arXiv:2602.07947v1 [cs.NI] 8 Feb 2026

DHEA-MECD：一种基于具身智能、用于高维特征水下环境中自主水下航行器跟踪的深度强化学习算法

Kai Tian、Chuan Lin，*IEEE*会员，Guangjie Han，*IEEE*会士，Chen An、Qian Zhu、Shengchao Zhu、Zhenyu Wang

摘要—近年来，自主水下航行器（AUV）系统在复杂海洋探索中展现出巨大潜力。然而，在具有高维特征的真实水下环境中，有效的自主水下航行器跟踪仍面临挑战，这些特征包括耦合的运动学状态、空间约束以及时变环境扰动等。为应对这些挑战，本文提出了一种分层具身智能（EI）架构，用于复杂水下环境中的自主水下航行器多目标跟踪。在此架构基础上，我们提出双头编码器-注意力多专家协同决策（DHEA-MECD），这是一种旨在支持高效、稳健多目标跟踪的新型深度强化学习（DRL）算法。具体而言，DHEA-MECD设计了一种双头编码器-注意力信息提取框架，对原始感知观测进行语义解构，并显式建模空间构型、运动学状态、结构约束和随机扰动等异构特征间的复杂依赖关系。在此基础上，引入一种采用Top-k 专家选择策略的运动阶段感知型多专家协同决策机制，以支持阶段自适应决策。此外，我们提出一种基于DHEA-MECD的水下多目标跟踪算法，实现自主水下航行器智能、稳定且抗干扰的多目标跟踪。大量实验结果表明，与主流的基于深度强化学习的方法相比，所提出的方法取得了更高的跟踪成功率、更快的收敛速度和更优的运动最优性，尤其是在复杂且扰动丰富的海洋环境中。

索引词—自主水下航行器、深度强化学习、具身智能、多专家协同决策、水下多目标跟踪。

I. 引言

AQUATIC 资源 [1]是人类生存和社会发展的根本，因此对其进行可持续探索和保护已成为全球关注的重大议题。随着水下通信、感知和机器人技术的迅速发展， [2], [3],自主水下航行器已成为执行各类海事任务的主要移动计算与感知代理，任务范围从环境监测到高风险资源侦察 [4]。在此类任务中，持续监测动态实体（如生物集群或漂移异常）的能力至关重要。因此，自主多目标跟踪是一项关键能力 [5], [6], ，要求自主水下航行器持续监视水下空间中的动态目标。

通讯作者：*Guangjie Han*Kai Tian、Chuan Lin、Qian Zhu 和 Zhenyu Wang 就职于中国沈阳东北大学软件学院。（电子邮箱：kaitian901@gmail.com；chuanlin1988@gmail.com；zhuq@swc.neu.edu.cn；larrywang1019@outlook.com）。Guangjie Han 和 Shengchao Zhu 就职于河海大学教育部海事智能网络信息技术重点实验室。（电子邮箱：hanguangjie@gmail.com；zhushengchao77@gmail.com。）Chen An 就职于中国沈阳东北大学计算机科学与工程学院。（电子邮箱：20235854@stu.neu.edu.cn）。

从根本上说，此类自主跟踪需要持续环境感知与精确智能体控制的紧密结合 [7], [8]。在此背景下，具身智能已成为一种前景广阔的新兴范式，可用于解决复杂的智能体控制问题 [9], ，通过整合感知与执行来抵御水下干扰。在具身智能范式中，深度强化学习被广泛用作自主智能体的核心决策框架 [10]。然而，传统深度强化学习算法应用于复杂海洋环境时，往往难以兼顾鲁棒性与适应性；这类环境充斥着不可预测的干扰和噪声感知数据。因此，由于存在两大技术瓶颈，实现精确的多目标跟踪仍然是一项挑战：

1）高维特征异质性与感知碎片化：水下观测天然具有高维且异质的特点 [11],，涵盖三维空间几何、运动学状态，以及时空洋流和多源声学噪声等复杂环境变量。传统深度强化学习算法通常将这些输入作为整体扁平向量处理 [12], ，无法揭示子特征之间内在的语义相关性。缺乏结构归纳偏置会导致“感知碎片化”，使智能体难以从冗余且含噪的高维数据中提取与任务相关的信号 [13],，最终造成策略不稳定或无法收敛。

2）动态多阶段任务中的决策适应性不足：水下环境具有显著的非平稳性，要求自主水下航行器频繁切换不同的运动模式，例如快速追踪、精确定点保持和紧急避碰 [14], [15]。传统深度强化学习架构通常依赖参数固定的单一策略网络。然而，单一策略往往难以兼顾高速机动的爆发性与精细控制的精确性 [16], ，尤其是在处理自适应水下追踪所需的混合动作空间（离散动作与连续动作相结合） [17] 时。

受深度强化学习驱动的具身智能架构在复杂机器人控制领域取得的变革性成功启发，本文提

出了一种由三个集成层构成的分层 EI 架构：具身感知层、具身决策层和具身执行层。不同于传统的单体深度强化学习，该架构将高层战略跟踪逻辑与低层流体动力学驱动解耦。在此 EI 架构基础上，本文提出**双头编码器注意力多专家协同决策（DHEA-MECD）**算法，作为自主水下跟踪的认知核心。所提出的 DHEA-MECD算法将双头编码器注意力机制与多专家协同决策机制相结合，以确保稳健而精准的跟踪。具体而言，该算法采用专用多头编码器和自注意力机制，对异构状态子空间进行语义分解与融合，从而促进基于高维特征的稳健表示学习。在此基础上，由 Top-k 专家选择策略驱动的运动阶段感知决策机制，可在专用专家之间动态切换，确保在不同跟踪阶段做出精准决策。该算法显著提升了机动性与稳定性，同时降低了高维特征水下环境中的计算开销。

本文的主要贡献总结如下：

- 我们为自主水下航行器提出了一种分层 EI 架构，通过将任务（例如多目标跟踪）解耦为专门的层级，重新定义了从感知到决策的流程，从而增强了系统在具有高维特征的水下环境中的鲁棒性。
- 基于所提出的 EI 架构，我们提出了 DHEA-MECD算法，该算法结合了双头编码器-注意力机制与多专家协同决策机制。它采用专门设计的编码器和自注意力机制来提炼任务所需的关键特征，并通过 top-k 专家选择策略，实现更高的效率和精度。
- 我们提出了一种基于 DHEA-MECD 的多目标跟踪算法，专为具有高维特征的水下环境中的自主水下航行器设计。评估结果表明，所提出的算法能够实现平稳、鲁棒且精准的跟踪，性能优于其他主流算法。

本文其余部分的组织结构如下：第II节综述相关工作；第III 节介绍预备知识；第IV节描述所提出的分层 EI 架构；第V 节介绍所提出的 DHEA-MECD算法；第VI 节提出一种基于 DHEA-MECD 的多目标跟踪算法；第VII 节给出评估结果；第VIII 节总结全文。

II. 相关工作

本节综述相关领域的最新进展，重点关注三个方面：1）基于自主水下航行器的水下目标跟踪；2）面向自主水下航行器的基于深度强化学习的决策；3）面向机器人的基于具身智能的决策。

A. 基于自主水下航行器的水下目标跟踪

自主水下航行器在水下目标跟踪中发挥着至关重要的作用。为了在复杂的流体动力学环境中可靠运行，同时保持与移动目标的精确接近，自主水下航行器需要稳健的控制策略。在 [19], 中，提出了一种数据驱动的滚动时域控制策略，用于优化声呐网络内目标跟踪任务中的自主水下航行器航向。通过集成声学感知模块，该控制器能够自主调整自主水下航行器的轨迹，以最大化信噪比，从而显著降低目标估计误差，并增强其在噪声环境中的跟踪稳定性。在 [20],中，作者提出了一种软件定义控制架构，以提升自主水下航行器的自适应跟踪能力。该框架将控制逻辑与航行器执行过程解耦，使自主水下航行器能够通过增量学习机制动态优化其跟踪策略。该方法无需模型重新训练，即可在保持较高运行灵活性的同时，确保持续稳定的目标监视。在 [21], 中，提出了一种基于行为决策的路径规划方法（PPM-BBD），用于优化自主水下航行器在下潜与接近阶段的轨迹。通过将角度修正策略融入 L-SHADE 进化算法，该方法确保生成的所有跟踪路径严格遵守自主水下航行器的物理机动性限制。这种控制感知规划可节省超过 9% 的能源，有助于实现高效、长时间的跟踪任务。

总体而言，上述传统的基于控制的策略能够有效地在理想环境中调度自主水下航行器。然而，它们往往难以适应高度随机的水下环境。为应对这些适应性挑战，研究人员日益转向深度强化学习，以实现端到端决策。

B. 基于深度强化学习的自主水下航行器决策

深度强化学习已成为自主水下航行器决策的有力工具，使自主航行器能够在高度非线性且不确定的水下环境中学习最优控制策略。在 [22], 中，提出了一种ACL-SAC算法，用于优化自主水下航行器的偏航和推进力，以实现动态跟踪。该框架将注意力卷积长短期记忆网络（ACL）与软演员-评论家（SAC）相结合，增强了自主水下航行器抵御洋流和声学不确定性的能力。与传统方法相比，这种深度强化学习方法能够快速收敛，并更好地适应随机环境扰动。在 [23], 中，开发了一种可快速部署的基于DDPG的控制器，用于欠驱动X型舵自主水下航行器的姿态控制和轨迹跟踪。该方法有效解决了临界偏航角下的控制失效问题，实现了精确的三自由度跟踪控制（俯仰角、偏航和速度）。这种深度强化学习驱动的策略便于快速部署，并能稳定跟踪各种方位的目标。在 [24], 中，提出了一种中断式软件定义深度强化学习架构（ISD-MARL），以加快自主水下航行器的决策。该框架采用S-MADDPG算法和动作优化模型，显著

缩短了策略收敛时间。结合中断式路径规划方案，该方法使自主水下航行器能够在动态水下场景中兼顾跟踪精度和时间效率。

总之，深度强化学习显著提升了自主水下航行器应对环境不确定性的能力。然而，大多数深度强化学习模型缺乏对物理交互和环境推理的深入理解。这一局限促使研究者探索具身智能，以更好地实现物理形态与认知过程的协同。

C. 基于具身智能的机器人决策

具身智能将机器人范式从纯数据驱动学习转向物理交互，使机器人能够在结构化和非结构化环境中进行推理与决策。在 [25]，文中，研究者提出了一种仿生机器人控制器，将环境感知、自主规划和运动控制集成到统一框架中。通过部署 YOLOv5_OBB 网络进行目标识别，并采用改进的 RRT-Growth-Angle算法进行轨迹规划，该控制器能够模拟人类的决策方式，完成多目标重排等复杂任务。这种模块化集成确保了在定制化制造场景中的双机器人精确协调与无碰撞执行。在 [26]，该研究中，作者通过开发用于手术机器人学习的交互式仿真平台，探究了人类参与的具身智能。研究强调了人类示范在通过强化学习改进策略学习方面的作用。该框架借助物理输入设备促进高质量的人机交互，使手术机器人能够更高效地学习复杂控制策略，展现了具身智能在安全关键和高精度领域的应用潜力。在 [27],这项研究凸显了具身智能如何通过弥合数字数据与物理行动之间的鸿沟，重新定义机器人操作。凭借增强的推理与泛化能力，具身智能使机器人能够应对非结构化环境的复杂性。最终，作者指出，具身智能是通用人工智能的基础构件，为在实际场景中部署智能自主系统提供了必要的理论基础。这些进展表明，具身智能能够连接数字推理与物理行动。然而，现有具身智能框架往往训练成本高昂，并且忽视了水下运动的具体约束。为弥合这一差距，我们提出了一种用于水下智能跟踪的分层 EI 架构，以提升其在高维特征水下环境中的效率和适应性。

