

USIM and U0: A Vision-Language-Action Dataset and Model for General Underwater Robots

Junwen Gu^{1,3†}, Zhiheng Wu^{2†}, Pengxuan Si³, Shuang Qiu^{1,3}, Zhentao Zhang^{1,3}, Yukai Feng^{1,3},
Luoyang Sun^{1,3}, Laien Luo^{1,3}, Lianyi Yu^{1,3}, Jian Wang^{1,3}, and Zhengxing Wu^{1,3*}

Abstract—Underwater environments pose unique challenges for robotic navigation and manipulation. While existing research has primarily focused on task-specific methods, studies on general-purpose intelligence for multi-task execution remain scarce. To address this gap, we propose a unified framework for general-purpose underwater robots that integrates perception and action driven by language instructions. First, we develop a data synthesis pipeline to construct USIM, a simulation-based dataset which comprises over 905K frames from 2275 trajectories, totaling approximately 25 hours of BlueROV2 interactions. Furthermore, we propose U0, a vision-language-action (VLA) model capable of executing various tasks from obstacle-avoidance navigation to three-dimensional mobile manipulation. The model features a convolution-attention-based perception (CAP) module, which incorporates target pose estimation as an auxiliary task to explicitly bolster the model’s spatial awareness. For evaluation, we establish a systematic assessment framework and an automated pipeline encompassing both offline metrics and online task execution. Experimental results demonstrate that the USIM dataset significantly empowers existing VLA models to adapt to underwater scenarios. Notably, our U0 model achieves state-of-the-art performance: it reduces the offline mean action prediction error to 0.0359 and achieves an overall online success rate of 43.1%, marking a 5.5% improvement over existing competitive baselines (below 37.6%), with navigation tasks reaching as high as 87.5%. These results validate the feasibility of general-purpose intelligence in underwater robotics, providing a foundation for scalable dataset synthesis and aquatic embodied agents.

I. INTRODUCTION

The underwater environment poses growing demands, encompassing diverse applications such as marine ecological surveys, resource exploitation, inspection of pipelines, and underwater salvage [5], [6]. However, in-situ human diving operations are far more challenging and hazardous than terrestrial or aerial tasks, which necessitating the use of underwater robots to execute these missions. Nevertheless, the development of these systems is hindered by distinctive environmental constraints, including complex hydrodynamics, limited visibility, and restricted communication. Enhancing the autonomy and intelligence of underwater robots is crucial

to overcoming these challenges, as it empowers them to execute complex missions with minimal human intervention.

Existing studies have developed task-specialized frameworks for underwater teleoperation [7], navigation [8], and manipulation [9]. Despite their success, the lack of a unified perspective results in fragmented data islands and complicates the integration of multi-task capabilities. Regarding the development of general embodied intelligence, prior works have explored reinforcement learning based approaches to achieve multi-task capability [10]. In recent years, vision-language-action (VLA) models based on large-scale pre-training have catalyzed significant breakthroughs. By offering high autonomy and generalization, VLA models are inherently suited for complex underwater missions that demand multi-task capabilities. However, current VLA applications are primarily concentrated in robotic arm manipulation and quadrupedal navigation [11]–[14]. Since these methods are predominantly trained on terrestrial data biased toward two-dimensional (2D) planar motion, they are difficult to transfer directly to underwater robots with distinct configurations and three-dimensional (3D) degrees of freedom. Moreover, the vast discrepancies between terrestrial and aquatic scenes regarding visual signals pose significant challenges to a model’s perception [15]. Furthermore, data collection in real underwater environments is prohibitively costly; for instance, even experienced operators require extensive practice to achieve proficiency in teleoperated grasping tasks [16]. Overall, the scarcity of unified frameworks and datasets stands in stark contrast to the urgent demand for higher autonomy and multi-task intelligence in underwater robotics.

