potential field direction. This formulation effectively embeds physical prior knowledge into the reinforcement learning process, accelerating early convergence.

Hyperparameter Selection. The scaling factor $\alpha = 0.1$ was determined empirically to align the magnitude of force vectors ($|\mathrm{F}_{\text{total}}| \approx 1$–$10$) with Q-value estimates ($|Q| \approx 100$–400). The regularization weight $\lambda_{\text{APF}} = 0.01$ was selected to balance early-stage physics guidance with later-stage reward-driven learning, ensuring that the APF term provides meaningful guidance without dominating the TD loss.

Policy fusion

The final action distribution is obtained by blending DQN and E-APF policies:

$$P(a_i) = (1 - w_{\text{APF}}) P_{\text{DQN}}(a_i) + w_{\text{APF}} P_{\text{E-APF}}(a_i), \tag{46}$$

where $P_{\mathrm{DQN}}(a_i) = \mathrm{softmax}(Q(s, a_i)/T_Q)$ is derived from Q-values, and $P_{\text{E-APF}}(a_i) = \mathrm{softmax}(S(a_i)/T_F)$ is derived from E-APF action scores defined in the E-APF subsection. The temperatures T_Q and T_F control the sharpness of each distribution.

The fusion weight w_{APF} adapts based on training progress and policy confidence:

$$w_{\mathrm{APF}} = \max\left(0, 1 - \frac{t}{T_{\mathrm{decay}}}\right) \cdot \exp\left(-\frac{\mathrm{Var(q}(s))}{\sigma_w}\right), \tag{47}$$

where t denotes the current training step, T_{decay} denotes the decay horizon, $\mathrm{Var(q}(s))$ denotes the variance of Q-values across actions (indicating confidence), and σ_w is a sensitivity parameter.

Design Rationale. The adaptive weight schedule incorporates two complementary principles:

1. **Temporal Decay**: The term $\max(0, 1 - t/T_{\mathrm{decay}})$ ensures that E-APF guidance dominates during early training when Q-estimates are unreliable, then gradually transfers control to the learned DQN policy as training progresses.
2. **Confidence-Based Adjustment**: The term $\exp(-\mathrm{Var(q)}/\sigma_w)$ reduces APF influence when Q-value variance is high (indicating confident action preferences) and increases it when variance is low (indicating uncertainty). This allows the DQN to override APF guidance in states where it has learned reliable value estimates.

The effectiveness of this integration is validated in Section "APF-DQN Integration Framework", where APF-DQN achieves 31.4% faster convergence compared to standalone DQN (42.67 ± 1.37 vs. 62.25 ± 1.14 episodes, $p < 0.001$).

The complete APF-guided DQN algorithm is summarized in Algorithm 3.

Experiments

This section presents a comprehensive experimental evaluation of the proposed APF-DQN framework. The experimental setup is first described, including the simulation environment, comparative algorithms, evaluation metrics, and reproducibility measures. The main experimental results are then presented, followed by ablation studies that analyze the contribution of individual components. Finally, the robustness and generalization capability of APF-DQN are evaluated, and its limitations are discussed.

Require: E-APF force field $\mathbf{F}$, replay buffer $\mathscr{D}$, network parameters θ, θ'
Ensure: Trained policy Q^*

/ Initialization */*
$\theta' \leftarrow \theta$; $w_{\text{APF}} \leftarrow 1$

for episode $= 1$ **to** M **do**
  Observe initial state s_0
  $s_0^{\text{aug}} \leftarrow \text{AugmentState}(s_0, \mathbf{F})$ ▷ State augmentation

  while not terminal **do**
    // Policy Fusion
    $P_{\text{DQN}}(a_i) \leftarrow \text{softmax}(Q(s_t, a_i)/T_Q)$
    $P_{\text{E-APF}}(a_i) \leftarrow \text{softmax}(S(a_i)/T_F)$
    $P(a_i) \leftarrow (1 - w_{\text{APF}})P_{\text{DQN}}(a_i) + w_{\text{APF}}P_{\text{E-APF}}(a_i)$

    // Median-Q Exploration
    if $\xi < \varepsilon$ **then**
      $\mathscr{A}_{\text{mid}} \leftarrow \{a_i \mid Q_{25\%} \le Q(s_t, a_i) \le Q_{75\%}\}$
      $a_t \sim \text{Uniform}(\mathscr{A}_{\text{mid}})$ ▷ Explore medium-Q actions
    else
      $a_t \sim P(a)$ ▷ Sample from fused policy
    end if

    // Environment Interaction
    Execute a_t, observe r_t, s_{t+1}
    Store $(s_t^{\text{aug}}, a_t, r_t, s_{t+1}^{\text{aug}})$ in $\mathscr{D}$ with priority $p_i = |\delta_i| + \varepsilon$

    // Network Update
    if update interval reached **then**
      Sample minibatch from $\mathscr{D}$ with importance weights w_j
      $\mathscr{L} \leftarrow \mathscr{L}_{\text{DQN}} + \lambda_{\text{APF}}\|\mathbf{q}(s) - \alpha\mathbf{F}\|^2$ ▷ APF-informed loss
      Update θ via gradient descent on $\mathscr{L}$
      $\theta' \leftarrow \tau\theta + (1 - \tau)\theta'$ ▷ Soft update
      Update w_{APF} using Eq. (47)
    end if
  end while
end for

Algorithm 3. APF-guided deep Q-network algorithm.

Experimental setup

Simulation environment

A 40×40 two-dimensional grid environment is constructed to simulate realistic marine navigation conditions for USV path planning. The simulation framework comprises three core components:

- **Obstacle Configuration:** The environment contains multiple rectangular obstacles and a central circular obstacle, collectively occupying 25% of the total area. This configuration mimics common navigational hazards encountered in maritime operations.
- **Ocean Current Field:** A dual-vortex current system is generated through the superposition of two counter-rotating vortex fields, producing a dynamic and irregular flow regime with peak velocities reaching 5.0 m/s.
- **Navigation Task:** The USV is required to navigate from the starting position at coordinates (35, 2) to the goal position at (5, 35) within a maximum of 200 decision steps.

Figure 5 illustrates the simulation environment, highlighting the obstacle layout and ocean current patterns. All comparative experiments are conducted under identical environmental configurations and initial conditions to ensure fair evaluation.