III. 预备知识

本节介绍与本文相关的预备知识，包括洋流扰动建模、水下噪声，以及用于基于自主水下航行器的多目标跟踪的马尔可夫决策过程（MDP）。

A. 洋流扰动建模

为刻画环境流的时空变化，洋流速度场表示为一个基于高斯径向基函数（RBFs）的有限维线性参数化模型。每个基函数 $\phi_\ell(r)$ 定义如下：

$$\phi_\ell(r) = \exp\left(-\frac{\|r-\mu_\ell\|^2}{2\lambda_\ell^2}\right), \quad \ell = 1,\dots,L, \tag{1}$$

其中，$r \in \mathbb{R}^3$ 表示空间坐标，而 $\mu_\ell \in \mathbb{R}^3$和 $\lambda_\ell > 0$ 分别表示第 ℓ个基函数的中心和空间尺度。

位置 r 处、时间 t的海流速度 $v_c(r,t)$表示为：

$$v_c(r,t) = \Phi(r)\,\theta(t), \tag{2}$$

其中，$\theta(t) \in R^{3L}$ 为时变权重向量，回归矩阵 $\Phi(r) = [\phi_1(r)I_3,\dots,\phi_L(r)I_3] \in R^{3\times 3L}$ 由径向基函数构造，其中 I_3 表示 3×3 单位矩阵。

环境流会在自主水下航行器与周围流体之间引起相对运动。令 $p_a(t) \in R^3$ 表示自主水下航行器在惯性坐标系中的位置，$v_a(t) = \dot{p}_a(t)$表示其对应的惯性速度。相对于环境流的相对速度 $v_r(t)$定义如下：

$$v_r(t) = v_a(t) - \Phi\big(p_a(t)\big)\,\theta(t), \tag{3}$$

其中，$\Phi\big(p_a(t)\big)\,\theta(t)$表示在自主水下航行器当前位置处评估得到的局部流速。

相应地，自主水下航行器的平动动力学满足以下关系：

$$M\dot{v}_a(t) = f_u(t) + f_h\big(v_r(t)\big) + f_d(t), \tag{4}$$

其中，$M \in R^{3\times3}$ 为惯性矩阵，$f_u(t)$表示控制输入，$f_h(\cdot)$表示取决于相对速度的水动力，$f_d(t)$表示未建模动力学和外部扰动。

B. 水下噪声建模

瞬时水下声噪声 $n(t)$源于多个同时存在的噪声源的汇合，包括航行器诱导噪声 $n_{\text{veh}}(t)$、生物噪声 $n_{\text{bio}}(t)$、地质噪声 $n_{\text{geo}}(t)$以及湍流诱导噪声 $n_{\text{turb}}(t)$。总噪声 $n(t)$可表示为：

$$n(t) = \big[n_{\text{veh}}(t),\ n_{\text{bio}}(t),\ n_{\text{geo}}(t),\ n_{\text{turb}}(t)\big]^\top. \tag{5}$$

航行器诱导噪声 $n_{\text{veh}}(t)$源于水下平台固有的机械振动和水动力相互作用。这些现象通常表现为宽带波动 [28]，，其功率谱密度随运行强度增加而增大。为进行算法评估，我们将 $n_{\text{veh}}(t)$建模为零均值高斯过程：

$$n_{\text{veh}}(t) \sim \mathcal{N}\big(0,\sigma_{\text{veh}}^2\big), \tag{6}$$

其中，方差 σ_{veh}^2 为

$$\sigma_{\text{veh}}^2 = \sigma_{\text{veh},0}^2 \left(\frac{v_s}{v_0}\right)^{\alpha_v} \left(\frac{\rho_N}{\rho_0}\right)^{\alpha_\rho}, \tag{7}$$

其中，$\sigma_{\text{veh},\,0}^2$、$v_0$和 ρ_0 是参考值，而 α_v，$\alpha_\rho > 0$控制速度和密度的敏感度。

生物噪声 $n_{\text{bio}}(t)$ 源于海洋动物的发声和脉冲性声学事件，通常呈现重尾、非高斯特性。为捕捉这种脉冲性，我们采用对称的 α-稳定（$S\alpha S$）分布 [29]:

$$n_{\text{bio}}(t) \sim S\alpha S(\alpha, \gamma, 0), \tag{8}$$

其中，$\alpha \in (1, 2]$ 和 $\gamma > 0$ 分别表示脉冲性和尺度参数。

地质噪声 $n_{\text{geo}}(t)$以低频成分为主，并呈现显著的时间相关性。因此，我们将 $n_{\text{geo}}(t)$建模为通过对白高斯激励进行滤波而得到的有色高斯噪声:

$$\begin{cases} n_{\text{geo}}(t) = (h_{\text{geo}} \circledast w)(t); \\ w(t) \sim \mathcal{N}(0, 1), \end{cases} \tag{9}$$

其中，h_{geo} 是一个成形滤波器，$\circledast$ 表示卷积。为确保低频占主导地位，选择 h_{geo}，使所得功率谱密度遵循幂律谱:

$$|H_{\text{geo}}(\omega)|^2 \propto \omega^{-\beta_{\text{geo}}}, \tag{10}$$

其中，$H_{\text{geo}}(\omega)$是 h_{geo} 的频率响应，β_{geo}控制谱滚降。

湍流诱导噪声 $n_{turb}(t)$源于流动引起的传感器节点及结构上的压力波动和涡旋波动，导致幅度随时间发生相关变化 [30]。在离散时间索引 t处，我们将 $n_{turb}(t)$建模为强度随时间变化的随机过程:

$$\begin{cases} n_{\text{turb}}(t) = \rho\, n_{\text{turb}}(t-1) + \sigma_{\text{turb}}\big(U_{\text{turb}}(t)\big)\, \epsilon(t); \\ \epsilon(t) \sim \mathcal{N}(0, 1); \\ \sigma_{\text{turb}}\big(U_{\text{turb}}(t)\big) = \sigma_{\text{turb},0} \left(\frac{U_{\text{turb}}(t)}{U_0}\right)^{\alpha_u}, \end{cases} \tag{11}$$

其中，$|\rho| < 1$ 控制时间相关性，$U_{\text{turb}}(t)$表示（可能随时间变化的）湍流速度或等效的强度指标，而$\sigma_{\text{turb},\,0}$，U_0，$\alpha_u > 0$ 为参考参数和敏感度参数。具体而言，$U_{\text{turb}}(t)$在每个处理窗口内被视为缓慢变化，从而得到局部平稳模型。

C. 多目标跟踪的马尔可夫决策过程

本文将具有高维特征的水下环境中的基于自主水下航行器的多目标跟踪问题建模为马尔可夫决策过程，正式定义为 $\mathcal{M} = (\mathcal{S},\ \mathcal{A},\ P,\ R,\ \gamma)$，其中 $\mathcal{S}$ 为状态空间，$\mathcal{A}$ 为混合动作空间，P表示状态转移概率，R 表示奖励函数，$\gamma \in [0, 1)$为折扣因子。

状态空间 S：状态包含自主水下航行器决策过程所需的全部必要信息，包括自主水下航行器的空间信息（位置、速度和姿态）、目标信息（目标位置、目标间距离）、障碍信息（洋流、障碍物状态）以及水下噪声（航行器诱导噪声、生物噪声、地质噪声和湍流诱导噪声）

动作空间 $\mathcal{A}$：自主水下航行器跟踪在混合动作空间 $\mathcal{A} \triangleq \mathcal{D} \times \mathcal{C}$ 中实现，其中每个复合动作 $a_t = (d_t, c_t)$ 将离散动作 $d_t \in \mathcal{D}$ 与连续动作 $c_t \in \mathcal{C}$ 相结合。这里，d_t 用于选择离散动作（例如，加速度、偏航、转向或制动），而$c_t = [\dot{v}_t, \delta_{\psi,t}, \delta_{\theta,t}, b_t]^\top$ 则通过精确的控制量幅值对所选离散动作进行细化，包括加速度幅值、姿态角速度和制动力。

状态转移概率 P：状态转移动态由转移概率 $P(s_{t+1}|s_t,\ a_t)$描述，该概率刻画了自主水下航行器-目标状态的随机演化。

奖励函数 R：奖励函数 R 旨在确保精确的目标跟踪，具体细节将在第VI-A节介绍。

折扣因子 γ：折扣因子 $\gamma \in$ （0,1）用于平衡即时与长期性能，使自主水下航行器能够在跟踪任务中实现稳定且鲁棒的跟踪行为。

第四节 面向水下智能跟踪的分层具身智能架构

本文提出一种分层 EI 架构（如图1所示），以促进动态水下环境中多目标跟踪的精确且高效的决策。该架构通过协同实时感知与自适应决策，即使面临高维不确定性与严重的环境扰动，也能赋予自主水下航行器持续监视和精确的近距离接近能力。

所提出的架构包括具身感知层、具身决策层和具身执行层:

具身感知层：为应对复杂水下环境中的感知挑战，该层旨在构建鲁棒且高保真的环境状态表征。鉴于多源传感信息具有不同的物理属性与统计分布，该架构将原始状态向量解耦为四个专用子空间：空间、运动、障碍和噪声。每个子空间由专门针对其独特特性定制的专用编码器处理，以确保领域特定特征在互不干扰的情况下被提炼。该设计的主要目标是将这些异构子特征综合为一种状态表征。通过利用多头自注意力机制，该层主动建模不同信息流之间的跨模态相关性。这一过程产生统一且具备上下文感知的状态表征，为后续阶段实现鲁棒且自适应的决策提供高保真基础。

具身决策层：基于具身感知层处理后的信息，具身决策层采用多专家协同决策机制（详见第V-B节），作为水下应用的核心策略。它通过生成一种混合动作，将感知输入转化为可执行动作，该动作