To address these challenges, we establish a unified framework for general-purpose underwater robots, comprising the multi-task VLA dataset USIM and the corresponding VLA model U0, which collectively enable sophisticated underwater multi-task autonomy. Specifically, USIM is a simulation-based VLA dataset featuring both perception and control traces for the BlueROV2 platform. Developed via an automated data synthesis pipeline, USIM consists of 905K frames spanning 2275 trajectories and approximately 25 hours of robot-environment interactions. Covering 20 tasks across 9 scenarios ranging from navigation to mobile manipulation, it offers a rapid and cost-effective alternative to real-world data collection. By employing a stochastic exploratory strategy, USIM incorporates sub-optimal states and recovery behaviors, ensuring broad coverage of the state space—a critical factor for learning robust policies in dynamic aquatic environments. Building upon this data foundation,

Junwen Gu and Zhiheng Wu contributed equally to this work. Corresponding author: Zhengxing Wu.

¹The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: gujunwen2022@ia.ac.cn jianwang@ia.ac.cn; zhengxing.wu@ia.ac.cn).

²Baidu Inc., Beijing 100085, China (e-mail: wzh404.ai@gmail.com).

³The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

<https://vincentgu2000.github.io/u0project>

TABLE I: Comparison of Typical Underwater Datasets and USIM Dataset

Dataset	Type	N_{traj}	$N_{\text{fr.}}$	Dur.	Primary Tasks	Modality	Labels	Lang.
Aqualoc [1]	Real	17	–	1.7 h	Visual SLAM, VIO	Monochromatic, IMU, Depth	Est. pose	×
MIMIR-UW [2]	Sim	11	31K	0.4 h	SLAM, Seg., Depth Est.	RGBD, IMU	Pose, Sem. masks	×
SubPipe [3]	Real	5	–	0.7 h	SLAM, Seg., Det.	RGB, SSS, IMU, DVL, Depth	Est. pose, Sem. masks	×
MOUD [4]	Real	–	18K	–	Rec., Seg., Det.	RGB, LiDAR	Instance masks	×
USIM (Ours)	Sim	2275	905K	25 h	Navigation, Manipulation	RGB, IMU, DVL, Depth	Pose, Action	✓

Note: N_{traj} , number of trajectories; $N_{\text{fr.}}$, number of frames; Dur., duration; Lang., language; VIO, visual-inertial odometry; SSS, side scan sonar; Seg., segmentation; Est., estimation; Det., detection; Rec., reconstruction; Sem., semantic.

we propose U0, a 3B-parameter VLA model capable of simultaneously handling navigation and manipulation tasks. It utilizes a unified action space that directly operates low-level actuators, effectively addressing the unique embodiment of underwater robots that combines thruster locomotion with manipulator control. To tackle the challenges of expansive environments and sparse targets, a convolution-attention-based perception (CAP) module is designed to predict the poses of navigation points or grasping targets. This auxiliary training task explicitly bolsters the model’s spatial awareness by introducing geometric constraints. To rigorously verify our approach, we establish a systematic assessment framework and an automated pipeline encompassing both offline metrics and online closed-loop execution. Extensive evaluations demonstrate that fine-tuning on the USIM dataset significantly reduces the action prediction error (APE) across all models, with U0 achieving the lowest APE of 0.0359. Furthermore, online results show that our framework achieves a state-of-the-art (SOTA) 43.1% multi-task success rate, marking a 5.5% absolute improvement over competitive baselines. Notably, U0 reaches an 87.5% success rate in navigation tasks and 30.0% in challenging 3D mobile manipulation—a domain where existing VLA models fail to perform (0% success) without fine-tuning.

Overall, we present a unified framework that integrates a dataset with a VLA model tailored for diverse underwater missions. To the best of our knowledge, this represents the first effort to jointly address multi-task underwater perception and action guided by natural language instructions. The main contributions of this paper are summarized as follows:

- A scalable, automated data synthesis pipeline and the USIM dataset, which significantly enhances model adaptability and policy robustness in dynamic aquatic environments.
- A general-purpose underwater VLA model, U0, featuring enhanced spatial awareness for precise perception and execution across diverse navigation and manipulation missions.
- A systematic assessment framework and a unified multi-task autonomy pipeline, setting a new benchmark for underwater robotics and demonstrating the feasibility of general-purpose underwater intelligence.