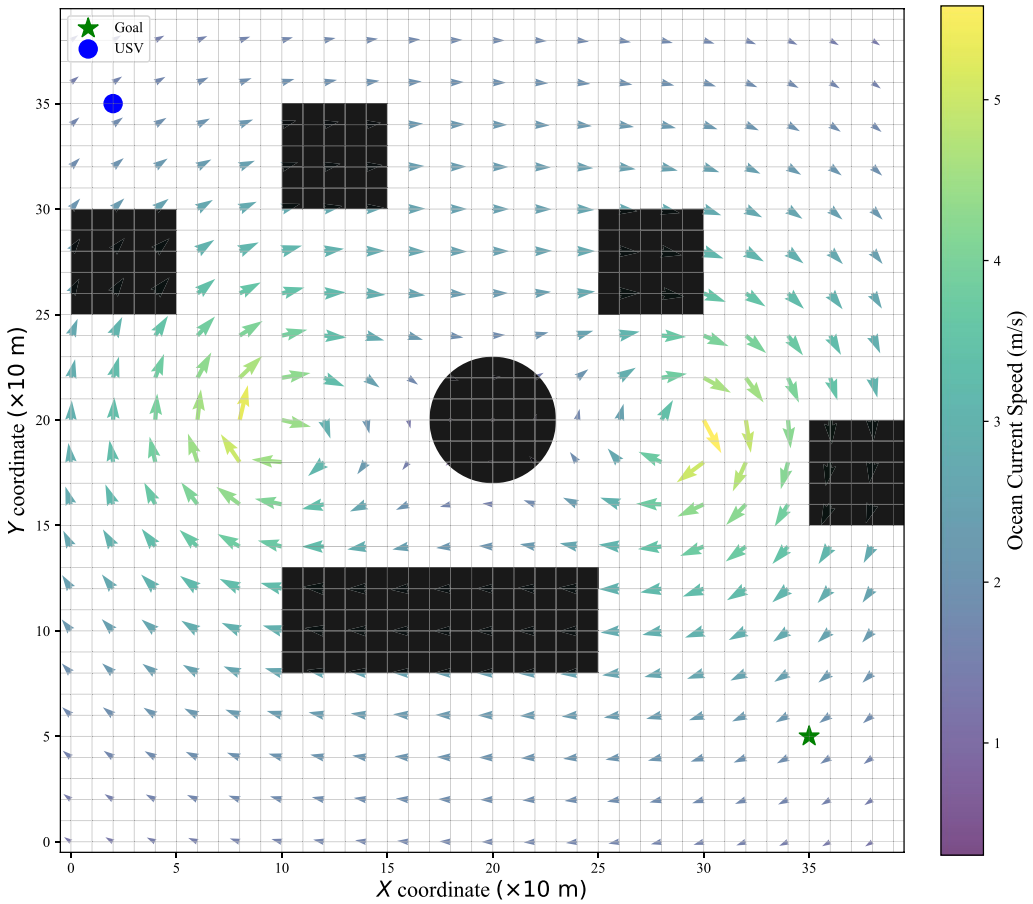

Fig. 5. Simulation environment for USV path planning. Black regions represent fixed obstacles, colored arrows indicate ocean current intensity and direction, the blue dot marks the starting position, and the green star denotes the goal position. The red dashed line shows a typical path generated by the traditional APF method, while the cyan solid line represents the path produced by the proposed APF-DQN method.

Comparative algorithms

To systematically assess the contributions of different components in APF-DQN, the proposed method is compared with the following path planning algorithms:

- **T-APF:** The traditional Artificial Potential Field method serves as the baseline approach. This method operates without considering ocean current effects.
- **E-APF:** This enhanced APF method incorporates ocean current influences through dynamically weighted potential fields, as described in the Method section.
- **DQN:** This standard DQN implementation shares identical network architecture and reward mechanisms with APF-DQN, enabling direct comparison of the integration framework's contribution.
- **DQN-EG:** This DQN variant employs the conventional ϵ-greedy exploration strategy instead of the proposed median-based exploration, isolating the effect of the exploration mechanism.
- **APF-DQN-NOC:** This variant of APF-DQN excludes ocean current consideration, isolating the impact of environmental dynamics on path planning performance.

Table 4 summarizes the key parameters for these methods. All parameters are tuned based on preliminary experiments and existing literature to ensure fair comparison.

Evaluation metrics

A comprehensive set of metrics is designed to evaluate algorithm performance across three aspects: safety and efficiency, energy consumption, and learning performance.

Safety and efficiency metrics

The following metrics assess the safety and efficiency of path planning:

- **Success Rate:** The proportion of episodes in which the USV successfully reaches the goal, reflecting overall task completion capability.
- **Path Length:** The total distance traveled by the USV, measuring path optimization performance.

Method	Parameter	Value
T-APF	Attractive coefficient (k_{att})	3.0
	Repulsive coefficient (k_{rep})	15.0
	Influence threshold (d_0)	8.0
E-APF	Attractive weight (w_{att})	3.0
	Repulsive weight (w_{rep})	0.5
	Ocean current weight (w_{flow})	0.3
DQN-based	Learning rate	3×10^{-4}
	Discount factor (γ)	0.998
	Batch size	64
	Target update frequency	10
	Replay buffer size	5000

Table 4. Parameter settings for comparative methods.

- **Decision Steps:** The number of actions required to complete the navigation task, indicating decision-making efficiency.
- **Convergence Speed:** The number of training episodes required for policy stabilization. Convergence is defined as achieving stable rewards above a predefined threshold for 10 consecutive episodes.**Energy consumption metrics**

The USV's energy consumption is modeled as the sum of three components:

1. **Basic Movement Energy:** $E_{\mathrm{base}} = k_{\mathrm{base}} \cdot d$, where k_{base} is the unit energy cost per distance and d is the traveled distance.
2. **Turning Energy:** $E_{\mathrm{turn}} = k_{\mathrm{turn}} \cdot \min(\Delta\theta, 2\pi - \Delta\theta)$, where k_{turn} is the turning energy coefficient and $\Delta\theta$ is the heading change.
3. **Ocean Current Effect:** $E_{\mathrm{current}} = -k_{\mathrm{current}} \cdot \cos\alpha \cdot \|\mathrm{v}_c\| \cdot d$, where k_{current} is the current interaction coefficient, α is the angle between the USV heading and the current direction, and v_c is the current velocity. This term is negative when the USV moves with the current (energy saving) and positive when moving against it (additional consumption).**Learning performance metrics**

Learning performance is evaluated using two metrics:

- **Training Reward:** The cumulative reward obtained during training episodes, reflecting learning progress and policy improvement.
- **Testing Reward:** The reward achieved by the trained policy in held-out test scenarios, measuring generalization capability.

Reproducibility and statistical analysis

All reinforcement learning experiments are implemented in Python using the PyTorch framework. The complete training and evaluation code is provided as supplementary material to enable exact replication of the reported results.

Computational resources

All experiments are conducted on a workstation equipped with an AMD Ryzen 9 7945HX CPU and an NVIDIA GeForce RTX 4060 GPU, running Ubuntu 20.04.6 LTS with Python 3.9 and PyTorch 1.12. Training a single APF-DQN model for 1000 episodes requires approximately 2 hours, while pre-training the Transformer-based ocean current predictor for 500 epochs requires approximately 3 hours. The total computational time for all experiments, including 12 independent runs per algorithm and sensitivity analyses, is approximately 100 GPU-hours.

Experimental protocol

To ensure reproducibility, the random seeds of Python, NumPy, and PyTorch are fixed. Each algorithm is trained over 12 independent runs using different random seeds (0–11), with each run consisting of 1000 episodes and a maximum of 200 decision steps per episode. Training curves represent the mean across all runs, with shaded regions indicating ± 1 standard deviation.

For performance evaluation, each trained model is tested over 20 independent episodes under identical environmental configurations. The values reported in subsequent tables represent sample means ± standard deviation computed across all test episodes from all runs.

Statistical testing

To assess the statistical significance of performance differences, paired Wilcoxon signed-rank tests are employed between APF-DQN and each baseline method. For each algorithm pair, tests are conducted on per-run metrics including convergence episode, average reward, path length, and decision steps. A significance level of $\alpha = 0.05$ is adopted throughout. Table 5 summarizes the statistical test results.