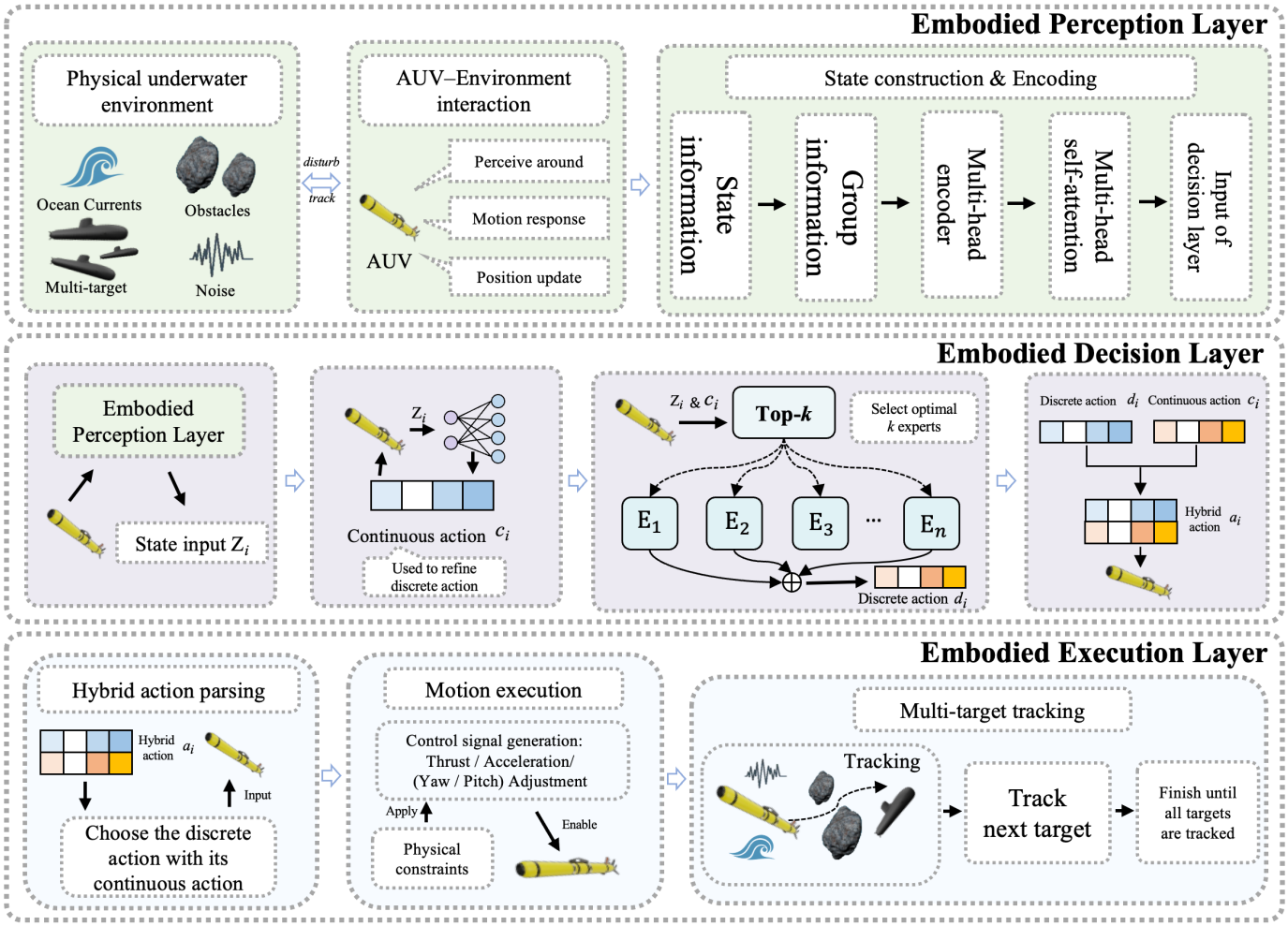


图1：所提出的分层 EI 架构

融合了离散动作与连续动作。借助采用 Top-k 专家选择策略的多专家协同决策机制，该层确保了计算效率、策略适应性和较高的决策准确率，使自主水下航行器能够在具有高维特征的水下环境中执行复杂任务。

具身执行层：执行层充当物理接口，将高层战略意图与自主水下航行器的水下动力学特性连接起来。它将混合动作 a_t 解析为专为该航行器推进系统设计的离散-连续控制指令。通过将定性机动（例如，加速动作）与定量控制量幅值（例如，加速度大小）分离，该层确保将策略输出精确映射为物理驱动。

V. 基于双头编码器-注意力的多专家协同决策算法

基于分层 EI 架构，本文提出了一种新颖的深度强化学习方法：面向具有高维特征的水下环境中自主水下航行器精确多目标跟踪的**双头编码器-注意力多专家协同决策算法**（DHEA-MECD）。受 Transformer 架构启发，所提出的方法将多头编码器、多头自注意力模块与多专家协同决策机制相结合。该设计利用多头编码器与自注意力机制，促进对自主水下航行器感知状态空间中的异构环境特征进行深度融合。为将这些表征转化为精确动作，本文引入 Top-k 专家选择策略，根据实时运动阶段动态选择最相关的专家。该选择策略实现了面向运动阶段的协同决策，在具有高维特征的水下环境中显著提升了跟踪精度和计算效率。

A. 双头编码器-注意力驱动的信息提取

准确的环境感知是智能决策的基石，使自主水下航行器能够在具有高维特征的水下环境中执行精确跟踪。传统方法通常将原始状态向量直接输入神经网络，但这往往无法捕捉异构语义特征之间的内在关联。为在此类场景中实现稳健决策，

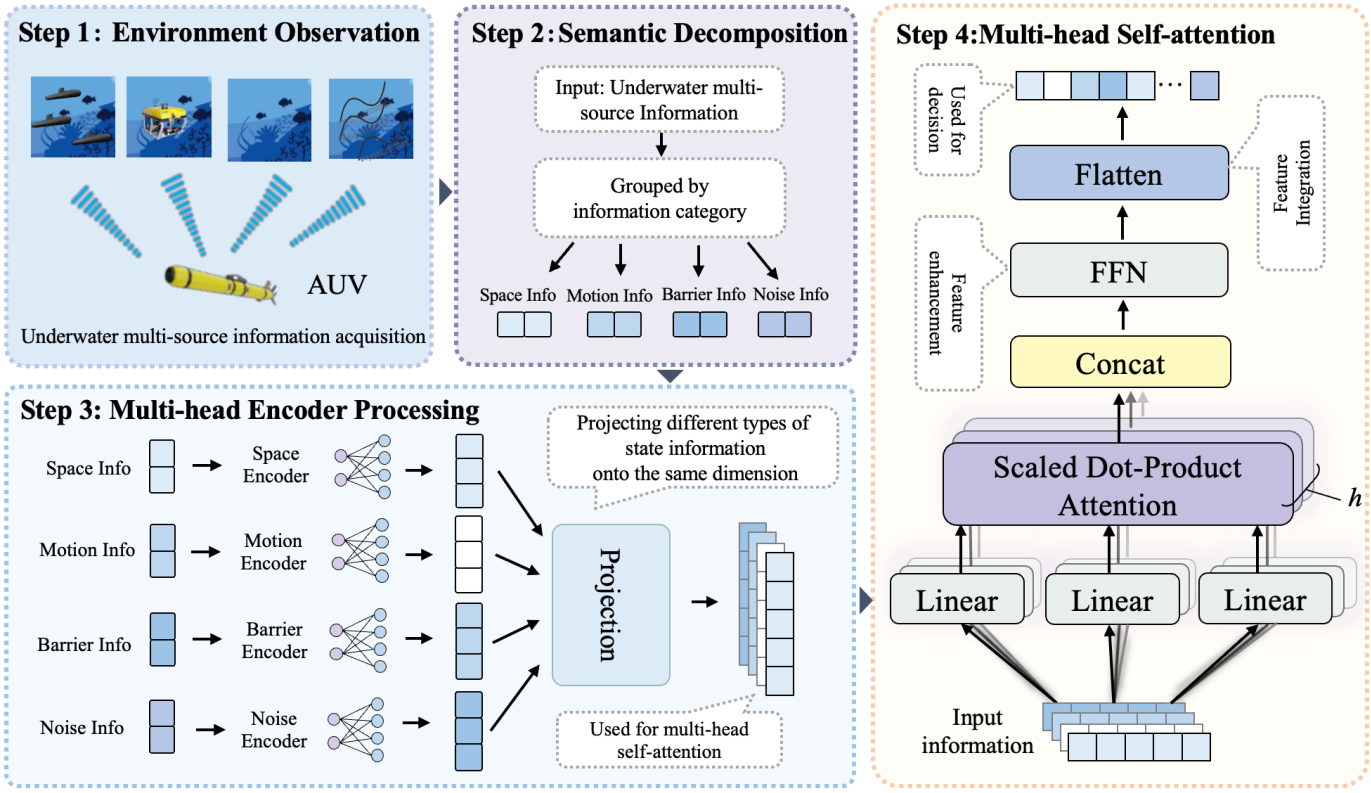


图 2：DHEA-MECD 中用于高维特征的拟议信息提取框架

我们提出了一种结构化状态表示，并设计了一种多头编码器架构，用于从感知观测中提取有意义的潜在特征。如图所示2，集成处理流程由以下四个连续步骤组成：

步骤 1 - 环境观测：在每个决策步骤中，自主水下航行器获取一个高维观测向量 s，其中包含来自水下环境的多源信息。该向量包括空间几何信息、运动动力学、约束条件，以及由感知不确定性和外部扰动引起的随机扰动。

步骤 2 - 语义分解：
为确保数值稳定性，观测向量 s 的每个分量首先被归一化到范围 $[0,1]$。状态向量被分解为四个语义独立的子向量：$s=[s_{\mathrm{sp}};\ s_{\mathrm{mot}};\ s_{\mathrm{bar}};\ s_{\mathrm{noise}}]^{\top}$。首先，$s_{\mathrm{sp}}$ 表示空间信息，包括自主水下航行器的位置 $p_a(t)$、所有目标的位置 $j\in\mathcal{J}$（记为 $p_j(t)$）、欧几里得距离$d_{a,j}(t)$以及捕获指示变量 $I(d_{a,j}(t)<r_{\mathrm{target}})$。其次，$s_{\mathrm{mot}}$ 描述运动状态，包括平移速度 $v_a(t)$、偏航-俯仰角的正弦编码（$\psi_a(t)$ $\theta_a(t)$）以及归一化时间索引 $t_{\mathrm{ratio}}=t/T_{\max}$。第三，$s_{\mathrm{bar}}$ 表示障碍信息，其封装了阻碍自主水下航行器跟踪的外部约束。该分量包括局部洋流速度障碍 $v_c(p_a(t)\ t)$ 以及一组 N 个周围障碍物，对于 $i=1,\ldots\ldots,\ N$记为$p_{o,i}(t)$。最后，s_{noise}是噪声信息，用于刻画影响感知与驱动的随机扰动。与水下噪声模型见第III-B节，它由噪声状态向量$n(t)=[n_{\mathrm{veh}}(t)\ n_{\mathrm{bio}}(t)\ n_{\mathrm{geo}}(t)\ n_{\mathrm{turb}}(t)]^{\top}$表示，该向量将多种噪声机制聚合为结构化随机输入。

为处理这些子向量中异质的物理属性，将它们输入四个专用编码器。该模块化设计确保通过有针对性的提取来提炼领域特定特征，然后将其整合到全局策略表示中。

步骤 3 - 多头编码器处理：为进一步处理信息，我们设计了一个由四个专用子编码器组成的多头编码器。每个模块执行特定的特征提取，并生成对齐的潜在表示。

空间编码器采用带有 GELU [31]激活函数和层归一化的前馈网络：

$$\begin{cases}\widetilde{h}_s^{(l)}=\mathrm{GELU}(W_s^{(l)}h_s^{(l-1)}+b_s^{(l)});\\ h_s^{(l)}=\mathrm{LayerNorm}(h_s^{(l-1)}+\widetilde{h}_s^{(l)}),\end{cases}\tag{12}$$

其中，$h_s^{(l)}\in\mathbb{R}^d$ 表示第l层的空间表示，$W_s^{(l)}\in\mathbb{R}^{d\times d}$ 和 $b_s^{(l)}\in\mathbb{R}^d$ 为可学习参数，$\widetilde{h}_s^{(l)}$为激活后的中间输出。请注意，层归一化会对每个样本的特征维度进行归一化，并稳定残差传播。