The remainder of this paper is organized as follows. Sec-

tion II reviews related work. Section III details the proposed methodology. Section IV presents experimental results and ablation studies. Section V concludes the paper.

II. RELATED WORK

A. Data-Driven Underwater Autonomy

Data-driven methods such as reinforcement learning (RL) and imitation learning (IL) have achieved significant progress in the underwater domain, covering path planning [17], multi-agent formation and obstacle avoidance [18], target tracking [19], [20], autonomous navigation [21], and mobile manipulation [16], [22]. However, most of these approaches are designed as expert models—highly specialized systems utilizing task-specific observation and action spaces. This specialization results in a lack of versatility when facing diverse natural language instructions and cross-task scenarios. While large language models [23] and vision-based deep learning methods [24] have been introduced to improve high-level cognition and human-robot interaction, they often lack adaptive mapping to low-level controllers. This fragmented technical trajectory limits the evolution of underwater robots toward high-level embodied intelligence. To bridge this gap, we integrate navigation and manipulation tasks into a unified state-action space, constructing U0 as a cohesive framework capable of understanding natural language and executing multiple tasks simultaneously.

B. Vision-Language-Action Models

Vision-language-action models have demonstrated exceptional generalization capabilities in terrestrial robotics by combining large-scale pre-training with embodied tasks. Representative models such as RT-2 [25], OpenVLA [13], the π series [12], [26], and GR00T [11] have proven that through end-to-end policy learning—utilizing architectures like vision-language models, mixture-of-experts, or dual systems—robots can achieve zero-shot generalization across diverse tasks. Recent studies have also extended VLA to unmanned aerial vehicle navigation by incorporating depth information and multi-stage training [27], [28]. However, VLA models specifically designed for underwater robotics remain exceedingly rare. The primary challenge lies in the significant perceptual domain shift caused by underwater conditions [15]. Addressing this, we introduce the CAP module in the training of U0, which incorporates target pose estimation as an auxiliary

