Never too Prim to Swim: An LLM-Enhanced RL-based Adaptive S-Surface Controller for AUVs under Extreme Sea Conditions

Guanwen Xie[1,*], Jingzehua Xu[1,*,✉], Yimian Ding[1], Zhi Zhang[1], Shuai Zhang[2] and Yi Li[1,✉]

***Abstract*— The adaptivity and maneuvering capabilities of Autonomous Underwater Vehicles (AUVs) have drawn significant attention in oceanic research, due to the unpredictable disturbances and strong coupling among the AUV's degrees of freedom. In this paper, we developed large language model (LLM)-enhanced reinforcement learning (RL)-based adaptive S-surface controller for AUVs. Specifically, LLMs are introduced for the joint optimization of controller parameters and reward functions in RL training. Using multi-modal and structured explicit task feedback, LLMs enable joint adjustments, balance multiple objectives, and enhance task-oriented performance and adaptability. In the proposed controller, the RL policy focuses on upper-level tasks, outputting task-oriented high-level commands that the S-surface controller then converts into control signals, ensuring cancellation of nonlinear effects and unpredictable external disturbances in extreme sea conditions. Under extreme sea conditions involving complex terrain, waves, and currents, the proposed controller demonstrates superior performance and adaptability in high-level tasks such as underwater target tracking and data collection, outperforming traditional PID and SMC controllers.** [3]

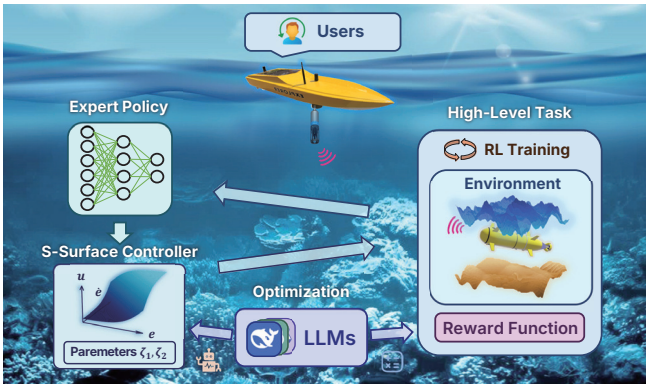


Fig. 1: **Illustration of an AUV conducting underwater tasks using the proposed controller**. The proposed controller utilizes an RL-based S-surface controller to enable effective control. LLMs assist the controller by optimizing the reward functions for RL training and tuning the parameters of the S-surface controller.

I. INTRODUCTION

The adaptive control and maneuvering capabilities of Autonomous Underwater Vehicles (AUVs) have drawn significant attention in oceanic research due to their substantial potential in maritime applications, including underwater resource exploration [1], shipwreck search [2], and underwater structure maintenance [3]. These capabilities contribute significantly to marine science and the economy [4], but require advanced control systems that provide task-adaptive and precise control of AUVs' position and attitude, particularly under extreme sea conditions [5]. However, achieving precise maneuvering control of AUVs is challenging due to their highly nonlinear dynamics [6], time-varying hydrodynamics, strong six-degree-of-freedom coupling, and environmental uncertainties [7]. During ocean navigation, AUVs encounter unpredictable external disturbances [8], requiring continuous high-precision trajectory tracking and obstacle avoidance during tasks such as coral reef ecosystem monitoring [9], which necessitates balancing multiple objectives [10]. Additionally, position uncertainties caused by extreme sea conditions [11] require additional control compensation.

Researchers have developed various control methodologies for AUVs, including PID controllers, sliding mode control (SMC) [12], fuzzy control [13], and model predictive control (MPC) [14]. While these methods demonstrate advantages in most scenarios, they exhibit limited adaptability in extreme conditions. Specifically, PID controllers require time-consuming parameter tuning for complex environments [15]. Fuzzy controllers provide good stability but are limited by the complexity of defining membership functions, inference methods, and fuzzy rules [13]. MPC predicts future behavior for optimized control but heavily relies on real-time computation and accurate system models, reducing its robustness in extreme conditions [14].

The S-Surface controller has shown promise in handling uncertainties and nonlinearities, which leverages a sigmoid plane-like surface to control AUVs' dynamic systems towards desired states [16]. However, it lacks the flexibility to adaptively adjust parameters and control strategies to handle the strong coupling between degrees of freedom [17]. The emergence of Reinforcement Learning (RL) has somehow addressed these issues. By training robots to learn adaptive control strategies through environmental interactions, RL has shown promising results in various applications including drone control [18], legged robot navigation [19], and other autonomous systems [20]. Although RL faces challenges like reward function design, its strong learning ability enables AUVs to develop expert-level control strategies that

*These authors contribute to this work equally.

✉Corresponding authors.

[1]G. Xie, J. Xu, Y. Ding, Z. Zhang and Y. Li are with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China. E-mail: {xgw24, xjzh23, dingym24, zzhang23}@mails.tsinghua.edu.cn, liyi@sz.tsinghua.edu.cn.

[2]S. Zhang is with Department of Data Science, New Jersey Institute of Technology, NJ 07102, USA. E-mail: sz457@njit.edu.

[3]The accompanying videos, details about prompts and LLM responses, and source code are available at the website **https://360zmem.github.io/AUV-RSControl** .

autonomously map high-level 6-DoF commands to end-to-end control signals, including thruster commands [7]. Also, with the assistance of the Large Language Model (LLM), AUVs can adaptively adjust controller's parameters while optimizing RL reward functions [21], enhancing AUVs' ability to balance multi-objective optimization and improve task-oriented control and maneuvering capabilities in extreme marine environments [7], [14].