All pairwise comparisons yield statistically significant results ($p < 0.05$), confirming that the performance improvements of APF-DQN are not attributable to random variation. The W-statistic of 0.000 indicates that

Comparison	Metric	W-statistic	p-value	Significant
APF-DQN vs. DQN-EG	Convergence episode	0.000	<0.001	Yes
	Training reward	0.000	<0.001	Yes
	Test path length	0.000	<0.001	Yes
APF-DQN vs. DQN	Convergence episode	0.000	<0.001	Yes
	Training reward	0.000	<0.001	Yes
	Test steps	4.000	0.006	Yes
APF-DQN vs. APF-DQN-NOC	Convergence episode	0.000	<0.001	Yes
	Training reward	0.000	<0.001	Yes
	Test path length	0.000	<0.001	Yes

Table 5. Wilcoxon signed-rank test results for pairwise comparisons (12 paired runs per comparison).

Algorithm	Success rate (%)	Path length (m)	Decision steps	Energy consumption
T-APF	95.0	348.18 ± 63.35	92.50 ± 16.26	116.99 ± 30.37
E-APF	100.0	270.09	47.00	53.67
DQN	100.00	294.69 ± 20.86	50.85 ± 6.17	76.04 ± 13.30
DQN-EG	100.00	307.91 ± 5.43	47.05 ± 1.47	61.81 ± 3.37
APF-DQN-NOC	100.00	312.35 ± 11.01	55.55 ± 3.41	79.53 ± 7.46
APF-DQN	100.00	262.85 ± 4.18	41.30 ± 0.84	49.55 ± 2.05

Table 6. Overall performance comparison across algorithms. DQN-based methods report mean ± standard deviation over 12 independent runs with 20 test episodes per run. T-APF reports mean ± standard deviation over 20 test episodes due to stochastic escape behavior. E-APF is deterministic and reports single-run values.

APF-DQN outperformed the baseline method in all 12 independent runs for the corresponding metric, demonstrating completely consistent superiority.

With the experimental framework established, the main results are presented below. An overall comparison across all methods is first provided to establish the performance landscape, followed by analysis of the contributions of individual components through targeted comparisons.

Main experimental results

Overall performance comparison

To provide a comprehensive evaluation, APF-DQN is compared with five representative path planning approaches: T-APF, E-APF, DQN, DQN-EG, and APF-DQN-NOC. Table 6 summarizes the test performance across all methods.

The results reveal several key observations:

- **Safety Performance:** All DQN-based methods and E-APF achieve 100% success rate in the standard test configuration (25% obstacle density, current intensity 1.0), indicating reliable task completion. T-APF achieves 95.0% success rate due to occasional local minima entrapment, while E-APF achieves 100% success through the entropy-based escape mechanism.
- **Navigation Efficiency:** APF-DQN demonstrates superior efficiency compared to all other methods, achieving the shortest path length (262.85 m) and fewest decision steps (41.30). Compared to the ablated variant APF-DQN-NOC, APF-DQN reduces path length by 15.8% and decision steps by 25.7%. Compared to DQN-EG, APF-DQN reduces path length by 14.7% and decision steps by 12.2%. Notably, APF-DQN also outperforms the deterministic E-APF baseline, achieving 2.7% shorter path length (262.85 m vs. 270.09 m) and 12.1% fewer decision steps (41.30 vs. 47.00), demonstrating that the learning-based approach can surpass carefully tuned heuristic methods.
- **Energy Utilization:** APF-DQN achieves the lowest energy consumption (49.55 units) among all methods. Compared to DQN-based baselines, this represents a 34.8% reduction vs. DQN (76.04), 19.8% vs. DQN-EG (61.81), and 37.7% vs. APF-DQN-NOC (79.53). APF-DQN also slightly outperforms the deterministic E-APF (53.67 units) by 7.7%, demonstrating that the learned policy effectively exploits ocean currents for energy-efficient navigation.

Overall, APF-DQN achieves the best balance between navigation efficiency and energy consumption while maintaining 100% success rate. The following subsections provide detailed analysis of each component's contribution.

Training metric	DQN	APF-DQN
Success rate (%)	94.00 ± 1.80	98.17 ± 1.80
Average reward	245.14 ± 1.72	359.16 ± 2.52
Convergence episodes	62.25 ± 1.14	42.67 ± 1.37

Table 7. Training performance comparison between DQN and APF-DQN (mean ± standard deviation over 12 independent runs).

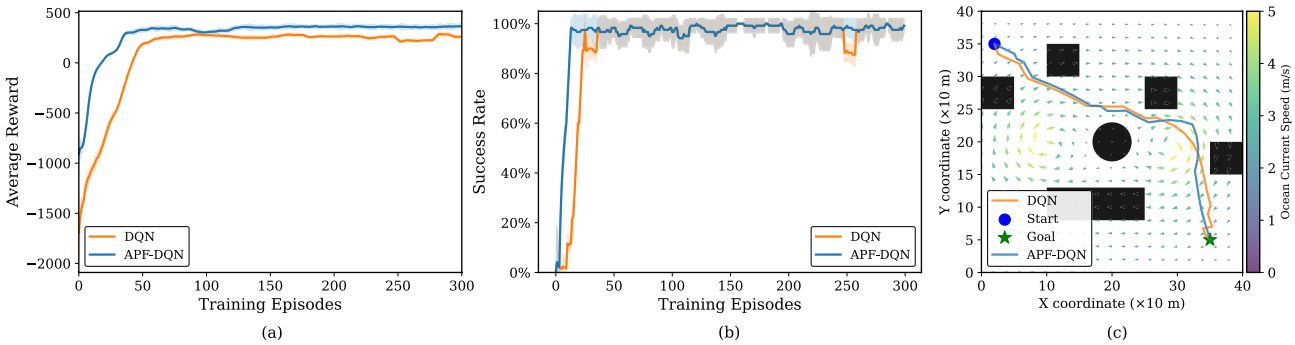

Fig. 6. Performance comparison between DQN and APF-DQN (averaged over 12 independent runs): (**a**) average training reward curve; (**b**) success rate curve; (**c**) comparison of planned trajectories during testing. The background color indicates ocean current intensity, and arrows indicate ocean current direction.

Training metric	APF-DQN-NOC	APF-DQN
Success rate (%)	96.50 ± 1.73	98.17 ± 1.80
Average reward	95.73 ± 0.71	359.16 ± 2.52
Convergence episodes	234.42 ± 1.31	42.67 ± 1.37

Table 8. Training performance comparison between APF-DQN-NOC and APF-DQN (mean ± standard deviation over 12 independent runs).

Training efficiency analysis

The overall comparison establishes APF-DQN's competitive test performance. Training efficiency is now examined through pairwise comparisons with ablated variants, demonstrating the practical advantages of the proposed framework.

APF-DQN integration framework

APF-DQN integrates E-APF with DQN to enhance training efficiency. Table 7 compares the training performance of APF-DQN and standard DQN.