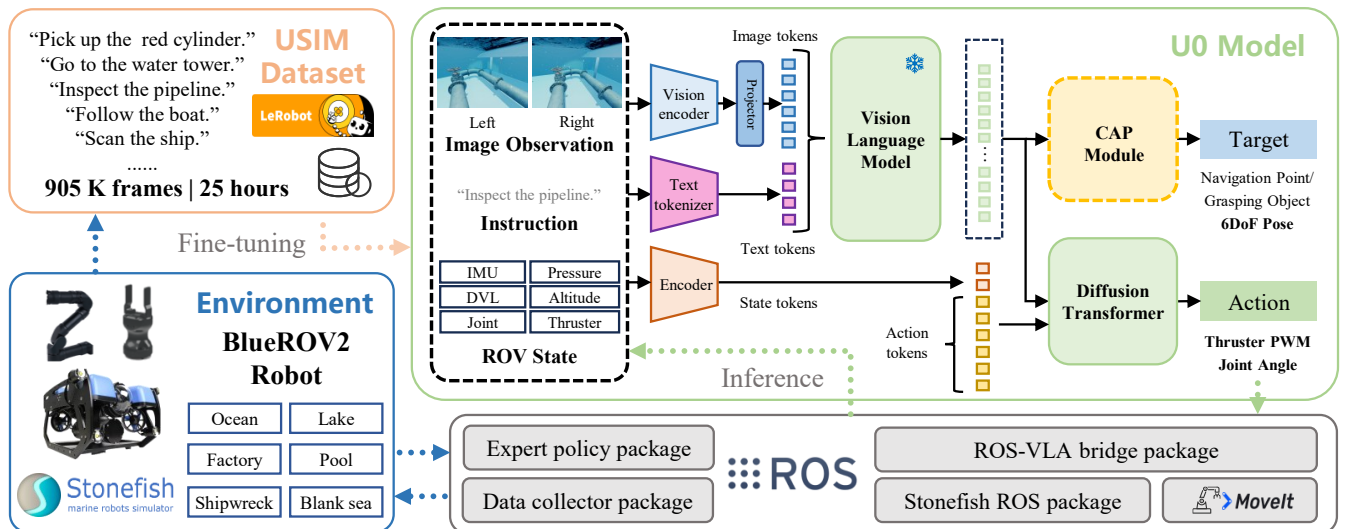


Fig. 1: Overall Framework. Diverse underwater scenarios and a BlueROV2 robot equipped with a manipulator are first constructed within the Stonefish simulator. Data collection and control are managed via ROS, yielding the USIM dataset of 905K frames (approximately 25 hours) across 20 tasks. The U0 model is then designed, which integrates a DiT action module and a CAP module, generating concurrent target perception and robot action outputs.

task, effectively enhancing the model’s spatial perception and its adaptability to aquatic environments.

C. Underwater Datasets

Current underwater datasets primarily focus on perception-only or isolated navigation tasks. For instance, datasets like Caves [29], AQUALOC [1], and AquaticVision [30] support simultaneous localization and mapping (SLAM), while UIEB [31] and the Marine Debris Dataset [32] target image enhancement and segmentation. To mitigate data scarcity, synthetic alternatives like VAROS [33] and MIMIR-UW [2] have also been introduced. Recently, multi-task datasets have emerged, including MOUD [4] (detection, segmentation, 3D reconstruction) and SubPipe [3] (SLAM and inspection). As categorized in Table I, while these pioneering datasets establish a solid foundation, scaled multi-task data encompassing both manipulation and diverse skills—akin to terrestrial counterparts like Open X-Embodiment [34]—remains a valuable next step for full underwater autonomy. Given the prohibitive costs of real-world underwater data collection, we offer a practical alternative by developing an automated data synthesis pipeline. This pipeline generates USIM, a simulation-based multi-task dataset comprising 20 distinct tasks and 2,275 trajectories, thereby enhancing the adaptability and versatility of VLA models in dynamic aquatic environments.

III. METHOD

The overall framework of this work is illustrated in Fig. 1, comprising an integrated underwater simulation environment, a multi-task dataset USIM, and the VLA model U0. Detailed descriptions of the task formulation, dataset construction, and model architecture are provided in the following.

A. Task Formulation

We aim to enable a single, unified policy π to execute a diverse array of underwater missions, including navigation, grasping, transport, and tracking. The policy is formulated as $\pi : (V_t, L, S_t, A_t) \rightarrow A_{t+1}$, where V_t denotes multi-view visual observations from two RGB cameras, L is the natural language instruction, and S_t represents the sensory data. This sensory input integrates inertial measurement unit (IMU), doppler velocity log (DVL), and pressure-based depth readings, as well as thruster commands and robotic arm joint states (angles and angular velocities). The action space A_t and A_{t+1} encompasses both desired propeller thrust intensities and robotic arm joint positions.

B. USIM Dataset Construction

Given the prohibitive costs associated with real-world underwater experiments, simulation offers an efficient alternative for data collection. Consequently, we developed a simulation-based pipeline to construct the USIM dataset using an automated collection process integrated with a stochastic exploration strategy. Specifically, we utilized the Stonefish simulator [35] to construct 9 distinct scenarios, including seabed environment, subsea pipeline, industrial pool, solar charging station, lake environment, open sea surface, underwater factory, modern shipwreck, and ancient shipwreck. Furthermore, a BlueROV2 platform, optionally equipped with a 4-degree-of-freedom (DoF) arm and a parallel gripper, was deployed. Leveraging the Stonefish ROS package [36], we established a standardized framework that remains compatible with other underwater simulators supporting ROS protocols.