运动编码器用于提取运动信息。在我们的方法中，运动特征通过

Swish 激活的多层感知机，后接组归一化 [32]:

$$\begin{cases}\widetilde{h}_m^{(l)} = \mathrm{Swish}(W_m^{(l)} h_m^{(l-1)} + b_m^{(l)}); \\ h_m^{(l)} = \mathrm{GroupNorm}(\widetilde{h}_m^{(l)}),\end{cases} \tag{13}$$

其中， $h_m^{(l)} \in \mathbb{R}^m$ 是第 l 层的运动特征向量， $W_m^{(l)}$和 $b_m^{(l)}$ 是可训练参数，组归一化在预定义的通道组上对特征进行归一化，使表示对批量大小不敏感，并适用于幅值变化的运动信号。

基于 ReLU 的残差多层感知机构建了**屏障编码器**:

$$\begin{cases}\widetilde{h}_b^{(l)} = \mathrm{ReLU}\left(W_b^{(l)} h_b^{(l-1)} + b_b^{(l)}\right); \\ h_b^{(l)} = \mathrm{LayerNorm}\left(h_b^{(l-1)} + \widetilde{h}_b^{(l)}\right),\end{cases} \tag{14}$$

其中， $h_b^{(l)} \in \mathbb{R}^d$ 表示屏障特征， $W_b^{(l)}$ 和 $b_b^{(l)}$ 是相应的可训练参数， $\widetilde{h}_b^{(l)}$ 是激活的中间输出。需要明确的是，残差路径确保了通常具有高方差的约束相关特征能够保持稳定的梯度流。

噪声编码器 专门通过使用 ReLU 激活的多层感知机，并在其后接入批归一化，从噪声中提取有意义的信息:

$$\begin{cases}\widetilde{h}_n^{(l)} = \mathrm{ReLU}\left(W_n^{(l)} h_n^{(l-1)} + b_n^{(l)}\right); \\ h_n^{(l)} = \mathrm{BatchNorm}\left(\widetilde{h}_n^{(l)}\right),\end{cases} \tag{15}$$

其中，向量 $h_n^{(l)} \in \mathbb{R}^d$ 表示噪声信息， $W_n^{(l)}$和 $b_n^{(l)}$是可学习参数，批归一化利用批次统计量对激活值进行归一化，从而有效抑制噪声测量值中的随机波动。

在每个编码器的最后一层之后，模态特定表示 $h_i \in \mathbb{R}^d (i \in \{s, d, e, n\})$ 被映射到共享潜在空间 t_i:

$$t_i = W_i h_i + b_i, \qquad t_i \in \mathbb{R}^n, \tag{16}$$

其中， $W_i \in \mathbb{R}^{n \times d}$ 和 $b_i \in \mathbb{R}^n$ 是每种模态的投影参数。四个投影嵌入构成一个特征序列

$$T = [t_s; t_d; t_e; t_n] \in \mathbb{R}^{4 \times n}, \tag{17}$$

随后，该序列被输入多头自注意力模块。

步骤 4 - 多头自注意力： 为显式建模四个特定向量之间的依赖关系，我们采用多头自注意力模块。对于每个注意力头，查询Q、键 K 和值 V 矩阵的计算方式如下:

$$Q = TW_Q; \quad K = TW_K; \quad V = TW_V, \tag{18}$$

其中， T 是输入信息矩阵， W_Q， W_K， W_V 分别是用于查询、键和值的可学习投影矩阵。因此，逐头注意力输出可表示为:

$$A_h = \mathrm{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_h}}\right) V_h, \tag{19}$$

其中， A_h 表示第 h个注意力头的注意力输出， Q_h， K_h， V_h 分别表示各注意力头对应的查询、键和值矩阵， d_h 表示每个注意力头的维度。将所有注意力头的输出拼接后进行线性融合:

$$M = \mathrm{Concat}(A_1, \ldots, A_6) W_O, \tag{20}$$

其中， A_1，……， A_6 是全部六个注意力头的输出，Concat(·) 表示沿特征维度进行拼接， W_O 是输出投影矩阵。

随后，采用前馈增强块进一步细化融合后的信息表示 M:

$$\begin{cases}z_1 = \mathrm{GELU}\left(W_1^f M + b_1^f\right); \\ z_2 = \mathrm{Dropout}\Big(\mathrm{GELU}\left(W_2^f z_1 + b_2^f\right)\Big); \\ Z = \mathrm{LayerNorm}\left(W_3^f z_2 + b_3^f\right) + M,\end{cases} \tag{21}$$

其中 W_1^f、 W_2^f、 W_3^f和 b_1^f、 b_2^f、 b_3^f是前馈块的可学习参数。变量 z_1 和 z_2 表示中间特征变换。GELU(·) 是高斯误差线性单元激活函数，Dropout(·)用于正则化，而LayerNorm(·)表示层归一化。引入与输入 M 的残差连接，以稳定优化并促进梯度传播。

随后，将增强后的信息序列 Z 展平，并通过一系列全连接层投影为统一的策略嵌入:

$$\begin{cases}h_1 = \mathrm{LayerNorm}(\mathrm{GELU}(W_1^p z_f + b_1^p)); \\ h_2 = \mathrm{LayerNorm}\Big(\mathrm{Dropout}(\mathrm{GELU}(W_2^p h_1 + b_2^p))\Big),\end{cases} \tag{22}$$

其中， z_f表示 Z的展平形式， W_1^p、 W_2^p和b_1^p 、 b_2^p 是投影层的参数。最终，策略嵌入表示为

$$z = W_3^p h_2 + b_3^p, \tag{23}$$

其中，向量 z 编码了高维特征之间的跨结构依赖关系，并为决策层提供稳定且连贯的环境表示。

总体而言，所提出的多头编码器和多头自注意力模块以层次化方式整合信息交互与全局结构依赖关系，为策略学习生成紧凑且稳定的表示。

B. 多专家协同决策机制

基于所提出的信息提取框架提取的多模态环境表示，本节介绍所提出的多专家协同决策机制。该机制采用一种运动阶段感知的 Top-k策略，以实现在复杂水下环境中的自适应跟踪。通过根据 k 实时运动阶段选择性地调用最相关的专家，该机制在确保专业化决策的同时，最大限度地降低每步计算负担。图3 展示了整体算法，随后将详细介绍其核心步骤。

步骤 1 - 信息获取： 在进行决策之前，原始多源观测 s通过第 V-A 节详述的信息提取框架进行处理。

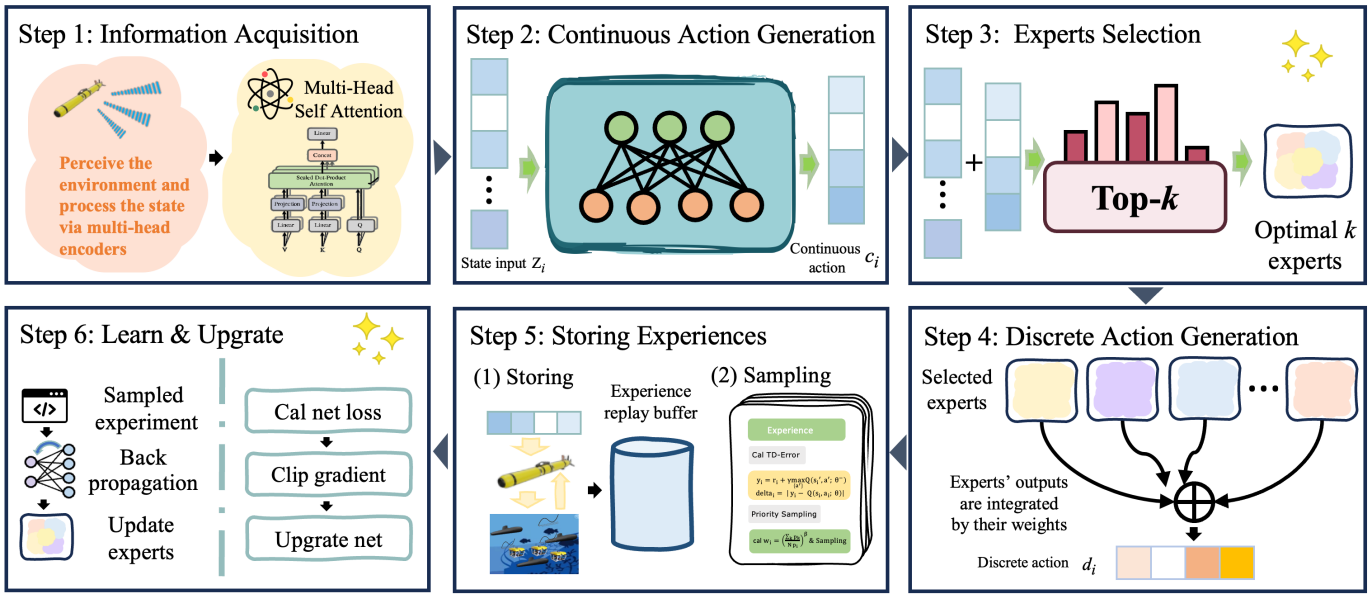


图 3：DHEA-MECD 中提出的多专家协同决策机制

由此得到紧凑的潜在状态向量 z_t，该向量封装了融合的时空依赖关系，并作为后续混合动作生成的基础输入。

步骤 2 - 连续动作生成：为实现超越离散动作的精细控制，我们设计了一个将状态 z_t 映射为连续动作参数的网络，从而构成混合动作空间。从结构上看，该网络采用门控循环单元来捕捉时间动态，随后使用多层感知机进行非线性特征变换。

首先，门控循环单元更新隐藏状态 h_t 以建模历史上下文：

$$h_t = \mathrm{GRU}(z_t, h_{t-1}), \tag{24}$$

其中，h_t 作为紧凑的时间表示。随后，h_t 经由一个两层 ReLU 激活的多层感知机处理，并通过双曲正切输出层投影到动作空间：

$$\begin{cases} f_1 = \mathrm{ReLU}(W_1^c h_t + b_1^c); \\ f_2 = \mathrm{ReLU}(W_2^c f_1 + b_2^c); \\ c_t = \tanh(W_o^c f_2 + b_o^c), \end{cases} \tag{25}$$

其中，$W_{(\cdot)}^c$ 和 $b_{(\cdot)}^c$ 分别表示各层的可学习权重和偏置。tanh 激活函数将输出限制在 $(-1,\ 1)$ 内，从而确保控制有界。

由此得到的连续动作向量 $c_t \in \mathbb{R}^4$ 编码了归一化加速度 $\dot{v}_t$、偏航调整 $\delta_{\psi,\ t}$、俯仰调整 $\delta_{\theta,\ t}$ 以及制动强度 b_t，从而有效地细化自主水下航行器的运动执行。

步骤 3 - 专家选择：为动态识别与当前运动阶段最相关的专家，我们采用 Top-k 专家选择策略。对于每个专家 i，引入一个可学习的键向量 k_i，用于编码其功能专长。将状态 z_t 投影到专家空间后，计算匹配分数 g_i，并通过 Top-k 选择策略来确定选定专家子集 Ω_t：

$$\begin{cases} g_i = z_t^\top W_g k_i + b_i; \\ \Omega_t = \mathrm{Topk}(\{g_1, \ldots, g_n\},\ k). \end{cases} \tag{26}$$

随后，对 Ω_t 应用稀疏 Softmax，以获得归一化专家权重：

$$\alpha_i(z_t) = \frac{\exp(g_i)}{\sum_{j \in \Omega_t} \exp(g_j)}, \quad i \in \Omega_t. \tag{27}$$