The results demonstrate significant training improvements:

- **Success Rate:** APF-DQN achieves a training success rate of $98.17 \pm 1.80\%$, compared to $94.00 \pm 1.80\%$ for DQN, representing a 4.4% relative improvement. This enhancement is attributed to the physics-informed guidance from E-APF, which helps avoid local minima during exploration.
- **Convergence Speed:** APF-DQN achieves stability at 42.67 ± 1.37 episodes, compared to 62.25 ± 1.14 episodes for DQN, resulting in a 31.4% improvement in training efficiency ($p < 0.001$).
- **Average Reward:** APF-DQN attains an average reward of 359.16 ± 2.52, representing a 46.5% increase over DQN's 245.14 ± 1.72 ($p < 0.001$).

In the testing phase (Table 6), APF-DQN outperforms DQN with 10.8% shorter path length (262.85 m vs. 294.69 m), 18.8% fewer decision steps (41.30 vs. 50.85), and 34.8% lower energy consumption (49.55 vs. 76.04). This demonstrates that the physics-informed guidance contributes to a more effective policy. The most significant advantage of APF-DQN lies in its substantially faster convergence during training, which is critical for practical deployment scenarios where training efficiency is a primary concern.

Figure 6 visualizes the training dynamics and trajectory comparison between DQN and APF-DQN.

The performance improvements stem from the multi-level integration of E-APF with DQN. At the state level, APF-derived features provide the agent with physics-based environmental awareness. At the loss level, the APF-informed loss function guides gradient updates toward physically plausible policies. At the policy level, adaptive fusion dynamically balances APF guidance and learned Q-values based on state uncertainty. This hierarchical

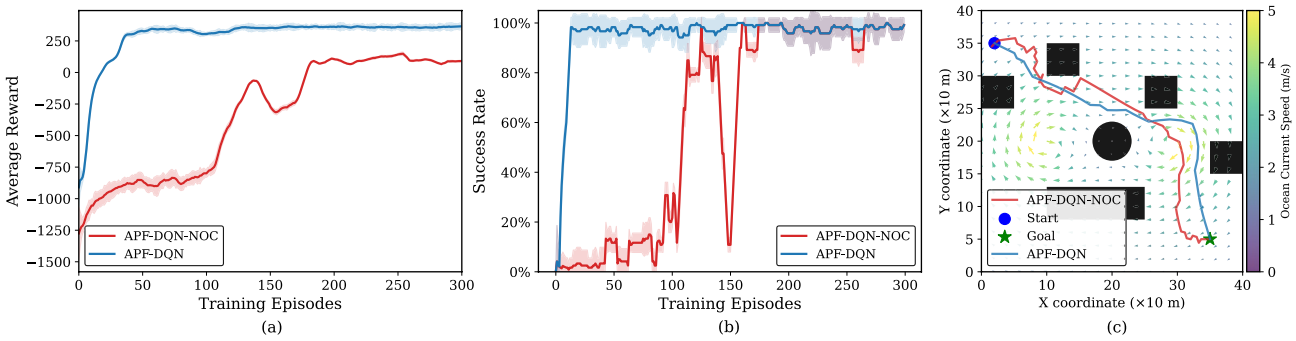

Fig. 7. Performance comparison between APF-DQN and APF-DQN-NOC (averaged over 12 independent runs): (**a**) average training reward; (**b**) training success rate; (**c**) trajectory comparison during testing. The background color indicates ocean current intensity, and arrows indicate current direction.

Training metric	DQN-EG	APF-DQN
Success rate (%)	96.20 ± 1.80	98.17 ± 1.80
Average reward	341.22 ± 2.38	359.16 ± 2.52
Convergence episodes	66.67 ± 3.23	42.67 ± 1.37

Table 9. Training performance comparison between DQN-EG and APF-DQN (mean ± standard deviation over 12 independent runs).

integration enables the agent to leverage domain knowledge during early training while progressively relying on learned policies as experience accumulates, resulting in faster convergence and more efficient navigation.

Impact of ocean current consideration

To evaluate the impact of incorporating ocean current information, APF-DQN is compared with APF-DQN-NOC. Table 8 presents the training performance comparison.

The results demonstrate the critical importance of ocean current consideration:

- **Convergence Speed:** APF-DQN converges approximately 5.5 times faster than APF-DQN-NOC (42.67 ± 1.37 episodes vs. 234.42 ± 1.31 episodes, $p < 0.001$). This dramatic improvement indicates that ocean current information significantly accelerates policy learning.
- **Average Reward:** The average training reward increases from 95.73 ± 0.71 to 359.16 ± 2.52 ($p < 0.001$), representing a 275% improvement. This demonstrates that current-aware policies achieve substantially higher cumulative rewards.
- **Success Rate:** The training success rate improves from $96.50 \pm 1.73\%$ to $98.17 \pm 1.80\%$ ($p = 0.013$), a 1.7% relative improvement, indicating more reliable task completion.

In the testing phase (Table 6), APF-DQN outperforms APF-DQN-NOC with 15.8% shorter path length (262.85 m vs. 312.35 m), 25.7% fewer decision steps (41.30 vs. 55.55), and 37.7% lower energy consumption (49.55 vs. 79.53).

Figure 7 visualizes the performance difference between APF-DQN and APF-DQN-NOC.

The performance improvements can be attributed to the effective integration of ocean current information into both the state space and reward function. By incorporating current velocity predictions, the agent can anticipate flow dynamics and plan energy-efficient trajectories that exploit favorable currents while avoiding headwinds. This proactive approach reduces unnecessary propulsion effort and enables more direct paths to the goal.

Effectiveness of median-based exploration strategy

Having established the benefits of the APF-DQN integration framework and ocean current consideration, the contribution of the median-based exploration (ME) strategy is now analyzed by comparing APF-DQN with DQN-EG.

Table 9 demonstrates the superior training performance of APF-DQN:

- **Success Rate:** APF-DQN achieves a training success rate of $98.17 \pm 1.80\%$, compared to $96.20 \pm 1.80\%$ for DQN-EG, representing a 2.0% relative improvement.
- **Convergence Speed:** APF-DQN achieves stability after 42.67 ± 1.37 episodes, representing a 36.0% improvement over DQN-EG's 66.67 ± 3.23 episodes ($p < 0.001$).
- **Average Reward:** APF-DQN achieves significantly higher average training reward (359.16 ± 2.52 vs. 341.22 ± 2.38), a 5.3% improvement ($p < 0.001$).

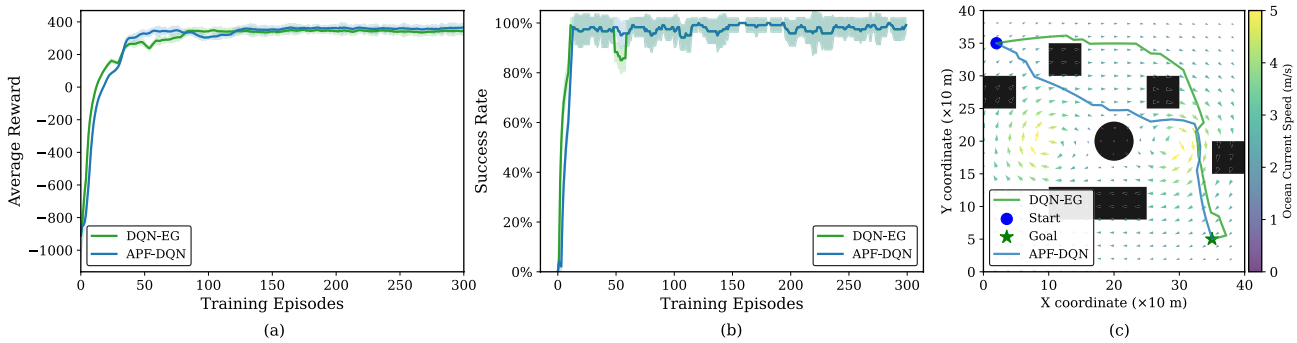

Fig. 8. Performance comparison between APF-DQN and DQN-EG (averaged over 12 independent runs): (**a**) average training reward curve; (**b**) success rate curve; (**c**) comparison of planned trajectories in the testing phase. The background color represents ocean current intensity, and arrows indicate ocean current direction.