To ensure data diversity, we implemented scene randomization and exploratory collection strategies. Specifically,

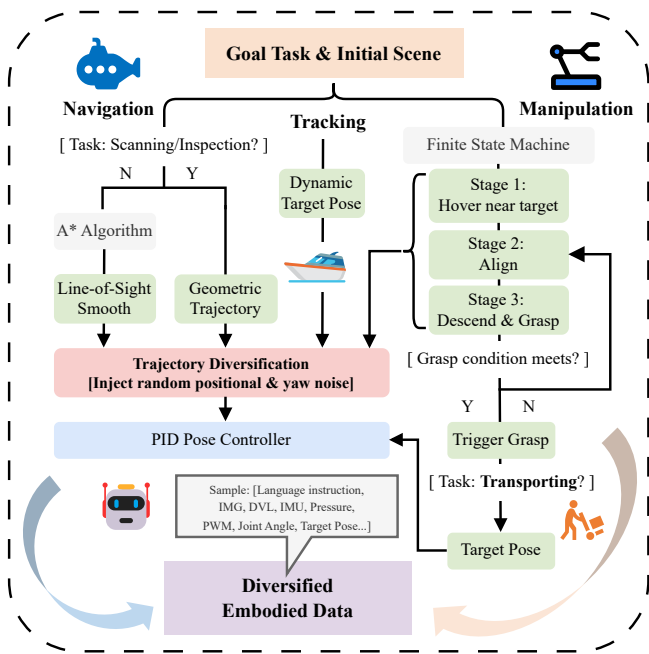


Fig. 2: Automated data collection pipeline. Hand-crafted expert policies generate diverse navigation and manipulation trajectories in simulation. Stochastic noise injection is integrated to automatically capture sub-optimal behaviors and recovery sequences.

scene randomization modules were utilized to vary object placement and lighting conditions. For data collection, we employed hand-crafted expert strategies that leverage ground-truth state information available only in simulation, as shown in Fig. 2. These strategies enable fully automated data collection and, by introducing stochastic exploration, ensure that the dataset encompasses diverse trajectories, including sub-optimal states and recovery behaviors, which are critical for training robust policies.

Under this framework, we constructed the USIM dataset, comprising 20 tasks across 9 scenarios, with a total of 905K frames and 2275 trajectories (sampled at 10 Hz, approximately 25 hours). The 20 tasks include 12 grasping, 6 navigation, 1 dynamic tracking, and 1 transport task. Table II provides a detailed breakdown of the task distributions. We intentionally weighted the dataset toward grasping tasks due to the inherent complexity of 3D mobile manipulation. Since trajectories of mobile manipulation involve significant robot locomotion, they concurrently serve to improve the model’s fundamental motion control, thereby benefiting navigation performance. Table I highlights the unique advantages of USIM, specifically its unprecedented scale, diverse task coverage, and the seamless integration of language descriptors with precise action labels, which are absent in previous underwater datasets.

C. U0 Model Architecture

U0 is a general-purpose underwater VLA model initialized with the pre-trained weights of Isaac-GR00T N1.5 [11], as illustrated in Fig. 1. The model accepts multi-modal data

TABLE II: Data Distribution of USIM

Category	Total frames	Episodes	Average duration
Grasping	585K (64.6%)	1560	37.5 ± 8.7 s
Navigation	231K (25.5%)	520	44.4 ± 24.3 s
Transport	66K (7.3%)	130	51.1 ± 5.0 s
Tracking	23K (2.5%)	65	35.5 ± 2.5 s
Total	905K (100.0%)	2275	39.8 ± 14.3 s

as input, where visual images and language instructions are processed through a VLM and mapped to 2048-dimensional embedding. Simultaneously, additional sensor modalities are encoded via a multi-layer perceptron (MLP), while robot action data are mapped to features of the same dimension through an MLP following the addition of positional embeddings. These features are then fed into a diffusion transformer (DiT) action module to generate action chunks. Concurrently, the VLM features are passed through the CAP module. By leveraging convolution-attention mechanisms, this module enhances visual feature representations to improve the model’s ability to correlate underwater motion states with task objectives. Details regarding multi-sensor integration, the CAP module, and training objectives are elaborated below.