Based on above analysis, we develop an LLM-enhanced RL-based adaptive S-surface controller for AUVs to effectively execute high-level tasks in extreme sea conditions. The contributions of this paper mainly include three parts:

- We develop a novel AUV controller that employs RL to train an expert-level control strategy for high-level task execution and control command generation, while the S-surface controller produces control signals, ensuring cancellation of nonlinear effects and external disturbances under extreme sea conditions.
- We utilize LLMs for joint optimization of RL reward function and controller parameters, utilizing multimodal task execution logs and combining contextual information such as environmental descriptions to enhance the final task performance and adaptability.
- The proposed controller demonstrates superior robustness and flexibility compared to conventional PID and SMC controllers in challenging marine conditions characterized by waves, currents, and complex terrain. It exhibits exceptional performance in advanced 3D tasks, including underwater target tracking and data collection tasks.

II. RELATED WORK

A. S-Surface Controller for AUV Control

S-Surface controller and its variants leverage the principles of smooth surfaces and dynamic control, significantly enhancing AUV maneuverability and environmental disturbance responsiveness. Li *et al*. [6] implemented the controller on MOOS-IvP, demonstrating robust lake test results despite buoyancy variations. Lakhekar *et al*. [8] combined disturbance-observer-based control with fuzzy-adaptive S-Surface control for trajectory tracking, effectively compensating for disturbances without prior knowledge of uncertainty bounds. Jiang *et al*. [22] enhanced the S-Surface controller with a sliding mode variable structure to handle static load and high-speed motion, with stability confirmed by Lyapunov analysis.

B. Reinforcement Learning for Control

RL methods demonstrate promising results in controlling complex robotic systems, especially in challenging environments. Meger *et al*. [23] employed an RL-based approach to control a flipper-based underwater vehicle, using a Gaussian process model to predict state distributions. Hadi *et al*. [24] investigated RL for learning 2-DoF control (yaw, speed) in a simulator. Lu *et al*. [25] applied domain randomization to enhance RL-based control for a 4-DoF AUV. Notably, RL is often applicable to various settings without requiring in-situ tuning [26].

C. Large Language Model for Multi-Objective Optimization

LLMs excel in multi-objective optimization, serving as high-level semantic planners for robotic tasks [27], learning complex manipulation tasks, and generating structured outputs for sequential decision-making [28]. Ma *et al*. [29] showed that LLM-generated rewards outperformed human-engineered ones across various robotic tasks. Xie *et al*. [30] utilized LLMs for creating interpretable, dense reward codes, enabling iterative refinement for multi-objective tasks with human feedback. Zarzà *et al*. [31] used GPT-3.5-turbo for instantaneous PID system updates, highlighting its network control potential. Guo *et al*. [32] leveraged LLMs to encode expert knowledge, emulating human-like gradual tuning of controller parameters to meet stability requirements.

III. CONTROLLER DESIGN

In this section, we detail our proposed controller, describing its overall design architecture and explaining the workflow and principles of its three main modules.

A. Structure of the Proposed Controller

Fig. 2 illustrates the overall design of our controller. To fully leverage the advantages of the LLM-enhanced RL-based S-Surface controller, while achieving simulation and perception of extreme marine conditions to evaluate the disturbance rejection performance, we decompose the proposed framework into three core modules. Specifically, the **RL-based S-Surface Controller Module** employs RL policies focusing on high-level task decision-making, and the S-Surface controller utilized to achieve precise 6-DoF control. The **LLM-enhanced Iterative Joint Optimization Module** performs joint optimization of the RL reward function and controller parameters guided by domain-specific guidelines. It systematically analyzes environmental summaries, numerical computations, and multi-modal task feedback to enhance adaptation to dynamic marine environments. The **Simulation and Environment-Aware Module** executes physical ocean modeling with 6-DoF control dynamics for extreme scenario simulation, and fuses multisource sensor data for active disturbance mitigation.

B. RL-based S-Surface Controller Module

Through RL training, we aim to learn expert-level control policies optimized for end-to-end performance in control-constrained systems. The policies should demonstrate disturbance rejection capabilities and generate optimal reference signals for subordinate S-Surface controllers to enable AUVs to accomplish high-level tasks.

Markov decision process modeling: We define the RL training process using a Markov decision process (MDP) with control-affine dynamics, represented as the tuple $\mathcal{M} \triangleq (\mathcal{X}, \mathcal{A}, \mathcal{U}, C, f, g, d, \mathcal{R}_\pi, \gamma)$. Here, $\mathcal{X} \subseteq \mathbb{R}^n$ represents the state space, $\mathcal{A} \subseteq \mathbb{R}^a$ denotes the action space, while $\mathcal{U} \subseteq \mathbb{R}^m$