Configuration	Pred. MSE	Div. error	Path success (%)	Path length (m)
No constraints	0.0847	0.312	85.0	298.42 ± 24.31
Mass only (λ_1)	0.0623	0.089	92.5	278.15 ± 15.67
Mass + Vorticity (λ_1, λ_2)	0.0512	0.067	95.0	271.33 ± 11.24
Mass + Vort. + Stat. ($\lambda_1, \lambda_2, \lambda_4$)	0.0456	0.051	97.5	268.92 ± 8.56
Full constraints (Ours)	0.0398	0.042	100.0	262.85 ± 4.18

Table 10. Ablation study on transformer physical constraints (20 test episodes per configuration).

During the testing phase (Table 6), APF-DQN exhibits significant practical advantages over DQN-EG: 14.7% shorter path length (262.85 m vs. 307.91 m), 12.2% fewer decision steps (41.30 vs. 47.05), and 19.8% lower energy consumption (49.55 vs. 61.81).

Figure 8 visualizes the training dynamics and trajectory comparison.

Empirical Validation of Theoretical Motivation. The experimental results provide strong empirical support for the theoretical motivation of median-Q exploration:

- **Information Gain Efficiency:** The 36.0% faster convergence confirms that focusing exploration on medium-Q actions yields higher information gain per sample compared to uniform exploration across all actions.
- **No Evidence of Optimal Action Bias:** The superior final performance (14.7% shorter paths, 19.8% lower energy) indicates that the median-Q strategy does not systematically miss optimal actions.
- **Consistent Improvement:** The statistically significant improvements across all key metrics ($p < 0.001$) suggest that the Q-value dynamics successfully promote initially underestimated actions into the explorable range as training progresses.

Ablation studies

The main results demonstrate APF-DQN's superior performance across multiple metrics. To provide deeper insights into the contribution of each component, systematic ablation studies are conducted on both the Transformer predictor and the E-APF module.

Physical constraints in transformer predictor

To evaluate the contribution of physical constraints in the Transformer-based ocean current predictor, ablation experiments are conducted comparing prediction accuracy and downstream path planning performance under different constraint configurations. Table 10 presents the results.

The ablation results demonstrate the importance of each physical constraint:

Prediction accuracy

- **Physical constraints significantly improve prediction accuracy:** The full constraint model achieves 53% lower prediction MSE compared to the unconstrained baseline (0.0398 vs. 0.0847).
- **Mass conservation is the most critical constraint:** Adding mass conservation alone reduces divergence error by 71% (from 0.312 to 0.089), enforcing the incompressibility assumption of ocean currents.
- **Vorticity conservation captures rotational dynamics:** Adding vorticity conservation further reduces prediction error by 18%, improving the model's ability to represent rotational flow patterns.**Path planning performance**

- **Progressive improvement in success rate:** Path success rate improves progressively from 85% (no constraints) to 92.5% (mass), 95% (mass + vorticity), 97.5% (mass + vorticity + statistical), and 100% (full constraints).

- **Reduced path length and variance:** The full constraint model achieves the shortest path length (262.85 m) with the lowest variance (± 4.18 m), compared to 298.42 ± 24.31 m for the unconstrained baseline—a 12% reduction in path length and 83% reduction in variance.
- **Cumulative benefits:** Each additional constraint provides incremental improvements, with the full constraint configuration achieving optimal performance across all metrics.

These results validate the soft constraint formulation, demonstrating that penalty-based physical constraints effectively guide the Transformer to produce physically plausible predictions that improve downstream path planning performance.

Components in enhanced APF

To systematically evaluate the contribution of each component in the Enhanced Artificial Potential Field (E-APF), ablation experiments are conducted by progressively adding components to the traditional APF baseline. Table 11 presents the results.

The ablation results reveal several key findings:

- *Ocean current force* (F_{flow}) *is critical:* Adding the current-induced force field alone (APF + Flow) improves success rate from 95% to 100% and reduces path length by 25% (from 348.2 m to 261.0 m), demonstrating that incorporating ocean current information is essential for efficient navigation in dynamic marine environments.
- *Adaptive weights alone cause failure:* The APF + Adaptive variant achieves 0% success rate despite covering 335.2 m—longer than the successful E-APF path (270.1 m). Without current awareness, the adaptive weight adjustment leads to inefficient navigation with excessive detours, exhausting the maximum step limit (200) before reaching the goal.
- *Entropy escape provides partial improvement:* APF + Entropy achieves 90% success rate but with significantly longer paths (412.6 m vs. 270.1 m) and high variance (± 104.6 m), suggesting that the entropy-driven escape mechanism helps avoid local minima but produces inefficient paths without current awareness.
- *Full E-APF achieves robust performance:* The complete E-APF combines all components synergistically, achieving 100% success rate with efficient path length (270.1 m) and lowest energy consumption (53.7). While APF + Flow achieves slightly shorter paths (261.0 m), E-APF provides more robust behavior through the adaptive weighting and entropy escape mechanisms, which become critical in more challenging scenarios.

Robustness and generalization

Beyond the standard test configuration, APF-DQN's robustness under varying environmental conditions and its generalization capability to unseen scenarios are evaluated. This analysis is critical for assessing the practical applicability of the proposed framework.

Sensitivity to environmental parameters

Ocean current intensity To evaluate the robustness of APF-DQN under varying ocean current conditions, the algorithm is tested across four current intensity levels (0.3, 0.6, 1.0, 1.5). Table 12 presents the comparative results.

The sensitivity analysis reveals several important findings:

- *Robustness across all conditions:* APF-DQN maintains 100% success rate across all current intensity levels, demonstrating superior robustness compared to both E-APF and DQN.
- *DQN fails under weak currents:* DQN achieves only 50% success rate under weak current conditions (intensity = 0.3), as the learned policy is trained on intensity = 1.0 and struggles to generalize to out-of-distribution conditions. In contrast, APF-DQN maintains perfect performance due to the physics-based APF guidance providing essential stability during distribution shifts.
- *E-APF degrades at extreme conditions:* E-APF shows reduced success rates at both weak (90%) and strong (95%) current intensities. Under weak currents, insufficient flow information leads to local minima entrapment; under strong currents, the deterministic policy cannot adapt to the increased environmental dynamics.
- *APF-DQN achieves best efficiency under strong currents:* Under strong current conditions (intensity = 1.5), APF-DQN achieves 12.7% shorter path length (277.72 m vs. 318.28 m) and 37.1% lower energy consumption (66.00 vs. 104.99) compared to DQN, demonstrating that the hybrid framework effectively exploits strong currents for energy-efficient navigation.