1) *Multi-Source Sensing and Unified Action Space*: Unlike terrestrial robots, underwater robots operate predominantly in a floating state and must coordinate locomotion with manipulator actions. They also frequently execute depth changes, in contrast to most ground robots limited to horizontal planes. To address these challenges, we incorporate underwater-specific sensors into the state space for accurate localization, including depth sensor, IMU, and DVL. Furthermore, dual-view camera images are provided as input to enhance visual coverage. The action space of underwater robots also differs significantly from terrestrial counterparts. Beyond manipulator control, locomotion is primarily achieved through multiple thrusters, whose control signals occupy a distinct force space. We normalize the thruster PWM signals and concatenate them with the manipulator joint angles to construct a unified action space. Both thruster commands and joint states are included in the state space to improve the model’s temporal understanding of the robot’s current physical configuration.

2) *Convolution-Attention-based Perception Module*: Due to the vast spatial scales and sparse visual features in aquatic environments, we designed the CAP module to bolster visual comprehension and spatial awareness. Guided by VLM features, this module is optimized during training to predict a 6-DoF pose. Specifically, for grasping tasks, this represents the pose of the target objects; for navigation, it represents the next waypoint; for transport tasks, it predicts the object pose during grasping and the target waypoint during transfer; and for tracking, it predicts the dynamic pose of the tracked target. By explicitly supervising these points of interest, we direct the model’s attention toward task-relevant features. The architecture of the CAP module, illustrated in Fig. 3, consists of convolutional layers, convolution-based attention

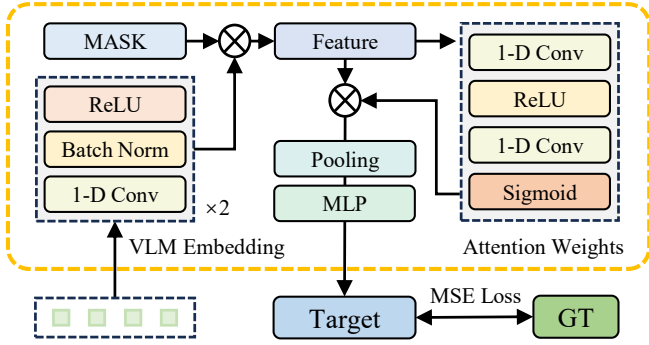


Fig. 3: The structure of the CAP module. Through combining convolutional layers and attention pooling, feature focus is refined for enhanced perception.

weighting, pooling operations, and a multilayer perceptron (MLP). The computation process is formulated as

$$\begin{aligned} \phi_t &= \text{VLM}(V_t, L), \quad F = \text{Conv}(\phi_t, \text{MASK}), \\ \text{Att} &= \text{Conv}(F), \quad T = \text{MLP}(\text{Pool}(F \cdot \text{Att})), \end{aligned} \quad (1)$$

where $\text{Conv}(\cdot)$ and $\text{Pool}(\cdot)$ refer to convolutional and pooling operations respectively, while MASK prevents computation over padding features. Att represents channel-wise attention, F is the intermediate feature map, and T is the predicted 6-DoF target.