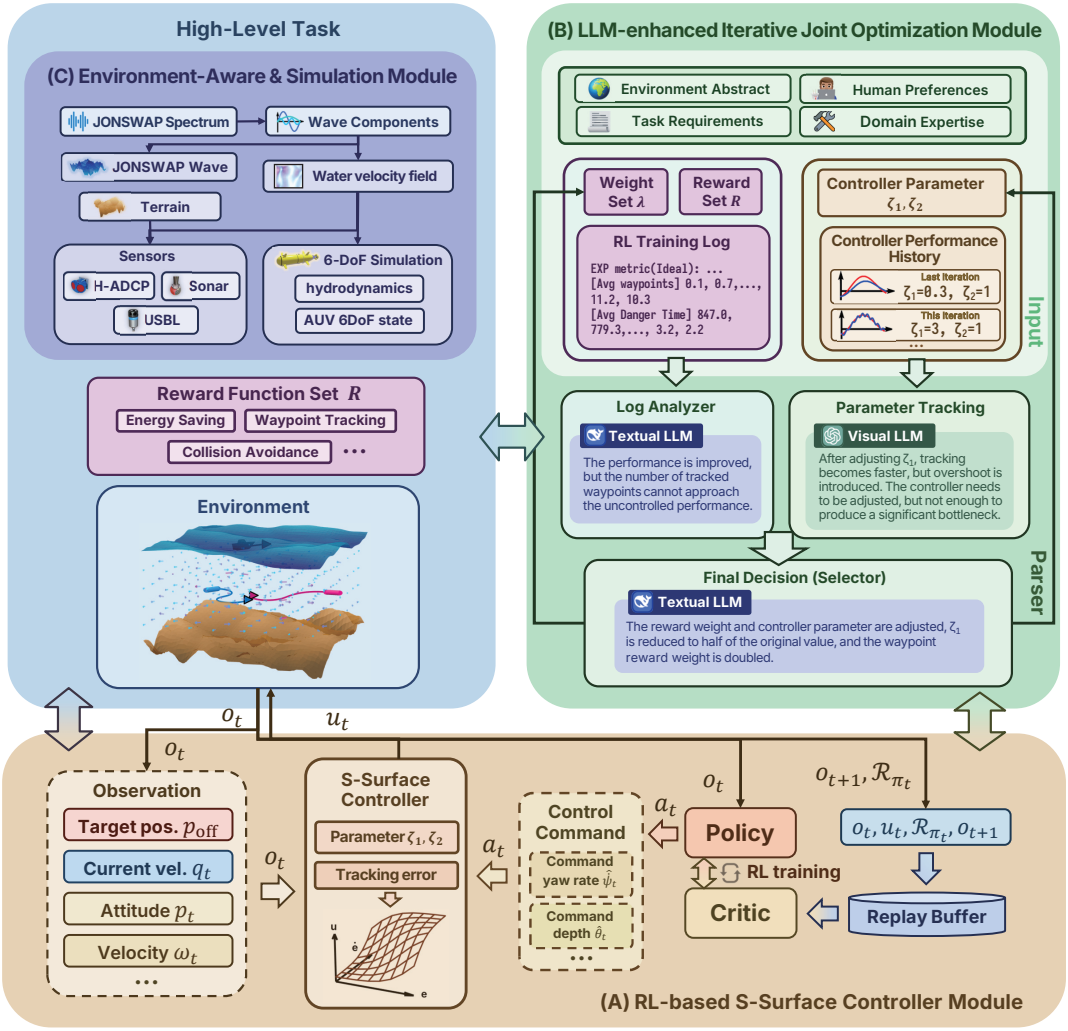


Fig. 2: **The overall framework of our proposed controller**, which comprises three modules: (A) RL-based S-Surface Controller Module. (B) LLM-Enhanced Iterative Joint Optimization Module. (C) Environment-Aware and Simulation Module.

denotes the control signal space. The state transitions in the MDP follow the control-affine system:

$$x_{t+1} = f(x_t) + g(x_t)\,C(a_t) + d(x_t)\,, \tag{1}$$

where $x_t \in \mathcal{X}$ represents the state at time step t. The high-level action signal is sampled from the distribution $\pi(a_t|x_t)$ according to a RL control policy π, and the control signal $u_t = C(x_t, a_t)$ is generated by the controller $C : \mathcal{X} \times \mathcal{A} \to \mathcal{U}$. The functions $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ characterize the known nominal model of the system. Additionally, $d : \mathbb{R}^n \to \mathbb{R}^n$ represents the unknown model component, such as environmental disturbances like ocean waves, which is continuous with respect to the state. The variable $\mathcal{R}_\pi$ denotes the reward functions, and $\gamma \in [0, 1]$ is the discount factor.

According to Eq. (1), the transition probability is represented as $P(x_{t+1}\,|x_t, a_t)$. The closed-loop transition probability under policy π is expressed as $P_\pi\,(x_{t+1}|x_t) \triangleq \int_{\mathcal{U}} \pi\,(\,a_t\,|\,x_t\,)\,P\,(\,x_{t+1}\,|\,x_t, a_t\,)\,da_t$. Furthermore, the closed-loop state distribution at time step t is denoted by $\upsilon(x_t|\rho,\ \pi,\ t)$. This distribution can be computed iteratively using the following formula $\upsilon\,(x_{t+1}|\rho,\ \pi,\ t+1) = \int_{\mathcal{X}} P_\pi(x_{t+1}|x_t)\upsilon(x_t|\rho,\ \pi,\ t)dx_t$, $\forall t \in \mathbb{N}$, with the initial condition $\upsilon(x_0|\rho,\ \pi,\ 0) = \rho$, which represents the initial state distribution.

Observations, actions, and rewards: Similar to [7], we implement a three-layer multilayer perceptron (MLP) policy, which processes an observation vector comprising both task-independent and task-relative components:

$$\vec{o} = \{\vec{p}_{\text{off}}, \vec{v}_{\text{cur}}, h_d, q_t, \vec{\omega}_t, \vec{o}_{\text{obs}}, \vec{o}_{\text{task}}\}\,, \tag{2}$$

where $\vec{p}_{\text{off}}$ denotes the positional offset between the target position and the AUV's current location, $\vec{v}_{\text{cur}}$ represents the water velocity, h_d indicates the water depth, q_t specifies the orientation quaternion, and $\vec{\omega}_t$ represents the measured angular velocities. All positional variables are defined in the AUV's body-fixed coordinate system to ensure goal-oriented control. Additionally, $\vec{o}_{\text{obs}}$ facilitate obstacle avoidance, while $\vec{o}_{\text{task}}$ represents task-specific observations (e.g. positions for other AUVs, for multi-AUV tasks).