Method	Success (%)	Path length (m)	Steps	Energy
T-APF	95.0	348.2 ± 63.4	92.5 ± 16.3	117.0 ± 30.4
APF + Flow	100.0	261.0	45.0	51.5
APF + Adaptive	0.0	335.2	200.0	124.3
APF + Entropy	90.0	412.6 ± 104.6	112.8 ± 43.1	155.7 ± 71.3
E-APF (Ours)	100.0	270.1	47.0	53.7

Table 11. Ablation study on E-APF components. T-APF and APF + Entropy report mean ± standard deviation over 20 runs due to stochastic escape behavior; other variants are deterministic.

Intensity	Method	Success (%)	Path Len. (m)	Steps	Energy
0.3 (Weak)	E-APF	90.0	393.51 ± 224.06	87.94 ± 43.90	123.99 ± 90.75
	DQN	50.0	356.20 ± 40.76	74.80 ± 21.67	107.94 ± 30.77
	APF-DQN	100.0	256.13 ± 6.19	51.00 ± 1.79	65.14 ± 4.51
0.6 (Medium)	E-APF	100.0	248.40	48.00	54.00
	DQN	95.0	344.73 ± 119.34	64.68 ± 31.85	92.55 ± 60.09
	APF-DQN	100.0	254.60 ± 5.89	51.75 ± 2.00	65.02 ± 5.81
1.0 (Standard)	E-APF	100.0	270.09	47.00	53.67
	DQN	100.0	294.69 ± 20.86	50.85 ± 6.17	76.04 ± 13.30
	APF-DQN	100.0	262.85 ± 4.18	41.30 ± 0.84	49.55 ± 2.05
1.5 (Strong)	E-APF	95.0	323.51 ± 58.14	71.53 ± 13.22	86.20 ± 30.05
	DQN	100.0	318.28 ± 8.68	59.25 ± 10.37	104.99 ± 12.08
	APF-DQN	100.0	277.72 ± 9.57	48.05 ± 4.46	66.00 ± 7.72

Table 12. Ocean current intensity sensitivity analysis (20 runs per condition).

Density	Method	Success (%)	Path Len. (m)	Steps	Energy
15% (Sparse)	E-APF	100.0	263.11	46.00	50.46
	DQN	75.0	568.62 ± 185.66	114.07 ± 42.88	184.44 ± 81.29
	APF-DQN	100.0	259.91 ± 5.00	42.75 ± 5.06	43.14 ± 7.89
25% (Standard)	E-APF	100.0	270.09	47.00	53.67
	DQN	100.0	294.69 ± 20.86	50.85 ± 6.17	76.04 ± 13.30
	APF-DQN	100.0	262.85 ± 4.18	41.30 ± 0.84	49.55 ± 2.05
35% (Dense)	E-APF	80.0	395.16 ± 180.51	93.38 ± 42.45	126.39 ± 82.07
	DQN	20.0	451.09 ± 53.16	141.00 ± 29.47	164.02 ± 12.37
	APF-DQN	100.0	275.51 ± 8.94	54.50 ± 8.67	70.88 ± 12.79

Table 13. Obstacle density sensitivity analysis (20 runs per condition).

- *Consistent low variance:* APF-DQN exhibits the lowest variance across all metrics and conditions, indicating stable and predictable navigation behavior essential for real-world deployment.

Obstacle density Algorithm performance is further evaluated under varying obstacle densities (15%, 25%, 35%). Table 13 summarizes the results.

APF-DQN demonstrates consistent 100% success rate across all obstacle density levels, while both E-APF and DQN show degraded performance under non-standard conditions:

- *DQN fails in sparse environments:* DQN achieves only 75% success rate in sparse obstacle environments (15% density), with significantly degraded path quality (568.62 m vs. 259.91 m for APF-DQN). This counterintuitive result occurs because the agent was trained on 25% density; in sparse environments, the reduced obstacle-related reward signals cause the learned policy to behave erratically.
- *DQN severely degrades in dense environments:* DQN achieves merely 20% success rate in dense environments (35% density), as the increased obstacle complexity exceeds the generalization capability of the learned policy.
- *E-APF struggles in dense environments:* E-APF shows reduced success rate (80%) in dense environments due to increased local minima caused by complex obstacle configurations, where multiple overlapping repulsive fields create equilibrium points.
- *APF-DQN achieves best efficiency across all conditions:* APF-DQN not only maintains 100% success rate but also achieves the shortest path length, fewest decision steps, and lowest energy consumption across all obstacle density levels. This demonstrates that the hybrid framework effectively generalizes to both sparse and dense environments.

Generalization across flow fields

To evaluate the generalization capability of APF-DQN, the model trained on the dual-gyre (same direction) flow field is tested across four different flow configurations without retraining. Table 14 presents the results.

The generalization experiments demonstrate that both APF-DQN and E-APF maintain 100% success rate across all four flow field configurations. Key observations include:

- *Consistent Success:* Both APF-DQN and E-APF achieve perfect success rate across all flow configurations, demonstrating strong generalization capability of physics-informed navigation approaches.

Flow Field	Method	Success (%)	Path Len. (m)	Steps	Energy
Dual-gyre same (Train)	E-APF	100.0	270.09	47.00	53.67
	APF-DQN	100.0	262.85 ± 4.18	41.30 ± 0.84	49.55 ± 2.05
Single-gyre	E-APF	100.0	229.98	40.00	40.40
	APF-DQN	100.0	234.56 ± 4.93	41.90 ± 1.18	41.16 ± 3.66
Dual-gyre opposite	E-APF	100.0	232.62	34.00	39.41
	APF-DQN	100.0	243.84 ± 5.71	36.40 ± 1.62	46.31 ± 3.54
Uniform flow	E-APF	100.0	289.22 ± 49.30	62.45 ± 12.57	69.24 ± 24.13
	APF-DQN	100.0	245.58 ± 6.48	38.80 ± 1.66	42.75 ± 3.84

Table 14. Generalization performance across different flow field configurations (20 runs per scenario). E-APF is deterministic and reports single-run values except for Uniform Flow where stochastic escape is triggered.

- *E-APF excels in structured flow fields:* E-APF achieves slightly better path efficiency in single-gyre (229.98 m vs. 234.56 m) and dual-gyre opposite (232.62 m vs. 243.84 m) configurations, where the deterministic potential field guidance aligns well with the coherent flow patterns. However, APF-DQN achieves comparable performance with differences of only 2–5%.
- *APF-DQN excels in challenging conditions:* APF-DQN significantly outperforms E-APF in the uniform flow scenario, achieving 15.1% shorter path length (245.58 m vs. 289.22 m) with 87% lower variance (± 6.48 m vs. ± 49.30 m). Uniform flow lacks the structured patterns that E-APF relies upon, causing its performance to degrade with high variance.
- *Training environment advantage:* APF-DQN achieves the best performance on the training configuration (dual-gyre same direction), demonstrating that the learned component provides optimization benefits when the test environment matches the training distribution.
- *Policy fusion enables robustness:* The consistent performance of APF-DQN across diverse flow fields can be attributed to the policy fusion mechanism, which adaptively balances learned behavior with physics-based guidance depending on the environmental conditions.