3) *Ground Truth and Training Objectives*: Based on our task formulation, since the policy outputs a unified action space, we construct a standardized ground truth (GT) across all tasks to supervise training. For action module, We adopt an action-chunking approach where actions for H consecutive timesteps are treated as the GT. Each action includes the thruster intensities and joint angles for that moment. Following the flow-matching objective, we sample a noise vector and a denoising timestep τ to construct the noisy action

$$A_t^\tau = \tau A_t + (1 - \tau)\epsilon, \quad (2)$$

where ϵ is the sampled noise and A_t is the GT action. The velocity field corresponding to this noisy action is defined as $v_\tau = A_t - \epsilon$. The DiT action module V_θ is trained to predict this velocity field, resulting in the action loss

$$L_{\text{action}} = \mathbb{E}_\tau [\|V_\theta(\phi_t, A_t^\tau, S_t) - v_\tau\|^2]. \quad (3)$$

During inference, a Gaussian noise sample is used as the initial action input, and the final action is obtained through K -step forward Euler integration

$$A_t^{\tau+1/K} = A_t^\tau + \frac{1}{K} V_\theta(\phi_t, A_t^\tau, S_t). \quad (4)$$

This requires the DiT to perform K forward passes. In this work, we set $H = 16$ and $K = 4$.

For CAP module, target poses are treated as the GT. To enable effective reasoning relative to the robot's current state, we represent target poses in a robot-centric coordinate system. Let the 6-DoF poses of the target and the robot in the world coordinate system be

$$p_g = (R_g, \mathbf{t}_g), \quad p_r = (R_r, \mathbf{t}_r), \quad (5)$$

where $R \in SO(3)$ denotes the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ the translation vector. The target pose in the robot-centric frame is

$$T_{\text{gt}} = (R_r^\top R_g, R_r^\top (\mathbf{t}_g - \mathbf{t}_r)). \quad (6)$$

The CAP training loss is

$$L_{\text{CAP}} = \mathbb{E} [\|T - T_{\text{gt}}\|^2]. \quad (7)$$

The total training objective is to minimize L , which is the weighted sum of the action and perception losses

$$L = L_{\text{action}} + \alpha L_{\text{CAP}}, \quad (8)$$

where the weighting factor α is set to 1.0.

IV. EXPERIMENTS

In this section, we define a comprehensive set of evaluation metrics and conduct experiments to assess the performance of the proposed framework across multiple tasks. To demonstrate the efficacy of our approach, we compare our framework against SOTA VLA methods and perform ablation studies to validate the contribution of each component.

A. Data Split

We allocated 77% (697K frames) of USIM for training and 23% (208K frames) for testing. To ensure evaluation integrity, both sets were collected under identical environmental conditions and follow the same statistical distribution.

B. Implementation Details

U0 was trained with a total batch size of 256 for 50000 steps on a server equipped with 8 NVIDIA H20 GPUs. In addition to U0, OpenVLA [13], $\pi_{0.5}$ [12], and GR00T N1.5 [11] are selected as the comparison methods. These models were also fine-tuned on the USIM training set.

C. Evaluation Metrics and Setup

We evaluate model performance across both offline and online dimensions, adopting standard metrics widely established in vision-language-action and vision-language navigation tasks. Offline evaluation assesses policy imitation accuracy on the test dataset, primarily focusing on the action prediction error (APE), which is the mean squared error between the predicted and ground-truth actions. Online evaluation measures closed-loop control efficacy within the simulation environment using five complementary metrics: success rate (SR), stage success rate (SSR), success-weighted path length (SPL) [37], average success duration (ASD), and mean target distance (MTD). For the deployment setup, evaluations are conducted on a server equipped with an Intel Xeon Silver 4214R CPU and four NVIDIA RTX 3090 GPUs. To ensure optimal performance, we customize the control and inference loops tailored to the architectural characteristics of each model. Specifically, for GR00T N1.5 and U0, the policy operates at a control frequency of 10 Hz, with model inference triggered every 1.6 s to predict 16-step action chunks. For $\pi_{0.5}$, the model performs recurrent inference at 10 Hz, with the robot executing only the first step of the predicted action chunk at each interval. Conversely, for