For high-level decision-making, the policy generates reference control signals:

$$\vec{a} = [\theta_t, \dot{\psi}_t, n_t], \tag{3}$$

where θ_t denotes the target pitch angle, $\dot{\psi}_t$ represents the target yaw rate, and n_t specifies the target rotational speed of thrusters for velocity control. These reference signals enable direct comparison with observations, providing actionable inputs for S-Surface controllers.

The reward function provides performance feedback for policy optimization. To facilitate subsequent LLM-based adaptation, the study defines a weighted reward structure:

$$\mathcal{R}_\pi = \boldsymbol{\lambda}^T \boldsymbol{R} = \sum_{i=1}^{p} \lambda_i R_i, \tag{4}$$

where $\boldsymbol{R} = \{R_1, R_2, \ldots, R_p\}$ represents distinct objectives (e.g., positional accuracy, orientation control, and energy efficiency), and $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_p\}$ denotes their corresponding weights.

S-Surface Controller: The RL policy generates adaptive reference signals, requiring precise tracking by S-Surface controllers. Each S-Surface controller computes the control signal u_t based on the error e and its derivative $\dot{e}$ between the reference and actual states:

$$u_t = \frac{2}{1+\exp(-\zeta_1 e - \zeta_2 \dot{e})} - 1 + \underbrace{\Delta u}_{\text{disturbances}}, \tag{5}$$

where ζ_1 and ζ_2 are positive constants that serve as surface coefficients. The term Δu accounts for environmental disturbances identified by the environment-aware module. The S-Surface's nonlinear exponential component ensures finite-time convergence and provides a smooth control signal.

C. LLM-enhanced Iterative Joint Optimization Module

For RL-driven control systems to achieve effective performance, both the controller and the reward function must provide explicit performance feedback [21], [32], as their coupled relationship presents significant tuning challenges. To address this, we propose a joint optimization of the reward function and controller parameters. The optimization objective is formulated as follows:

$$\underset{\boldsymbol{\lambda},\zeta_1,\zeta_2}{\operatorname{argmax}} \lim_{T\to\infty} \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t \Upsilon \left(\boldsymbol{R}(\pi, \boldsymbol{\lambda}, \zeta_1, \zeta_2) \right) \right], \tag{6}$$

where $\Upsilon : \mathbb{R}^p \to \mathbb{R}$ is a utility function that maps multi-dimensional rewards to a scalar value [33]. While the scalarization process is not fixed and varies with user needs across different scenarios and over time, we maximize Υ indirectly through performance logs, hard safety constraints, and task prioritization.

Module (B) of the Fig. 2 illustrates the LLM-Enhanced Iterative Joint Optimization Module. Environmental specifications and decomposed user requirements, such as performance metrics and safety constraints, form the context. RL training logs, including performance metrics, guide reward adjustments, while signal tracking performance guides controller adjustments. However, traditional controller tuning metrics, such as settling time and phase margin, struggle to handle RL-generated reference signals characterized by high variability and noise. Therefore, we use visual signal tracking results as inputs. The LLM analyzes tracking performance across critical signal phases, such as steady-state and transients, and diagnoses issues like overshoot, sluggish response, or oscillations. Controller parameters, specifically ζ_1 and ζ_2 for the S-surface controller, are adjusted based on their physical interpretations.

To mitigate context overload in LLM reasoning, we implement a memory-augmented parameter tracking module using separate visual LLMs. This submodule processes historical parameter-performance correlations, generates comparative summaries, and determines whether it is necessary to terminate optimization if enough optimization or controller limits are detected.

For efficient joint optimization, a bottleneck-driven synchronization strategy is introduced: the system identifies whether performance limitations stem from reward function misalignment or controller inadequacy, then prioritizes adjustments to reward parameters ($\boldsymbol{\lambda}$), controller parameters (ζ_1, ζ_2), or both. And finally, the LLM generates formatted output for parameter adjustment. Besides, the reward weights will undergo preliminary tuning based on training feedback under ideal environments (allowing the RL policy to directly adjust the positions of the AUVs without control characteristics), thereby speeding up the adjustment in the controll-constraint scenarios.

D. Environment-Aware and Simulation Module

To achieve realistic 6-DoF simulation, we utilize the Python Vehicle Simulator [34] based on Fossen's motion equations [35], which is capable of simulating real-world hydrodynamic and hydrostatic forces, while providing high-level control input interfaces.

To evaluate AUV disturbance rejection, we simulate marine environments including waves and currents. The fetch-limited JONSWAP (Joint North Sea Wave Project) spectrum is adopted to represent wave energy distribution [36]:

$$S(f) = \frac{\alpha g^2}{(2\pi)^4 f^5} \exp\left(-\frac{5}{4}\left(\frac{f_p}{f}\right)^4 \right) \gamma^{\exp\left(-\frac{(f-f_p)^2}{2\sigma^2 f_p^2}\right)}, \tag{7}$$

where α denotes the energy scale parameter, f_p represents the peak frequency, γ is the peak enhancement factor, and σ is the peak shape parameter, defined as $\sigma = \sigma_a$ for $f \le f_p$ and $\sigma = \sigma_b$ for $f > f_p$. The parameter values are listed in Table I. Then, Wave surfaces are generated through linear superposition[37]:

$$\eta(x, y, t) = \sum_{i,j} a_{ij} \cos(\varphi_{ij}), \tag{8}$$

$$\varphi_{ij} = k_{ij} x \cos\theta_j + k_{ij} y \sin\theta_j - \omega_i t + \phi_{ij}, \tag{9}$$

using directional spreading function $D(\theta_j) = \cos^2\theta_j$ and phase offsets ϕ_{ij} sampled from a Gaussian process. Component amplitudes derive from $a_{ij} = \sqrt{2S(f_i)D(\theta_j)\Delta f \Delta\theta}$,

TABLE I: Key parameters of the experimental setup.