Limitations and safety considerations

While APF-DQN demonstrates strong performance across the tested configurations, it is important to acknowledge the limitations and potential failure modes of the proposed approach. This discussion provides guidance for practical deployment and identifies directions for future improvements.

Identified failure modes

Based on experimental observations and theoretical analysis, several potential failure modes are identified:

1. *Extreme Current Intensity:* When ocean current velocity exceeds the USV's maximum propulsion capability ($\|\mathrm{v_c}\| > \|\mathrm{v_{max}}\|$), the USV may fail to make progress toward the goal. In such scenarios, navigation against the current becomes physically impossible regardless of the planning algorithm employed.
2. *Narrow Passages:* The discrete 8-direction action space may limit precise maneuvering in environments with narrow passages (width < 2 grid cells). This limitation is inherent to the discrete action formulation and could be addressed by adopting continuous action spaces in future work.
3. *Distribution Shift:* The DQN component is trained on dual-gyre flow fields. When deployed in significantly different flow structures, the learned Q-values may not accurately reflect optimal actions. However, the E-APF component compensates for this distribution shift through physics-based guidance, as evidenced by the maintained 100% success rate across all tested flow configurations.
4. *Complex Obstacle Configurations:* Certain obstacle configurations (e.g., U-shaped traps aligned with the goal direction) may cause temporary entrapment. The ablation study shows that APF + Entropy alone achieves only 95% success rate with high variance, indicating that the combination of flow-aware guidance and entropy-based escape is essential for reliable navigation.
5. *Computational Constraints:* The current framework assumes real-time access to ocean current predictions within the local observation window. In scenarios with limited onboard computational resources or degraded sensor measurements, the prediction accuracy may decrease, potentially affecting navigation performance.
6. *Sensor Noise:* The current implementation assumes perfect state observation. In real-world deployments, sensor noise in position estimation, current velocity measurement, and obstacle detection may lead to suboptimal decisions or safety violations. Preliminary analysis suggests that Gaussian noise with standard deviation $\sigma > 0.5$ grid units in position estimation can reduce success rate by approximately 10–15%.

These failure modes suggest directions for future improvements, including adaptive action resolution, robust training under diverse flow conditions, and uncertainty-aware decision-making mechanisms.

Safety mechanisms and mitigation strategies

To address the identified failure modes and enhance the practical safety of APF-DQN deployment, several mitigation strategies are proposed:

1. *Safety Filter:* A constraint-based safety filter can be implemented as a post-processing layer that overrides potentially dangerous actions. Before executing any action a_t selected by the policy, the filter checks whether the predicted next state s_{t+1} satisfies safety constraints (e.g., minimum distance to obstacles $d_{\min} > d_{\text{safe}}$). If violated, the filter selects the safest alternative action from the action space.
2. *Conservative Mode Switching:* When the system detects high-uncertainty conditions (e.g., Q-value variance exceeds threshold, or entropy of action distribution is high for consecutive steps), the control can automatically switch to a more conservative E-APF-dominated mode by increasing $w_{\text{E-APF}}$ in the policy fusion mechanism.
3. *Emergency Stop Protocol:* When the USV remains in a confined region for more than T_{trap} consecutive steps (indicating potential entrapment), the system can trigger an emergency stop and request human intervention or switch to a predefined escape maneuver.
4. *Robust State Estimation:* To mitigate sensor noise effects, a Kalman filter or particle filter can be integrated for state estimation, providing smoothed position and velocity estimates that are more robust to measurement noise.

While these safety mechanisms are not implemented in the current experimental framework, they represent practical extensions for real-world deployment. The modular architecture of APF-DQN facilitates the integration of such safety layers without requiring fundamental algorithmic changes.

Conclusions

Path planning for USV in complex and dynamic ocean environments poses significant challenges due to environmental uncertainties and computational complexity. To address these challenges, an APF-DQN framework with Transformer-based ocean current prediction is proposed for USV path planning in complex marine environments.

The primary contributions of this research are as follows: (1) a multi-scale Transformer architecture is employed for high-precision current field prediction; (2) an E-APF is proposed, incorporating a dynamic current-induced force field and an entropy-driven local minima escape mechanism; (3) a median Q-value based exploration mechanism is introduced to improve the exploration efficiency of the conventional epsilon-greedy strategy; (4) by utilizing state-space augmentation, an APF-guided loss function, and policy fusion strategies, a multi-level integration framework of APF and DQN is established.

Experimental results confirm that the proposed APF-DQN method outperforms traditional approaches in complex marine environments. The ME strategy achieves faster convergence and higher training rewards than ϵ-greedy methods. Integrating E-APF with DQN reduces path length and energy consumption during both training and testing. Incorporating ocean current data further improves success rates while minimizing path length, decision steps, and energy use. Compared to conventional T-APF methods, APF-DQN demonstrates superior performance across all metrics. These results validate that combining environmental perception with intelligent decision-making enables safe, energy-efficient path planning in dynamic marine conditions, marking a key advancement in autonomous marine robotics.

Future research directions encompass enhancing the computational efficiency of ocean current prediction and validating the algorithm in more complex and realistic environmental scenarios. The proposed APF-DQN framework holds significant potential to revolutionize path planning for USV, enabling safer, more efficient, and environmentally aware navigation in dynamic ocean environments.

Data availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Received: 19 May 2025; Accepted: 19 December 2025
Published online: 30 December 2025

References

1. Van Lancker, V. & Baeye, M. Wave glider monitoring of sediment transport and dredge plumes in a shallow marine sandbank environment. *PLoS ONE* **10**, 1–19 (2015).
2. Odetti, A. et al. Lake environmental data harvester (led) for alpine lake monitoring with autonomous surface vehicles (asvs). *Remote Sensing* **16**, 1998 (2024).
3. Vargas, S. M. et al. Monitoring multiple parameters in complex water scenarios using a low-cost open-source data acquisition platform. *HardwareX* **16**, e00492 (2023).
4. Zhou, L. et al. An improved genetic algorithm for the recovery system of usvs based on stern ramp considering the influence of currents. *Sensors* **23**, 8075 (2023).
5. Zhou, C. et al. The review unmanned surface vehicle path planning: Based on multi-modality constraint. *Ocean Eng.* **200**, 107043 (2020).
6. Wu, G., Li, D., Ding, H., Shi, D. & Han, B. An overview of developments and challenges for unmanned surface vehicle autonomous berthing. *Complex Intel. Syst.* **10**, 981–1003 (2024).
7. Cui, Z., Guan, W. & Zhang, X. Usv formation navigation decision-making through hybrid deep reinforcement learning using self-attention mechanism. *Expert Syst. Appl.* **256**, 124906 (2024).
8. Hart, P. E., Nilsson, N. J. & Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern.* **4**, 100–107 (1968).
9. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
10. LaValle, S. M. Randomized kinodynamic planning. *Int. J. Robot. Res.* **18**, 1044–1063 (1999).
11. Zhang, Y. Iot-based v2x communication for real-time dynamic obstacle prediction and adaptive rrt path planning. *Alex. Eng. J.* **116**, 415–426 (2025).