这种稀疏选择在确保计算效率的同时，将模型容量集中于最相关的专家。

在结构上，每个专家 i由一个 GRU-MLP 主干网络（参数独立）和一个运动阶段感知模块组成。专家输出 $E_i(z_t)$ 定义为：

$$E_i(z_t) = \mathrm{MLP}_i(\mathrm{GRU}_i(z_t)) + \underbrace{\sum_{j=1}^{d} w_{ij} f_{ij}(z_t^{(j)})}_{P_i(z_t)} + b_i^E, \tag{28}$$

其中，运动阶段感知模块 $P_i(z_t)$ 将加权变换 f_{ij} 应用于特定状态分量，通过有选择地强调对指定运动阶段至关重要的特征，增强专家级专门化。

步骤 4 - 离散化动作生成：基于步骤 2 中生成的连续动作 c_t，此步骤用于确定当前时间步 t的最优离散动作 d_t。由于c_t 已经细化了动作粒度，决策策略会将每个候选离散动作 $d \in \mathcal{D}$ 与当前状态 z_t 及连续动作 c_t配对后进行评估。

在给定选定专家子集 Ω_t 的条件下，特定配对 $(d,\ c_t)$ 的全局混合动作价值通过归一化专家权重聚合专门化评估得出：

$$Q_{\text{final}}(z_t, d, c_t; \omega) = \sum_{i \in \Omega_t} \alpha_i(z_t) \cdot Q_i(z_t, d, c_t; \omega_i), \tag{29}$$

其中，i 表示专家索引，ω 表示选定专家子集。最终，智能体会在当前专家共识下选择能够产生最高期望回报的离散动作：

$$d_t = \arg\max_{d\in\mathcal{D}} Q_{\text{final}}(z_t, d, c_t; \omega). \tag{30}$$

步骤5 - 经验存储：将转移封装为元组 $e_t = (z_t,\ d_t,\ c_t,\ r_t,\ z_{t+1},\ done_t)$，并存储在优先经验回放（PER）[33] 缓冲区 $\mathcal{B}$中。为提高样本效率，在采样过程中，根据时序差分（TD）误差更新每个转移的优先级 j： δ_j

$$\delta_j = \left|y_j - Q_{\text{final}}(z_j, d_j, c_j; \omega)\right|, \tag{31}$$

其中，j 表示小批量中采样转移的索引，y_j 是TD目标，c_j 是执行的连续动作。该策略会使采样过程偏向学习潜力更高的转移。

步骤6 - 学习与更新：训练目标是联合优化专家网络参数 ω 和连续动作网络参数 θ。从 $\mathcal{B}$ 中使用重要性采样权重w_j采样大小为B_{batch} 的小批量。

专家网络更新：多专家网络的主要目标是最小化贝尔曼预测误差。具体而言，转移 j 的TD目标 y_j 通过利用目标连续策略对离散动作进行最大化来构造：

$$y_j = r_j + \gamma(1 - done_j)\max_{d'\in\mathcal{D}} Q_{\text{final}}\left(z'_j, d', c^{\mu}(z'_j; \theta^{-}); \omega^{-}\right), \tag{32}$$

其中，z'_j 表示样本 j的下一状态，θ^{-}、ω^{-}表示目标网络参数。专家参数 ω通过最小化批次上的加权均方误差（MSE）进行更新：

$$\mathcal{L}_Q(\omega) = \frac{1}{B_{batch}}\sum_{j=1}^{B_{batch}} w_j\left(y_j - Q_{\text{final}}(z_j, d_j, c_j; \omega)\right)^2. \tag{33}$$

关键是，为实现功能专长，梯度仅反向传播至选定专家。对于特定专家 i 和样本 j，梯度贡献通过掩码处理如下：

$$\nabla_{\omega_i}\mathcal{L}_Q^{(j)} = \mathbb{I}(i \in \Omega_j)\cdot\alpha_i(z_j)\cdot\frac{\partial\mathcal{L}_Q^{(j)}}{\partial Q_{\text{final}}}\cdot\nabla_{\omega_i}Q_i(z_j, d_j, c_j; \omega_i), \tag{34}$$

其中，$\mathbb{I}(\cdot)$表示指示函数。该掩码机制可防止休眠专家（$i \notin \Omega j$）发生参数退化，确保它们保留针对特定运动流形的专长。

连续动作网络更新：连续动作网络 θ 通过确定性策略梯度（DPG）方法进行更新。损失函数对批次中的Q值取平均：

$$\mathcal{L}_\mu(\theta) = -\frac{1}{B_{batch}}\sum_{j=1}^{B_{batch}} Q_{\text{final}}(z_j, d_j, c^{\mu}(z_j; \theta); \omega). \tag{35}$$

在这里，固定的评论家网络充当可微代理，

算法1 DHEA-MECD的学习过程

输入：优先经验回放缓冲区 $\mathcal{B}$，批量大小 B_{batch}
初始化：专家参数 ω，连续策略 θ，目标网络 $\omega^{-} \leftarrow \omega, \theta^{-} \leftarrow \theta$
当训练**时**
  根据 $\mathcal{B}$ 中的 B_{batch} 转移采样一个小批量
  优先级权重 w_j
  通过信息提取框架提取潜在状态 z_j 和下一状态 z'_j（第V-A节所述）
  对于小批量**中的**每个转移 j **执行**
    根据公式（25）计算目标连续动作 c'_j
    根据公式（26）计算 z'_j 对应的选定专家Ω'_j
    使用公式（32）计算TD目标 y_j
  计算专家网络损失 $\mathcal{L}_Q(\omega)$根据公式（33）
  通过梯度下降更新 ω
  计算连续网络损失 $\mathcal{L}_\mu(\theta)$ 根据公式（35）
  通过梯度下降更新 θ
  软更新目标网络：$\omega^{-} \leftarrow \tau\omega + (1-\tau)\omega^{-}$，$\theta^{-} \leftarrow \tau\theta + (1-\tau)\theta^{-}$
  更新信息提取框架参数
  基于时序差分误差 δ_j更新 $\mathcal{B}$ 中的转移优先级根据公式（31）

引导 θ 输出最优调制参数。

最后，通过软替换更新目标网络：$\phi^{-} \leftarrow \tau\phi + (1-\tau)\phi^{-}$，其中 $\phi \in \{\omega, \theta\}$。

VI. 基于DHEA-MECD算法的自主水下航行器跟踪算法

基于所提出的DHEA-MECD算法，本节介绍一种面向具有高维特征的水下环境而设计的基于自主水下航行器的多目标跟踪算法。如前文所述，我们的自主水下航行器多目标跟踪框架采用EI架构，通过分层感知—决策—执行过程，实现复杂海洋环境中的高效目标获取。在这一流程中，采用第五节所述的信息提取框架，从高维环境状态中提炼紧凑的潜在表示；随后，所提出的多专家协同决策机制生成稳健自主水下航行器跟踪所需的混合（离散—连续）动作。

A. 奖励设计

为了实现自主水下航行器在具有高维特征的水下环境中稳定、高效且安全的多目标跟踪，我们设计了结构化奖励函数 r 如式（36）所示：

$$r(t) = r_d(t) + r_g(t) + r_c(t) + r_b(t), \tag{36}$$

其中，$r_d(t)$ 表示距离奖励，$r_g(t)$ 表示目标奖励，$r_c(t)$ 表示避碰奖励，$r_b(t)$ 表示工作空间遵循奖励。

距离奖励 r_d 鼓励自主水下航行器在每个时间步持续接近目标。奖励定义如式（37）：

$$r_d(t) = \alpha_d\big(d_{a,j}(t-1) - d_{a,j}(t)\big), \tag{37}$$

其中，$d_{a,j}(t) = \|p_a(t) - p_j(t)\|_2$ 表示在时刻 t自主水下航行器与目标之间的欧几里得距离，$d_{a,j}(t-1)$表示 $t-1$ 时刻的距离，而 $\alpha_d > 0$ 是一个缩放系数，用于调节距离奖励的幅度。

目标奖励 r_g 在自主水下航行器成功到达目标时提供稀疏但显著的终止信号。该奖励项旨在强化任务的成功完成，并引导策略实现全局收敛。奖励定义如式（38）：

$$r_g(t) = \alpha_g \cdot \mathbb{I}_{goal}(t), \tag{38}$$

其中，$\mathbb{I}_{goal}(t) \in \{0,1\}$ 是一个二值指示函数：若目标在时刻 t 被成功跟踪，则取值为1，否则为0；$\alpha_g > 0$ 表示缩放系数，用于调节终端任务达成激励的贡献。

避碰奖励 r_c 用于提供负向激励，以减少自主水下航行器与环境障碍物发生碰撞的情况。奖励定义如式（39）：

$$r_c(t) = -\alpha_c \cdot \mathbb{I}_{\text{collision}}(t), \tag{39}$$

其中，$\mathbb{I}_{\text{collision}}(t) \in \{0,1\}$ 是一个二值指示函数：若在时间 t 检测到碰撞，则取值为 1，否则取值为 0。项 $\alpha_c > 0$ 表示用于调节碰撞相关成本影响的缩放系数。

工作空间遵循奖励 r_b 用于对自主水下航行器的运动施加空间约束。当自主水下航行器超出边界时，该项会施加边界成本，从而确保智能体始终处于有效工作空间内。奖励定义如式（40）：

$$r_b(t) = -\alpha_b \cdot \mathbb{I}_{\text{boundary}}(t), \tag{40}$$

其中，$\mathbb{I}_{\text{boundary}}(t) \in \{0,1\}$ 是一个二值指示函数：若自主水下航行器超出工作空间边界，则取值为 1，否则取值为 0。系数 α_b 用于缩放这一负向激励，从而有效地将自主水下航行器限制在允许区域内。

B. 基于 *DHEA-MECD* 的多目标跟踪算法

基于EI驱动的DHEA-MECD的所提出多目标跟踪算法的完整执行过程总结于算法2中。具体而言，第1–6行初始化网络，并通过语义分解、多头编码器和基于多头自注意力的融合，将原始状态 s_t 转换为潜在表示 z_t 。随后，第7–14行通过动态选择Top-k 专家子集Ω_t 并融合其输出，生成混合动作，以确定连续动作 c_t 和离散动作 d_t。最后，第15–19行执行组合后的动作a_t，计算多目标跟踪奖励 r_t，并通过算法1 对感知与决策模块进行联合优化，从而确保稳健的跟踪性能。

算法2 基于 DHEA-MECD 的跟踪算法

输入： 环境状态 s_t，目标专家数量 k，缩放系数 α
初始化： 经验回放缓冲区 $\mathcal{B}$，分层编码器参数-、自注意力权重、连续网络和专家网-络
当任务处于活动状态**执行**
  观测原始观测 s_t 并归一化为 $[0,1]$
  分解 s_t into four 语义 子空间：$\{s_{\text{sp}}, s_{\text{mot}}, s_{\text{bar}}, s_{\text{noise}}\}$;
  提炼 使用专门的模态特定特征多头编码器
  通过多头自注意力建模跨模态依赖关系以生成潜在状态 z_t
  生成连续动作 c_t ，根据公式（25）
  计算匹配分数并选择排名前-k 位的专家子集 Ω_t 公式（26）
  对于 每个选定专家 $i \in \Omega_t$ **执行**
    根据公式计算专家权重 α_i （27）
    计算专家动作价值 $Q_i(z_t,\ d,\ c_t)$
  根据公式合成全局动作价值 Q_{final} （29）
  选择离散动作 d_t ，根据公式（30）
  组合混合动作 $a_t = (d_t,\ c_t)$
  通过 AUV推进系统执行 a_t
  获取奖励 r_t ，根据公式（36）并获取下一状态 s_{t+1}
  追加转移 $e_t = (z_t, d_t, c_t, r_t, z_{t+1}, done_t)$ 至优先级缓冲区 $\mathcal{B}$
  对编码器和策略进行联合优化通过算法 1