Parameters	Values
JONSWAP parameters α,f_p,γ, σ_a, σ_b	0.01,0.1,3.3,0.07,0.09
AUV maximum speed $v_{\max}$, $\omega_{\max}$	2.3m/s(4.5kts), 15deg/s(0.26rad/s)
Propeller maximum revolution	1525rpm
Water density ρ	1026kg/m^3
Control frequency	20Hz
LLM model	GPT-4o (VLLM) deepseek-V3 (Textual)
LLM parameters	temperature=0.5, Top P=1

where Δf and $\Delta\theta$ represent frequency/directional resolutions. The dispersion relation $(2\pi f)^2 = gk\tanh(kh)$ determines wave numbers k_{ij}. Horizontal wave-induced flows follow Airy theory:

$$\vec{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = \sum_{i,j} a_{ij}\omega_i \frac{\cosh\left[k_{ij}(z+h)\right]}{\sinh(k_{ij}h)} \cos(\varphi_{ij}) \begin{pmatrix} \cos\theta_j \\ \sin\theta_j \end{pmatrix}, \tag{10}$$

where h denotes the water depth. Although vertical flow disturbances are currently excluded, wave-induced coupling effects still pose challenges for 6-DoF control due to AUV motion dynamics. To address this, we design an **Environment-Aware Module**. The AUVs are equipped with horizontal acoustic Doppler current profilers (H-ADCPs) to measure water velocities and active sonar systems for terrain and obstacle detection to avoid collision. Additionally, unmanned surface vehicles (USVs) are utilized to estimate AUV positions via ultra-short baseline (USBL) acoustic positioning and facilitate inter-vehicle communication [38].

IV. EXPERIMENTS AND ANALYSIS

In the following, we first describe our simulation setup and then evaluate and analyze the adaptability and performance of our proposed LLM-enhanced RL-based adaptive S-surface controller through comprehensive experiments.

A. Experiment Setup

We validate the effectiveness of our proposed controller utilizing a REMUS 100 AUV (1.6 m in length, 31.9 kg in weight) with a maximum disturbance-free velocity of 2.3 m/s. The terrain data are derived from the East China Sea region (123°E–124°E, 28°N–29°N), and is post-processed to reduce depth variations, with the deepest water reaching 60m. Additionally, we use the TD3 as our RL algorithm with default settings [39]. Key experiment parameters and configurations are summarized in Table 1.

Within this specific setup, we introduce two high-level tasks, whose description is outlined as follows:

• **3D data collection task**: Employing the proposed controller, a single or multiple AUVs operate together to search and collect data from sensor nodes (SNs) scattered randomly. The main objectives contain conducting adaptive control of AUVs to optimize data collection rates, reducing energy consumption, and enhancing the capability to avoid collisions. We refer further details on this task to [10].

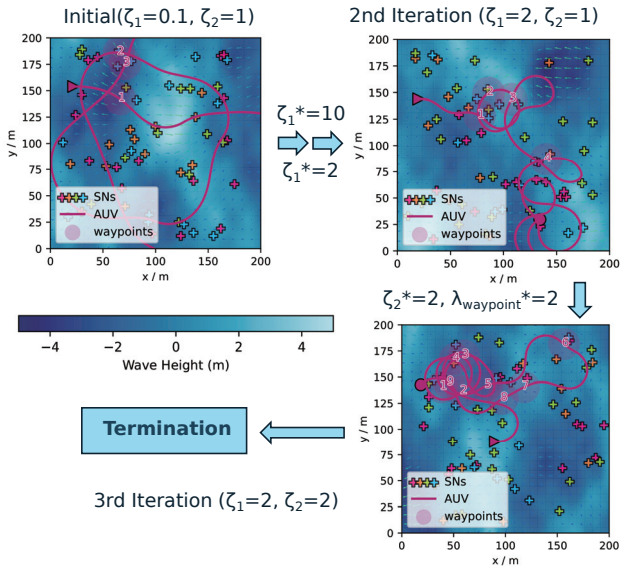


Fig. 3: Parameters for yaw tracking controller and reward weights, along with 2D projections of AUV trajectories from the 3D data collection tasks during the LLM optimization phase.

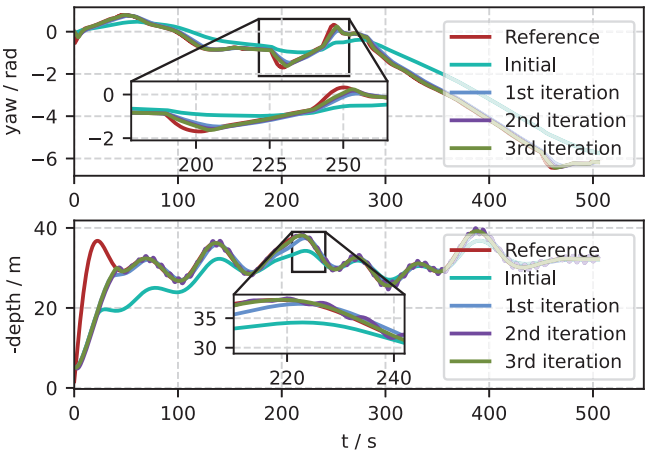
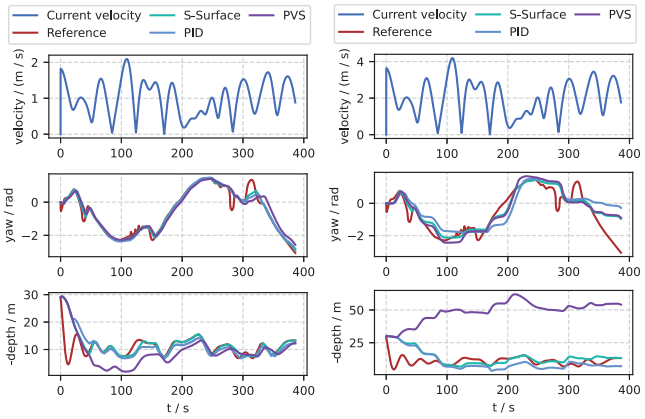