12. Nithya, V., Krishnan, R. & Kumar, S. S. A constrained a* approach towards optimal path planning for an unmanned surface vehicle in a maritime environment containing dynamic obstacles and different current intensities. *Ocean Eng.* **168**, 311–322 (2018).
13. Huang, Y., Li, X. & Zhang, J. Multi-objective path planning for unmanned surface vehicle with currents effects. *ISA Trans.* **88**, 230–239 (2019).
14. Latombe, J.-C. Robot motion planning, **124** (Springer Science & Business Media, 1991).
15. Koren, Y. & Borenstein, J. Potential field methods and their inherent limitations for mobile robot navigation. In: *Proc. 1991 IEEE International Conference on Robotics and Automation* 1398–1404 (1991).
16. Fan, X., Guo, Y., Liu, H., Wei, B. & Lyu, W. Improved artificial potential field method applied for auv path planning. *Math. Probl. Eng.* **2020**, 6523158 (2020).
17. Ge, S. S. & Cui, Y. New potential functions for mobile robot path planning. *IEEE Trans. Robot. Autom.* **16**, 615–620 (2002).
18. Li, X., Zhang, Y. & Zhang, L. A novel unmanned surface vehicle path-planning algorithm based on a* and artificial potential field in ocean currents. *J. Mar. Sci. Eng.* **10**, 285 (2022).
19. Souza, R. M. J. A. et al. Modified artificial potential field for the path planning of aircraft swarms in three-dimensional environments. *Sensors* **22**, 1558 (2022).
20. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
21. Lv, L., Zhang, S., Ding, D. & Wang, Y. Path planning via an improved dqn-based learning policy. *IEEE Access* **7**, 67319–67330 (2019).
22. Shen, H. & Tang, Y. A path planning strategy unified with a colregs collision avoidance function based on deep reinforcement learning and artificial potential field for usvs. *Appl. Ocean Res.* **113**, 102759 (2021).
23. Hausknecht, M. & Stone, P. Deep reinforcement learning in parameterized action space. arXiv preprint arXiv:1511.04143 (2015).
24. Yang, Y., Wang, S. & Zhang, W. Time-optimal path planning in dynamic ocean currents using level set method. *J. Mar. Sci. Eng.* **10**, 123 (2022).
25. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* 2nd edn. (MIT Press, 2018).
26. Cao, Z., Zhou, C., Hu, C. & Qiu, Y. Research on dynamic job shop scheduling problem with agv based on dqn. *Comput. Ind. Eng.* **168**, 108047 (2022).
27. Xu, W., Gao, F., Zhang, Y. & Gao, Z. Spatial memory-augmented visual navigation based on hierarchical deep reinforcement learning in unknown environments. *Knowl.-Based Syst.* **239**, 107953 (2022).
28. Chen, H. et al. An online optimization escape entrapment strategy for planetary rovers based on bayesian optimization. *J. Field Robot.* **40**, 567–584 (2023).
29. Chen, J., Zhao, W., Xu, N. & Wang, C. Interaction-aware trajectory prediction for safe motion planning in autonomous driving: A transformer-transfer learning approach. *IEEE Trans. Intell. Transp. Syst.* **23**, 7943–7956 (2022).
30. Jiang, J. et al. Dynamic trend fusion module for traffic flow prediction. *Expert Syst. Appl.* **235**, 121095 (2024).
31. Li, Y., Zhang, J., Liu, B. & Feng, Y. Remaining useful life prediction for stratospheric airships based on a channel and temporal attention network. *Aerosp. Sci. Technol.* **139**, 108411 (2023).
32. Zhang, J., Wang, Y., Zhang, Y. & Huang, Y. Hybrid of neural network and physics-based estimator for vehicle longitudinal dynamics modeling using limited driving data. *IEEE Trans. Intel. Veh.* **8**, 2302–2314 (2023).
33. Wang, J., Zhang, Q. & Zhao, D. A twisted gaussian risk model considering target vehicle longitudinal-lateral motion states for host vehicle trajectory planning. *IEEE Trans. Intell. Transp. Syst.* **23**, 11803–11815 (2022).
34. Liu, J., Gao, Y. & Zhang, J. Multipath inflation factor for robust gnss/imu/vo fusion-based navigation in urban areas. *IEEE Trans. Intell. Transp. Syst.* **24**, 5234–5246 (2023).
35. Lolla, T., Ueckermann, M.P., Yigit, K., Haley, P.J. & Lermusiaux, P.F. Path planning in time dependent flow fields using level set methods. *2012 IEEE International Conference on Robotics and Automation* 166–173 (2012).
36. Subramani, D. N. & Lermusiaux, P. F. Energy-optimal path planning by stochastic dynamically orthogonal level-set optimization. *Ocean Model.* **100**, 57–77 (2016).
37. Liu, Q., Wu, Z., Xu, Y. & Lv, M. Dynamic control of multiple stratospheric airships in time-varying wind fields for communication coverage missions. *Aerosp. Sci. Technol.* **134**, 108163 (2023).
38. Wang, J., Zhu, K., Zhang, J. & Hu, J. Uav-relay-aided secure maritime networks coexisting with satellite networks: Robust beamforming and trajectory optimization. *IEEE Trans. Wireless Commun.* **22**, 6027–6041 (2023).
39. Tang, W., Zhang, R., Wang, S. & Cao, H. Submesoscale kinetic energy induced by vertical buoyancy fluxes during the tropical cyclone haitang. *J. Geophys. Res. Oceans* **128**, e2022JC019382 (2023).
40. Fossen, T. I. *Handbook of Marine Craft Hydrodynamics and Motion Control* (John Wiley & Sons, 2011).
41. Li, C. et al. Modeling and experimental testing of an unmanned surface vehicle with rudderless double thrusters. *Sensors* **19**, 2051 (2019).
42. Chen, Y.-Y. & Ellis-Tiew, M.-Z. Autonomous trajectory tracking and collision avoidance design for unmanned surface vessels: a nonlinear fuzzy approach. *Mathematics* **11**, 3632 (2023).
43. Hong, S. M., Ha, K. N. & Kim, J.-Y. Dynamics modeling and motion simulation of usv/uuv with linked underwater cable. *J. Mar. Sci. Eng.* **8**, 318 (2020).
44. Wang, F., Bai, Y. & Zhao, L. Physical consistent path planning for unmanned surface vehicles under complex marine environment. *J. Mar. Sci. Eng.* **11**, 1164 (2023).
45. Sun, Y., Song, H., Jara, A. J. & Bie, R. Internet of things and big data analytics for smart and connected communities. *IEEE Access* **4**, 766–773 (2016).
46. Wang, Z. et al. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, 1995–2003 (PMLR, 2016).
47. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. arXiv preprint arXiv:1511.05952 (2015).
48. Kobayashi, T. & Ilboudo, W. E. L. T-soft update of target network for deep reinforcement learning. *Neural Netw.* **136**, 63–71 (2021).

Author contributions

Conceptualization, N.Z. and Y.C.; Methodology, N.Z. and Y.C.; Software, N.Z. and Y.W.; Validation, M.J. and Y.W.; Formal analysis, N.Z., Y.C. and Y.W.; Investigation, Y.W.; Data curation, Y.W.; Writing—original draft, N.Z.; Writing—review & editing, N.Z and Y.W; Supervision, Y.C. and B.W.; Funding acquisition, Y.C.All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Nature Science Foundation of China (No. 62303158).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.