VII. 评估

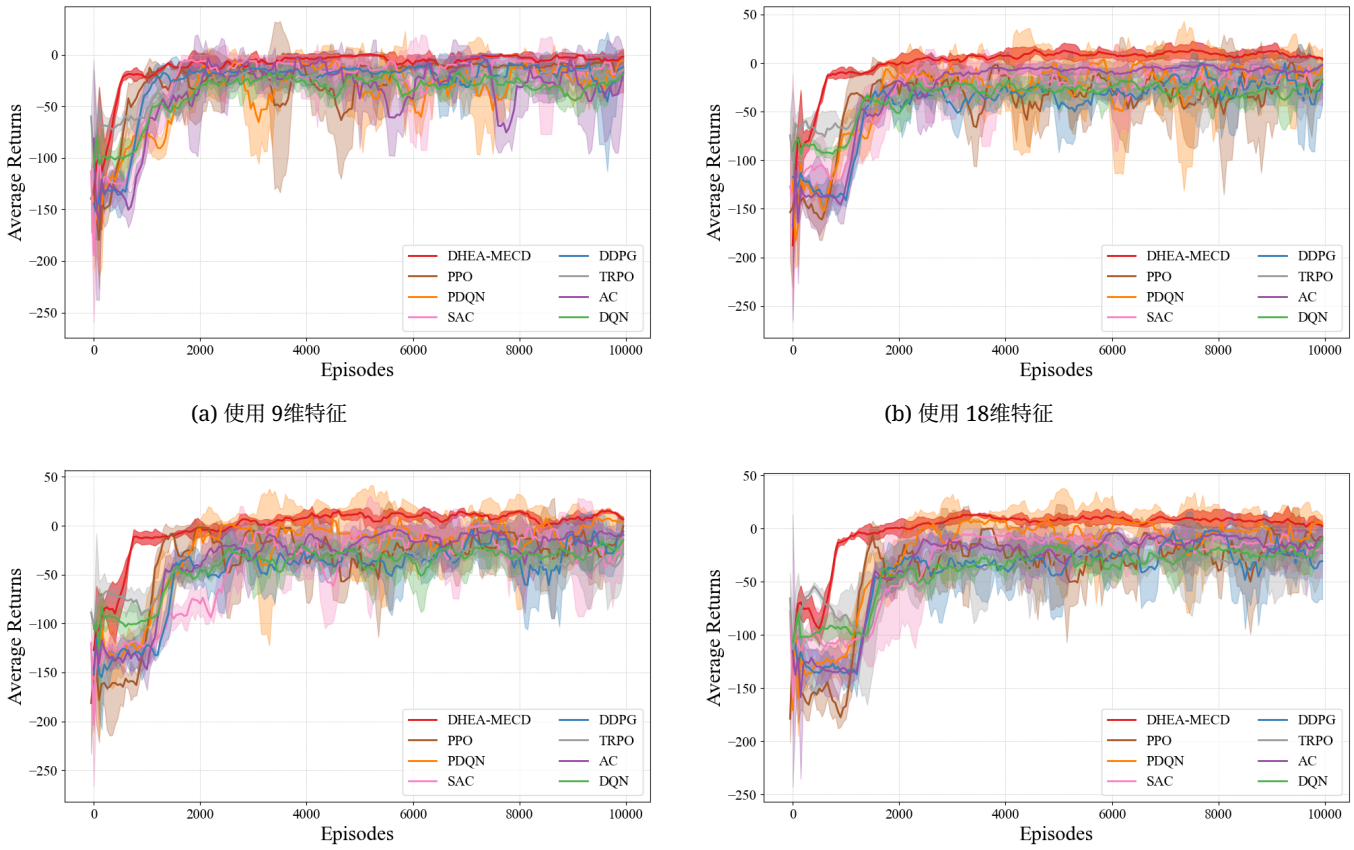
本节对所提出的方法与若干主流DRL驱动方法进行了全面的比较评估。

A. 仿真设置

所有仿真均在配备 Apple M3 处理器和 16 GB 内存的 MacBook Air（2024）上进行。在仿真中，自主水下航行器和目标实体均被抽象为在三维连续空间中运动的动态粒子。仿真中的所有参数详见表格。第四节。

表 I：仿真参数

参数	描述	值
$T_{\max}$	每回合的最大步数	800
N_{episodes}	训练回合总数	10,000
N_{expert}	专家数量	5
k	每步选定的专家数量	2
α	专家网络学习率	0.001
β	连续网络学习率	0.001
γ	折扣因子	0.95
T_{update}	目标网络更新频率	200
B_{batch}	批量大小	64
M	经验回放缓冲区容量	200,000
ϵ_0	初始探索率	1.0
λ_ϵ	探索衰减率	0.0002
$\|\nabla\|_{\max}$	梯度裁剪范数	1.0



(a) 使用 9维特征

(b) 使用 18维特征

(c) 使用 27维特征

(d) 使用 36维特征

图 4：收敛速度比较

B. 结果

为评估所提出的 DHEA-MECD 在多目标跟踪中的有效性，我们将其与一系列主流的深度强化学习算法驱动方法进行全面比较，包括近端策略优化 [34], PDQN [35],软演员-评论家 [36],深度确定性策略梯度 [37], 信任域策略优化 [38], 演员-评论家 [39],和深度Q网络 [40], , 比较指标分别为收敛速度、轨迹长度、跟踪精度和碰撞率。

1）收敛速度：为评估收敛速度，我们在不同状态特征维度的环境下开展了多次实验†，并使用带有96%置信区间的平均回报评估不同深度强化学习算法的收敛性。

如图4所示，在大多数场景中，所提出的 DHEA-MECD 始终比其他算法收敛更快。值得注意的是，随着状态特征维度的增加，DHEA-MECD 与对比算法之间的性能差距日益显著。具体而言，对比算法 PPO、SAC、DDPG、DQN 在高维空间中表现出显著的性能波动，这主要是由于它们在结构化特征提取方面的能力有限。因此，这些算法难以应对冗余和噪声表征，导致优化不稳定。

相比之下，DHEA-MECD在训练早期表现出显著更快的收敛速度。这种效率得益于多头编码器和注意力架构，它们将高维观测分解为连贯的潜在流形。通过提炼紧凑的嵌入 z_t，这种结构偏置能够有效过滤噪声与冗余，从而促进比对比方法更高效的探索。此外，DHEA-MECD中的 Top-k 专家选择策略通过将梯度传播限制在相关子网络中，加速了策略搜索。不同于对比深度强化学习算法的单体架构，我们提出的DHEA-MECD减少了复杂混合动作空间中的参数干扰。在训练后期，DHEA-MECD在优先经验回放机制的支持下保持平滑的轨迹，并将振荡降至最低。通过将梯度更新集中于高影响力的转移上，优先经验回放机制增强了价值估计稳定性。总体而言，结构化表征、稀疏协作与优先回放的协同作用确保了稳健收敛，使系统能够在动态、高维水下环境中实现可靠的跟踪性能。

2）轨迹长度：轨迹长度定义为自主水下航行器在跟踪目标过程中航行距离的累积值，并被用作评估

† 有关维度特征的详细规范见附录“附录”部分。

表 II：轨迹长度分布比较。

维度	采用9维特征				采用18维特征				采用27维特征				采用36维特征			
长度(m)	200-300	300-350	350-400	400-500	200-300	300-350	350-400	400-500	200-300	300-350	350-400	400-500	200-300	300-350	350-400	400-500
PPO	35.2%	24.2%	27.1%	13.5%	38.1%	24.2%	25.1%	12.6%	42.3%	26.9%	19.9%	10.9%	44.6%	26.7%	18.8%	9.9%
PDQN	44.4%	24.8%	19.8%	10.9%	46.4%	26.0%	17.7%	9.8%	49.7%	27.9%	14.9%	7.5%	53.8%	25.4%	14.2%	6.6%
SAC	11.7%	13.2%	9.2%	65.9%	12.9%	13.6%	9.0%	64.5%	13.2%	15.3%	8.7%	62.7%	15.7%	16.1%	8.4%	59.9%
DDPG	8.2%	10.9%	17.3%	63.6%	8.6%	11.9%	17.0%	62.5%	9.0%	12.3%	17.0%	61.7%	10.3%	13.5%	16.1%	60.1%
TRPO	4.9%	7.3%	17.1%	70.7%	5.3%	7.7%	16.9%	70.0%	5.7%	8.2%	16.4%	69.7%	6.3%	9.0%	15.9%	68.8%
AC	4.8%	13.0%	18.8%	63.3%	5.2%	13.5%	18.6%	62.7%	5.6%	14.1%	18.4%	61.9%	6.1%	14.6%	18.6%	60.7%
DQN	5.8%	7.2%	14.5%	72.5%	6.6%	8.1%	14.2%	71.1%	7.5%	9.0%	13.9%	69.7%	8.5%	9.7%	13.6%	68.2%
DHEA-MECD	**56.5%**	**21.7%**	**16.3%**	**5.4%**	**59.1%**	**22.7%**	**13.6%**	**4.5%**	**63.1%**	**22.5%**	**10.8%**	**3.6%**	**66.7%**	**24.1%**	**6.5%**	**2.8%**