Fig. 4: Comparative results of the S-surface controller in tracking reference signals taken from a target tracking task during the LLM optimization phase.

• **3D target tracking task**: A single or multiple AUVs are utilized to follow a dynamic underwater target whose position is unpredictable. Other task objectives include avoiding collisions with hazardous terrain and obstacles, maintaining a reasonable water depth, and maintaining communication between AUVs (if applicable). We refer more details to [40].

B. Experimental Results

To evaluate the joint optimization of the LLMs, we perform parameter adjustments that the controller parameters are previously set to under-regulation configurations. The results of 3D data collection tasks executed during optimization are illustrated in Fig. 3. For comparative analysis, we also utilize the S-Surface controller to track fixed reference control signals obtained from a target tracking task during optimization, with the comparative tracking performance shown in Fig. 4.



(a) ES condition (b) VES condition

Fig. 5: Comparative results of three controllers tracking reference signals taken from a target tracking task under ES and VES conditions.

For yaw control, a low ζ_1 parameter value results in a significantly slow response. Consequently, the LLM substantially increases ζ_1 during the first iteration. In the second iteration, while continuing to increase ζ_1, the adjustment magnitude is reduced due to improved tracking performance. By the third iteration, residual inadequate regulation in yaw control persists when the reference signal shows rapid change, indicating system steering limitations. The LLM responds by increasing ζ_2 before terminating the iteration to enhance stability. Concurrently, it enhances the reward weight for waypoint tracking, as the AUV struggles with flexible and accurate waypoint tracking in practical tasks. For depth control, high-frequency oscillation prompts the LLM to reduce ζ_1 while increasing ζ_2 before termination.

Also, We conduct comparative experiments between the LLM-optimized S-Surface controller and baseline controllers, with their parameters also been optimized by the LLM process mentioned before, including:

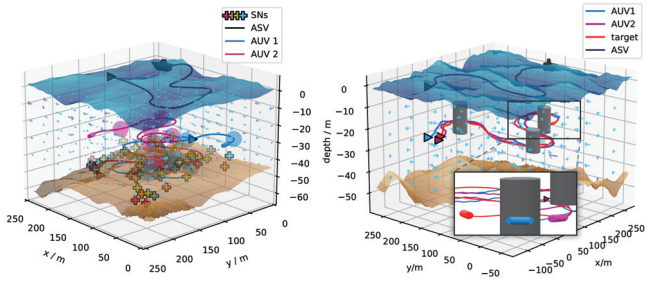
- **PID**: Conventional PID controllers for separate yaw and depth control.
- Original control implementation from Python Vehicle Simulator (denoted as **PVS**): The PVS employed a SMC controller with reference model compensation for yaw control and a PI controller for depth control.

These controllers are evaluated under two disturbance conditions: the extreme sea condition (**ES**) with a maximum water velocity of 2 m/s, and the very extreme sea condition (**VES**) with doubled water velocities (maximum 4 m/s), exceeding the AUV's maximum propulsion capability and requiring advanced compensation strategies.

Comparative results illustrated in Fig. 5 demonstrate the S-Surface controller's superior adaptability under disturbances. Specifically, both the PID and S-Surface controllers achieve stable reference tracking under ES conditions, while the PVS exhibits delayed yaw control response due to the inherent phase lag of its SMC controller with reference model architecture, along with significant depth overshoot from its basic PI controller. When transitioning to VES conditions, the controllers exhibit progressive performance deterioration, with the PID controller showing worse flexibility and stability compared to the S-Surface controller. Additionally, the PVS suffers a complete loss of depth regulation capability.

TABLE II: Performance metrics of different control methods evaluated during the data collection task under ES and VES conditions.

Metrics		SSN ↑	EC (W) ↓	DT (s) ↓
Ideal		15.7 ± 6.4	163.8 ± 32.0	0.0 ± 0.0
S-Surface	ES	14.0 ± 8.7	202.9 ± 38.5	0.0 ± 0.0
	VES	10.7 ± 7.0	227.2 ± 44.5	34.4 ± 20.7
PID	ES	13.8 ± 9.0	194.8 ± 41.9	0.0 ± 0.0
	VES	9.2 ± 6.1	231.3 ± 46.8	53.7 ± 23.3
PVS	ES	12.3 ± 8.4	205.1 ± 35.2	203.0 ± 98.8
	VES	6.5 ± 5.4	247.1 ± 56.3	517.8 ± 157.9



(a) 3D Data collection task (b) 3D target tracking task

Fig. 6: 3D visualizations of multiple AUVs performing data collection and target tracking tasks using the proposed RL-based S-surface controller.