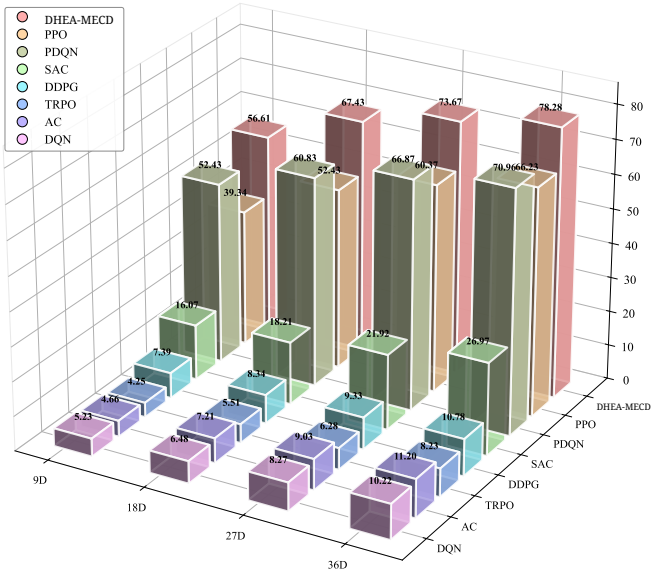


图 5：准确率对比

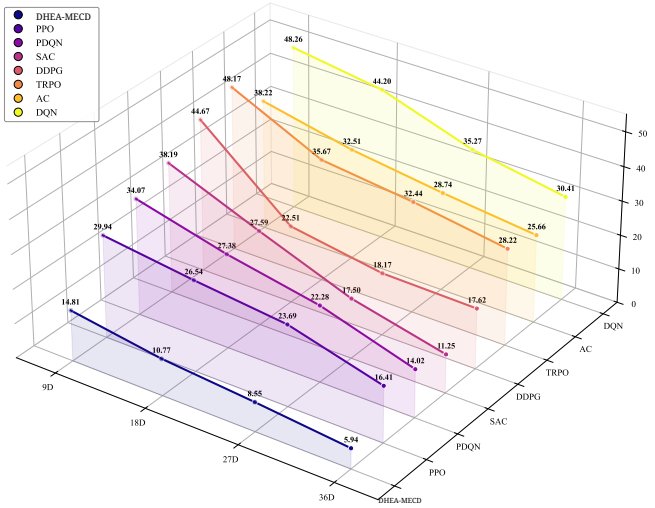
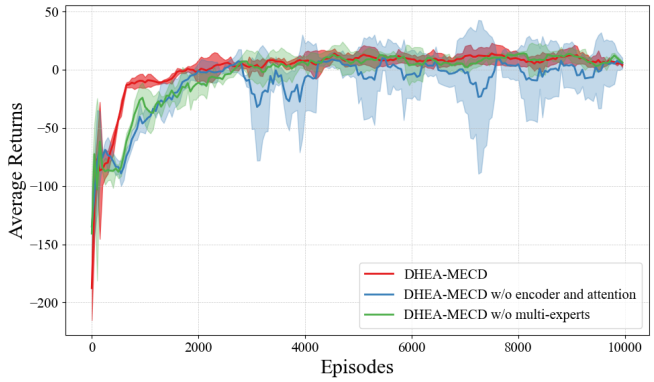
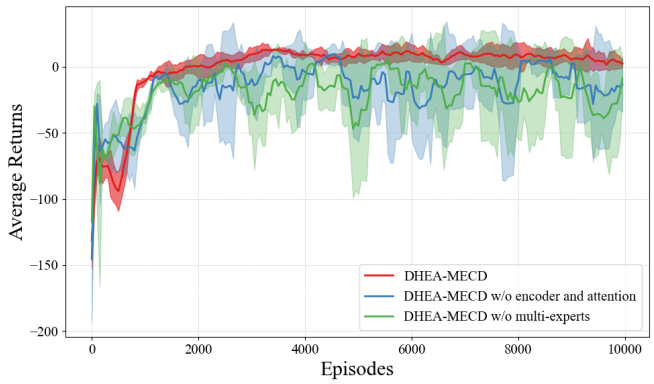


图 6：碰撞率对比

跟踪效率。表II 展示了不同算法在不同特征维度下成功跟踪目标所需轨迹长度的分布。在所有场景（特征维度各不相同）中，DHEA-MECD 在 200–300米区间内始终取得最高的成功回合集中度。尤为重要的是，随着维度从 9 增至 36，这一性能优势愈发显著。尽管其他对比算法



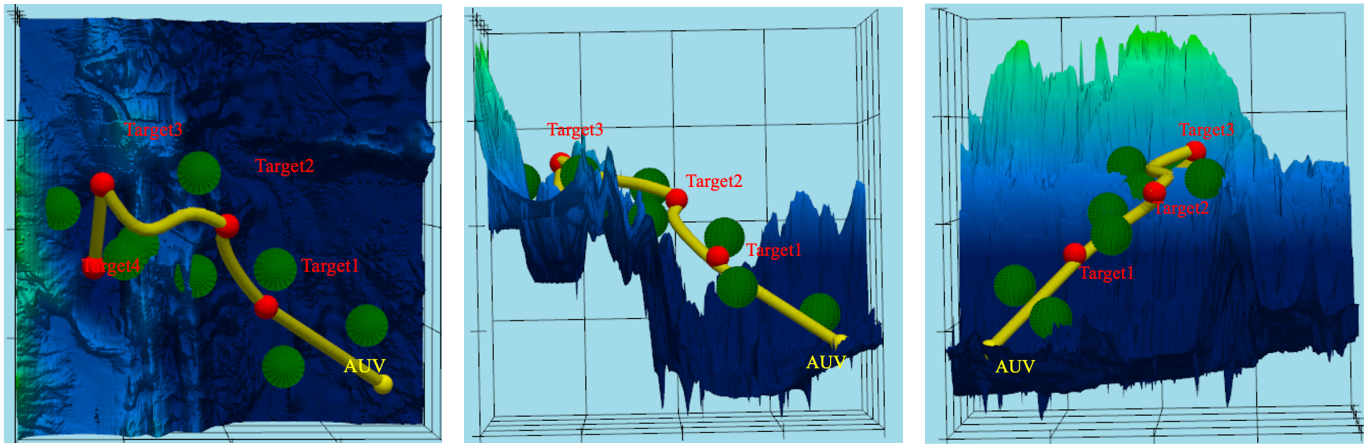
(a) 18维特征消融实验



(b) 36维特征消融实验

图7：消融研究

在高维环境中，其他方法的效率会显著下降，而 DHEA-MECD 仍能保持高质量表现。这得益于其结构化表示学习和多专家协同决策机制，能够有效应对复杂的维度问题并最大限度地减少冗余运动。DHEA-MECD 所实现的卓越轨迹，主要源于其层次化架构的结构优势。双头编码器—注意力模块将高维观测提炼为经过噪声抑制且语义对齐的潜在表征，从而能够更准确地推断时空依赖，并减少多余的纠正动作。同时，运动阶段感知的 Top-k 专家选择机制将决策集中于与当前阶段最相关的专家，而混合（离散—连续）动作



(a) 跟踪过程俯视图 (b) 跟踪过程侧视图 1 (c) 跟踪过程侧视图 2

图 8：可视化结果。

执行则产生更加平滑、方向一致性更强的轨迹。总体而言，这些机制使 DHEA-MECD 能够将更丰富的环境信息转化为果断且目标导向的行为，从而在所有特征维度的环境中生成更短、更直接的跟踪路径。

3) 跟踪精度：图5 展示了不同观测特征维度下各算法的成功率对比。随着特征维度从 9 增加到 36，所提出的 DHEA-MECD 始终取得所有对比方法中最高的成功率。值得注意的是，在高维环境中，其性能优势日益显著，体现出出色的可扩展性和鲁棒性。

相比之下，PPO、SAC 和 DDPG 等对比算法的成功率会随着观测维度增加而明显下降。这一性能下滑表明，它们处理复杂高维状态空间的能力有限，冗余特征和噪声观测会显著损害策略学习。尤其是在 36维环境中，对比方法的性能大幅下降，凸显了它们容易受到维数灾难影响。

不同于传统的单策略方法，DHEA-MECD 通过多专家协作机制，明确将结构化状态表示学习与混合动作决策分离。这一设计使其即使在具有挑战性的高维环境中，也能保持可靠的目标跟踪性能。

4）碰撞率：避碰是多目标跟踪任务中的一项关键要求，尤其是在具有高维特征的水下环境中。

图 6 展示了不同算法在不同观测特征维度下的碰撞率对比。在低维环境中，所有算法的碰撞率都相对较低。然而，随着特征维度的增加，各对比方法的碰撞率明显上升，表明在避碰方面面临的挑战日益增大在复杂的高维观测环境中。

相比之下，所提出的 DHEA-MECD 在所有不同特征维度的环境中，始终保持最低的碰撞率。值得注意的是，其碰撞率从 9 维到 36 维稳步下降，展现出强大的鲁棒性，即使在高维环境中也能实现可靠的避障性能。这凸显了 DHEA-MECD 在感知复杂度增加时学习稳定、安全导航行为的卓越能力。

特别是，PPO、SAC 和 DDPG 等对比算法表现不佳，原因在于它们在高维观测下协调运动规划与连续控制的能力有限。在相同的训练时长内，这些方法往往在学会有效的避障策略之前执行过多的转向或过于激进的运动指令，从而增加碰撞风险。相比之下，DHEA-MECD 明确将离散运动策略选择与连续控制参数优化解耦，使其在接近障碍物时能够更精确地调节速度、转向和制动行为。

5) 消融研究： 为评估所提出方法中各核心组件的独立贡献，通过将完整的 DHEA-MECD 与两个简化变体进行比较开展消融研究：(1) 移除多头编码器和多头自注意力模块的 DHEA-MECD；(2) 移除多专家协同决策机制的 DHEA-MECD。在不同观测维度下的学习曲线如图7所示。

如图所示 7a，在 18维观测 下，移除多头编码器和注意力模块会导致训练早期的收敛明显变慢，性能波动也更大。这表明，结构化表示学习能够提取紧凑且信息丰富的状态嵌入，在加速早期策略优化方面发挥着关键作用，从而促进更高效的探索。

相比之下，不具备多专家协同决策机制的变体在

早期学习行为方面表现相对相近，但在训练中后期振荡加剧，性能下降。这种退化表明，专家协作对于稳定决策和降低策略方差至关重要，尤其是在复杂的混合动作空间中。在更具挑战性的 36维环境 中也观察到类似趋势，如图 7b所示，缺少任一组件都会导致性能更加明显地下降。值得注意的是，完整的 DHEA-MECD 在所有训练阶段始终取得更高的平均回报，且收敛轨迹更加平滑。

最后，采用基于网格的方案构建水下环境，以模拟真实的运行条件。使用 `pyvista` 库对 DHEA-MECD 引导的多目标跟踪过程进行可视化。如图 8所示，红色球体表示目标，黄色球体表示自主水下航行器。实线曲线展示了不同跟踪场景下相应的目标和自主水下航行器轨迹，如图 8a 、图 8b 和图 8c 所示。结果表明，所提出的DHEA-MECD算法能够使自主水下航行器在复杂水下环境中生成平滑、稳定且自适应的跟踪轨迹，展现出其在动态条件下的有效性和鲁棒性。

VIII. 结论

本文提出了一种分层 EI 架构，以应对在具有高维特征的水下环境中基于自主水下航行器进行多目标跟踪所固有的复杂性。基于该 EI 架构，我们提出了DHEA-MECD算法，该算法利用多头编码器和多头自注意力模块，解决复杂海洋环境中的“感知碎片化”问题。通过采用具备运动阶段感知能力的 Top-k 专家选择策略的多专家协同决策机制，所提出的方法有效缓解了参数干扰，并确保较高的计算效率。大量实验评估表明，在不同特征维度下，DHEA-MECD的表现始终优于主流深度强化学习算法（包括近端策略优化、PDQN和深度确定性策略梯度）。具体而言，随着观测维度增加，我们的方法展现出更高的样本效率、渐近稳定性和跟踪精度，且性能优势显著扩大。消融研究进一步证实，结构化表示学习与协同决策对于实现稳健的水下多目标跟踪具有协同作用，二者缺一不可。

在未来的工作中，我们计划将该框架扩展至多自主水下航行器协同跟踪场景，并研究分布式专家协调策略，以应对大规模目标集群。此外，我们希望通过在真实水文条件下的实体自主水下航行器平台上验证所提出的算法，弥合“仿真到现实”的差距。

参考文献

[1] C. Frazão Santos, T. Agardy, D. Aheto, E. H. Allison, N. J. Bennett, J. L. Blythe, H. Calado, L. B. Crowder, J. C. Day, A. de Vos 等，“推进海洋可持续发展的跨学科知识”，*npj OceanSustainability*，第2卷，第1期，第18页，2023年。

[2] L. Sun、Y. Wang、X. Hui、X. Ma、X. Bai和M. Tan，“水下机器人及作业控制关键技术”，《仿生与仿生系统》，第5卷，第0089页，2024年。 [3] L. Christensen、J. de Gea Fernández、M. Hildebrandt、C. E. S. Koch和B. Wehbe，“水下机器人导航与控制中人工智能的最新进展”，《当代机器人学报告》，第3卷，第4期，第165–175页，2022年。 [4] J. Zhou、Y. Si和Y. Chen，“海底自主水下航行器技术综述”，《海洋科学与工程学报》，第11卷，1119，2023年第6期，第页。 [5] R. Zeng、C. Jiang、X. Wang和B. Li，“虚假数据注入攻击下无人机团队多任务学习的联合安全机制”，IEEE移动计算汇刊，2025年。 [6] S. Zhu 、 G. Han 和 C. Lin，“ 一种用于自主水下航行器集群网络以实现协同智能水下目标跟踪的基于软件定义多智能体强化学习的架构”，IEEE无线通信，第31卷，第6期，第56–62页，2024年。 [7] C. Ma、N. Wang、Q. A. Chen和C. Shen，“Slowtrack：利用对抗样本增加自动驾驶中基于摄像头的感知延迟”，载于AAAI人工智能会议论文集，第38卷，第5期，2024年，第4062–4070页。 [8] R. Zeng、C. Zeng、X. Wang、B. Li和X. Chu，“联邦学习中的激励机制及一种博弈论方法”，《IEEE网络》，第36卷，第6期，第229–235页，2022年。 [9] F. Sun、R. Chen、T. Ji、Y. Luo、H. Zhou和H. Liu，“具身智能综合综述：进展、挑战与未来展望”，《中国人工智能学会人工智能研究》，第3卷，第9150042期，第1页，2024年。 [10] W. Liang、R. Zhou、Y. Ma、B. Zhang、S. Li、Y. Liao和P. Kuang，“大模型赋能的具身人工智能：关于决策与具身学习的综述”，arXiv预印本arXiv:2508.10399，2025年。 [11] L. Lei、Y. Zhou和G. Yang，“基于多源信息融合的非规则海洋环境中水下航行器环境感知与动态模型”，《信息融合》，第94卷，第257–271页，2023年。 [12] D. Arce、J. Solano和C. Beltrán，“动态场景下室内自主导航的传统算法与基于深度强化学习的算法比较研究”，《传感器》，第23卷，第24期，第9672页，2023年。 [13] N. D. Merchant、R. L. Putland、M. André、E. Baudin、M. Felli、H. Slabbekoorn和R. Dekeling，“支持《欧洲海洋战略框架指令》的十年水下噪声研究”，《海洋与海岸带管理》，第228卷，第106299页，2022年。 [14] S. Zhu、G. Han、C. Lin和Q. Tao，“基于分层软件定义多自主水下航行器强化学习的水下目标跟踪：一种多自主水下航行器优势注意力行动者-评论家方法”，IEEE移动计算汇刊，第23卷，第12期，第13639–13653页，2024年。 [15] H. Ju、R. Juan、R. Gomez、K. Nakamura和G. Li，“面向机器人学将深度强化学习策略从仿真迁移到现实”，《自然·机器智能》，第4卷，第12期，第1077–1087页，2022年。 [16] D. Rawat、M. K. Gupta和A. Sharma，“机器人机械臂的智能控制：综合综述”，《空间信息研究》，第31卷，第3期，第345–357页，2023年。 [17] R. Lu、Z. Jiang、T. Yang、Y. Chen、D. Wang和X. Peng，“一种面向工业能源管理的新型基于混合动作的深度强化学习”，《IEEE工业信息学汇刊》，2024年。 [18] Y. Long、W. Wei、T. Huang、Y. Wang和Q. Dou，“面向手术机器人学习的人在回路具身智能与交互式仿真环境”，IEEE机器人与自动化快报，第8卷，第8期，第4441–4448页，2023年。 [19] G. Ferri、A. Munafo和K. D. LePage，“一种用于目标跟踪的自主水下航行器数据驱动控制策略”，《IEEE海洋工程学报》，第43卷，第2期，第323–343页，2018年。 [20] S. Zhu、G. Han、C. Lin和F. Zhang，“基于最小奖励参与嵌入式多智能体强化学习的水下多自主水下航行器协同目标跟踪”，IEEE移动计算汇刊，2024年。 [21] G. Chen、Y. Shen、N. Qu和B. He，“基于行为决策的自主水下航行器潜航过程路径规划”，《海洋工程》，第234卷，第109073页，2021年。 [22] D. Liang、J. Chu、Y. Cui、D. Liang和Y. Feng，“基于复杂环境仿真与ACL-SAC深度强化学习的自主水下航行器水下动态跟踪控制”，《IEEE智能车辆汇刊》，2024年。 [23] Y. Fang、Z. Huang、J. Pu和J. Zhang，“基于快速部署的深度强化学习方法的自主水下航行器位置跟踪与轨迹控制”，《海洋工程》，第245卷，第110452页，2022年。

[24] S. Zhu、G. Han、C. Lin和Y. Zhang，“基于中断式软件定义的多自主水下航行器强化学习的水下目标跟踪：一种多自主水下航行器省时多智能体强化学习方法”，*IEEE*移动计算汇刊，2024年。[25] Y. Gan、B. Zhang、J. Shao、Z. Han、A. Li和X. Dai，“具身智能：融合环境感知、自主规划与运动控制的仿生机器人控制器”，*IEEE*机器人与自动化快报，第9卷，第5期，第4559–4566页，2024年。[26] Y. Long、W. Wei、T. Huang、Y. Wang和Q. Dou，“面向手术机器人学习的人在回路具身智能与交互式仿真环境”，*IEEE*机器人与自动化快报，第8卷，第8期，第4441–4448页，2023年。[27] H. Song、L. Wang、X. Qiao、Y. Chen、D. Sun和Z. Sun，“面向机器人操作的具身智能：发展与挑战”，*Vicinagearth*，第2卷，第1期，第8页，2025年。[28] B. Mao、H. Yang、W. Li、X. Zhu和Y. Zheng，“自主水下航行器水下辐射噪声的多频降噪方法”，《海洋科学与工程学报》，第12卷，第5期，第705页，2024年。[29] K. P.aczek、D. Jelito、M. Pitera和A. Wyłoma´nska，“利用分位数条件方差比估计对称 α稳定分布的稳定性指数”，*TEST*，第33卷，第1期，第297–334页，2024年。[30] D. Hu、Y. Hu、S. Yi、X. Liang、Y. Li和X. Yang，“基于机器学习算法的矩形顶管隧道地表沉降预测”，《管道系统工程与实践学报》，第15卷，第1期，第04023061页，2024年。[31] D. Hendrycks，“高斯误差线性单元（GELU）”，*arXiv*预印本*arXiv:1606.08415*，2016年。[32] Y. Wu和K. He，“组归一化”，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第3–19页。[33] T. Schaul、J. Quan、I. Antonoglou和D. Silver，“优先经验回放”，*arXiv*预印本*arXiv:1511.05952*，2015年。[34] J. Schulman、F. Wolski、P. Dhariwal、A. Radford和O. Klimov，“近端策略优化算法”，*arXiv*预印本*arXiv:1707 06347*，2017年。。[35] M. Hausknecht和P. Stone，“参数化动作空间中的深度强化学习”，*arXiv*预印本*arXiv:1511.04143*，2015年。[36] T. Haarnoja、A. Zhou、K. Hartikainen、G. Tucker、S. Ha、J. Tan、V. Kumar、H. Zhu、A. Gupta、P. Abbeel 等，“软演员-评论家算法及其应用”，*arXiv*预印本*arXiv:1812.05905*，2018年。[37] H. Tan，“基于深度确定性策略梯度的强化学习”，载于*2021*年人工智能、大数据与算法国际会议（*CAIBDA*）。IEEE，2021年，第82–85页。[38] J. Schulman、S. Levine、P. Abbeel、M. Jordan和P. Moritz，“信赖域策略优化”，载于国际机器学习会议。PMLR，2015年，第1889–1897页。[39] V. Konda和J. Tsitsiklis，“演员-评论家算法”，神经信息处理系统进展，第12卷，1999年。[40] I. Osband、C. Blundell、A. Pritzel和B. Van Roy，“通过自举深度Q网络进行深度探索”，神经信息处理系统进展，第29卷，2016年。



Kai Tian 目前正在东北大学软件学院攻读学士学位，学校位于中国沈阳。他的研究兴趣包括具身智能、深度强化学习和机器学习。



Chuan Lin [S’17, M’20]目前是中国沈阳东北大学软件学院副教授。他于2011年在中国沈阳辽宁大学获得计算机科学与技术理学学士学位，2013年在中国沈阳东北大学获得计算机科学与技术理学硕士学位，并于2018年获得计算机体系结构博士学位。2018年11月至2020年11月，他在中国大连大连理工大学软件学院担任博士后研究员，中国。他的研究兴趣包括水下无线传感器网络、工业物联网和软件定义网络。



Guangjie Han [S’03-M’05-SM’18-F’22]目前任河海大学物联网工程系教授，工作地点在中国常州。他于2004年获得中国沈阳东北大学博士学位。2008年2月，他结束了在韩国光州全南大学计算机科学系的博士后研究工作。2010年10月至2011年10月，他在日本吹田大阪大学担任访问研究学者。2017年1月至2017年2月，他曾任中国香港城市大学访问教授。2017年7月至2020年7月，他曾任中国大连理工大学特聘教授。他目前的研究兴趣包括物联网、工业互联网、机器学习与人工智能、移动计算、安全与隐私。Han博士已发表500余篇同行评审期刊和会议论文，此外还拥有160项已授权和待授权专利。目前，他在谷歌学术上的H指数为65，i10指数为282。他的论文总引用次数超过 15500+ 次。Han博士是英国工程技术学会会士（FIET）。他曾担任多达10种国际期刊的编委会成员，包括IEEE工业信息学汇刊、IEEE认知通信与网络汇刊、IEEE车辆技术汇刊、IEEE系统汇刊等。他曾为多种IEEE期刊和杂志客座编辑特刊，包括IEEE通信领域精选期刊、IEEE通信杂志、IEEE无线通信、计算机网络等。Han博士还曾担任多个国际会议组织委员会和技术委员会主席。他曾获2020年IEEE系统期刊年度最佳论文奖和2017—2019年IEEE ACCESS杰出副编辑奖。他是IEEE会士。



Chen An 目前正在中国东北大学计算机科学与工程学院攻读理学学士学位。她的研究兴趣包括情感计算、大语言模型中与情感表达相关的幻觉以及模型可解释性。



Qian Zhu 现任中国沈阳东北大学软件学院副教授。她于2006年获得中国沈阳东北大学信息与计算科学理学学士学位，于2008年获得东北大学运筹学与控制理论理学硕士学位，并于2018年获得中国东北大学通信与信息系统博士学位。她的研究兴趣包括人工智能优化算法、无人飞行器（UAV）技术及其应用软件开发。



Shengchao Zhu 于2023年获得中国常州河海大学物联网工程理学学士学位。目前，他正在中国南京河海大学计算机科学与技术系攻读博士学位。他目前的研究兴趣包括群体智能、群体海洋、多智能体强化学习。



Zhenyu Wang 目前就读于中国沈阳东北大学软件学院，攻读学士学位。他的研究兴趣包括强化学习、物联网和软件定义网络。