Then, we utilize the controllers above to conduct 3D data collection task to evaluate the task-specific performance. We introduce three metrics: the total number of served sensor nodes (**SSN**, quantifying yaw control capability), energy consumption (**EC**, measuring actuation efficiency, calculated using equations from [41]), and danger time (**DT**, representing the cumulative duration of unsafe seafloor proximity below 10 m, quantifying depth control capability). An idealized control setting (**Ideal**) is additionally introduced, which removes hydrodynamic limitations and the RL policy can directly change the AUVs' positions. The results are presented in Table II. Under the ES condition, the S-Surface controller achieves performance close to the ideal setting, while the PVS exhibits significantly longer danger time due to poor depth control. Under the VES condition, the PID controller exhibits significantly greater performance degradation compared to the S-Surface controller, while the PVS experiences serious control failure.

Finally, Fig. 6 visualizes the 3D data collection and target tracking tasks performed by two AUVs utilizing S-Surface control. In the former case, the AUV must judiciously control its direction to efficiently serve sensor nodes due to its restricted tuning capability, while the latter requires more real-time control capabilities. Thanks to the powerful optimization capability of RL and the flexible execution of

the controller, the AUVs can plan optimal routes as much as possible, achieving performance close to ideal control conditions. In the latter case, the AUVs also demonstrate high maneuverability in response to target turns.

V. CONCLUSIONS

In this study, we develop an LLM-enhanced RL-based adaptive S-surface controller for AUVs under extreme sea conditions. This controller utilizes LLMs to iteratively optimize controller parameters and reward functions, while leveraging RL to train the AUV to acquire an expert-level control strategy. The strategy autonomously generates control commands for S-surface controllers in high-level tasks, which further convert them into low-level control signals. Comprehensive simulation experiments on representative high-level tasks demonstrate the superior performance and adaptability of the proposed controller, which outperforms PID and SMC controllers under extreme sea conditions. Future work will focus on implementing the proposed controller on AUVs and conducting field experiments to realize the sim2real process, aiming to minimize the gap between simulation and reality.

VI. ACKNOWLEDGEMENT

The authors gratefully acknowledge the anonymous reviewers for their constructive comments. We also extend our sincere thanks to Dr. Xiangwang Hou, Prof. Yong Ren, Prof. Daoyi Chen, and Prof. Juntian Qu from Tsinghua University for their insightful discussions and guidance. Additionally, we appreciate the encouragement and recognition from Prof. Nare Karapetyan at the Woods Hole Oceanographic Institution, Prof. Xiaofan Li at the University of Hong Kong, and Prof. Xiaomin Lin at the University of South Florida.

REFERENCES

[1] L. Hawkes, O. Exeter, S. Henderson, C. Kerry, A. Kukulya, J. Rudd, S. Whelan, N. Yoder, and M. Witt, "Autonomous underwater videography and tracking of basking sharks," *Animal Biotelemetry*, vol. 8, 08 2020.

[2] J. Rutledge, W. Yuan, J. Wu, S. Freed, A. Lewis, Z. Wood, T. Gambin, and C. Clark, "Intelligent shipwreck search using autonomous underwater vehicles," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6175–6182.

[3] F. L. Peña, F. Orjales, and A. Deibe, "Development of a collaborative host-guest unmanned underwater vehicle docking system for inspection and maintenance of offshore structures," in *OCEANS 2023 - MTS/IEEE U.S. Gulf Coast*, 2023, pp. 1–5.

[4] D. Wei, C. Huang, X. Li, B. Lin, M. Shu, J. Wang, and M. Pan, "Power-efficient data collection scheme for auv-assisted magnetic induction and acoustic hybrid internet of underwater things," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 11 675–11 684, 2022.

[5] G. V. Lakhekar, L. M. Waghmare, and R. G. Roy, "Disturbance observer-based fuzzy adapted s-surface controller for spatial trajectory tracking of autonomous underwater vehicle," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 4, pp. 622–636, 2019.

[6] B. Li, Y. Xu, C. Liu, and W. Xu, "Simulation and preliminary experimental results on s-surface control of an autonomous underwater vehicle based on moos-ivp," in *2014 Oceans - St. John's*, 2014, pp. 1–6.

[7] L. Cai, K. Chang, and Y. Girdhar, "Learning to swim: Reinforcement learning for 6-dof control of thruster-driven autonomous underwater vehicles," 2024. [Online]. Available: https://arxiv.org/abs/2410.00120

[8] A. R. Thomas, L. P. P S, and H. K. R, "Disturbance estimation and rejection in an underwater autonomous vehicle," in *2024 International Conference on E-mobility, Power Control and Smart Systems (ICEMPS)*, 2024, pp. 1–6.

[9] X. Lin, N. Karapetyan, K. Joshi, T. Liu, N. Chopra, M. Yu, P. Tokekar, and Y. Aloimonos, "Uivnav: Underwater information-driven vision-based navigation via imitation learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5250–5256.

[10] Z. Zhang, J. Xu, G. Xie, J. Wang, Z. Han, and Y. Ren, "Environment- and energy-aware auv-assisted data collection for the internet of underwater things," *IEEE Internet of Things Journal*, vol. 11, no. 15, pp. 26 406–26 418, 2024.

[11] C.-W. Chen, N.-M. Yan, J.-X. Leng, and Y. Chen, "Numerical analysis of second-order wave forces acting on an autonomous underwater helicopter using panel method," in *OCEANS 2017 - Anchorage*, 2017, pp. 1–6.

[12] A. Mitchell, E. McGookin, and D. Murray-Smith, "Comparison of control methods for autonomous underwater vehicles," *IFAC Proceedings Volumes*, vol. 36, no. 4, pp. 37–42, 2003, iFAC Workshop on Guidance and Control of Underwater Vehicles 2003, Newport, South Wales, UK, 9-11 April 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474667017366545

[13] B. Hu, H. Tian, J. Qian, G. Xie, L. Mo, and S. Zhang, "A fuzzy-pid method to improve the depth control of auv," in *2013 IEEE International Conference on Mechatronics and Automation*, 2013, pp. 1528–1533.

[14] Z. Yan, P. Gong, W. Zhang, and W. Wu, "Model predictive control of autonomous underwater vehicles for trajectory tracking with external disturbances," *Ocean Engineering*, vol. 217, p. 107884, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801820308490

[15] S. D. Joshi and D. B. Talange, "Integer & fractional order pid controller for fractional order subsystems of auv," in *2013 IEEE Symposium on Industrial Electronics & Applications*, 2013, pp. 21–26.

[16] Y. Xu, "S control of automatic underwater vehicles," *Ocean Engineering*, 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:112961246

[17] L. Ji-qing and W. Lei, "The heel and trim adjustment of manned underwater vehicle based on variable universe fuzzy s surface control," in *2011 International Conference on Electronics, Communications and Control (ICECC)*, 2011, pp. 1122–1125.

[18] A. Loquercio, E. Kaufmann, R. Ranftl, A. Dosovitskiy, V. Koltun, and D. Scaramuzza, "Deep drone racing: From simulation to reality with domain randomization," *IEEE Transactions on Robotics*, vol. 36, no. 1, pp. 1–14, 2020.

[19] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:58031572

[20] Z. Wang, H. Huang, J. Tang, and L. Hu, "A deep reinforcement learning-based approach for autonomous lane-changing velocity control in mixed flow of vehicle group level," *Expert Syst. Appl.*, vol. 238, no. PD, Mar. 2024. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.122158

[21] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *ICLR*, 2024.

[22] C. Jiang, J. Lv, L. Wan, J. Wang, B. He, and G. Wu, "An improved s-plane controller for high-speed multi-purpose auvs with situational static loads," *Journal of Marine Science and Engineering*, vol. 11, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2077-1312/11/3/646

[23] D. Meger, J. C. G. Higuera, A. Xu, P. Giguère, and G. Dudek, "Learning legged swimming gaits from experience," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2332–2338.

[24] B. Hadi, A. Khosravi, and P. Sarhadi, "Deep reinforcement learning for adaptive path planning and control of an autonomous underwater vehicle," *Applied Ocean Research*, vol. 129, p. 103326, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141118722002589

[25] W. Lu, K. Cheng, and M. Hu, "Reinforcement learning for autonomous underwater vehicles via data-informed domain randomization," *Applied Sciences*, vol. 13, no. 3, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/3/1723

[26] A. Molchanov, T. Chen, W. Hönig, J. A. Preiss, N. Ayanian, and G. S. Sukhatme, "Sim-to-(multi)-real: Transfer of low-level robust control policies to multiple quadrotors," in *2019*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE Press, 2019, p. 59–66. [Online]. Available: https://doi.org/10.1109/IROS40897.2019.8967695

[27] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "PROGPROMPT: Generating situated robot task plans using large language models," in *International Conference on Robotics and Automation (ICRA)*, may 2023.

[28] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9493–9500.

[29] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2210.00030

[30] T. Xie, S. Zhao, C. H. Wu, Y. Liu, Q. Luo, V. Zhong, Y. Yang, and T. Yu, "Text2reward: Reward shaping with language models for reinforcement learning," in *International Conference on Learning Representations*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:270063618

[31] I. de Zarzà, J. de Curtò, G. Roig, and C. T. Calafate, "Llm adaptive pid control for b5g truck platooning systems," *Sensors*, vol. 23, no. 13, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/13/5899

[32] X. Guo, D. Keivan, U. Syed, L. Qin, H. Zhang, G. Dullerud, P. Seiler, and B. Hu, "Controlagent: Automating control system design via novel integration of llm agents and domain expertise," 2024. [Online]. Available: https://arxiv.org/abs/2410.19811

[33] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, p. 26, 2022.

[34] T. I. Fossen, "Python vehicle simulator," 2021. [Online]. Available: https://github.com/cybergalactic/PythonVehicleSimulator

[35] T. I. Fossen, *Handbook of marine craft hydrodynamics and motion control*. John Wiley & Sons, 2011.

[36] Y. Goda, "A comparative review on the functional forms of directional wave spectrum," *Coastal Engineering Journal*, vol. 41, no. 1, pp. 1–20, 1999. [Online]. Available: https://doi.org/10.1142/S0578563499000024

[37] R. G. Dean and R. A. Dalrymple, *Water wave mechanics for engineers and scientists*. world scientific publishing company, 1991, vol. 2.

[38] J. Xu, G. Xie, X. Wang, Y. Ding, and S. Zhang, "Usv-auv collaboration framework for underwater tasks under extreme sea conditions," *arXiv preprint arXiv:2409.02444*, 2024.

[39] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*, 2018, pp. 1582–1591.

[40] J. Xu, G. Xie, Z. Zhang, X. Hou, D. Ma, S. Zhang, Y. Ren, and D. Niyato, "Is fisher all you need in the multi-auv underwater target tracking task?" *arXiv preprint arXiv:2412.03959*, 2024.

[41] D. Steinberg, A. Bender, A. Friedman, M. Jakuba, O. Pizarro, and S. Williams, "Analysis of propulsion methods for long-range auvs," *Marine Technology Society Journal*, vol. 44, no. 2, pp. 46–55, 2010.