TABLE III: Offline APE Results on USIM Test Set

Model	Setting	Navigation	Grasping	Transport	Tracking	Overall
OpenVLA	Zero-shot	0.5672	0.1872	0.2160	0.3810	0.2920
	Fine-tuned	0.3281	0.0768	0.0666	0.0827	0.1410
$\pi_{0.5}$	Zero-shot	0.2363	0.1217	0.1181	0.0591	0.1496
	Fine-tuned	0.2087	0.0423	0.0475	0.0528	0.0861
GR00T N1.5	Zero-shot	0.3459	0.1220	0.1452	0.1906	0.1834
	Fine-tuned	0.0974	0.0161	0.0158	0.0272	0.0374
U0 (Ours)	Fine-tuned	0.0935	0.0154	0.0141	0.0310	0.0359

OpenVLA, the policy operates in a closed-loop manner where the model executes inference as rapidly as possible and immediately applies the predicted action to the robot.

D. Offline Evaluation and Comparison

We compared the proposed model with several baselines across 525 trajectories in the USIM test set, with results summarized in Table III. In the zero-shot setting, OpenVLA, $\pi_{0.5}$, and GR00T N1.5 all exhibit severe APE. This reveals a substantial perceptual and action-space domain shift between terrestrial and complex underwater environments, highlighting the difficulty of direct cross-domain transfer. However, after fine-tuning on our USIM dataset, the APE of all baseline models improves drastically, yielding an average error reduction of 59.7% across the three competitive methods. This performance leap demonstrates the effectiveness and necessity of the USIM dataset. It successfully bridges the underwater perceptual barrier, empowering existing VLA models with cross-modal adaptability and policy generalization across diverse aquatic multi-task scenarios.

Among all the fine-tuned models, our proposed U0 model achieves the SOTA performance, reducing the overall APE to a minimum of 0.0359. Although GR00T N1.5 already exhibits exceptionally low error due to its robust design, U0 further pushes the performance boundary (0.0359 vs. 0.0374). This is attributed to the integration of the CAP module and auxiliary pose estimation tasks, which allow the model to capture sparse underwater features more accurately and enhance 3D spatial awareness.

On the whole, the offline evaluation results validate the necessity and feasibility of utilizing the USIM dataset to train underwater embodied policies, while fully demonstrating the superiority of the CAP-based auxiliary training strategy in improving high-precision aquatic control.

E. Closed-Loop Autonomy and Efficiency Analysis

We conducted closed-loop evaluations in the simulation environment. As the first VLA framework designed for multi-task underwater operations, we utilize U0 to establish the initial online performance benchmark. Specifically, we conducted 40 independent trials for each task, except for the 4 inspection and 1 tracking tasks, which involved 20 trials each. Generous maximum execution times, ranging from 180 to 450 s, were tailored to each mission based on its difficulty, with any exceeding duration resulting in task failure. The positions of the robot, environmental obstacles, target objects, and waypoints were randomized to ensure

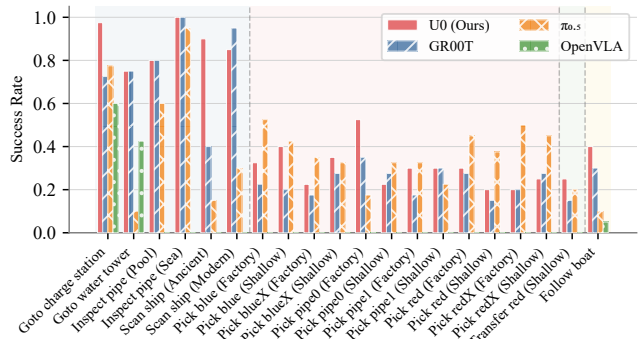


Fig. 4: Breakdown of success rates across 20 individual tasks.

statistical significance before each trial. Physically grounded success criteria were also defined. Grasping requires the robot to maintain a stable grip for at least 3s after relocation; navigation and transport success is defined by the robot or target object reaching a task-specific radius of the goal; and tracking success requires maintaining the dynamic target within the field of view and within a safety distance threshold for a required cumulative duration.

The success rates across 20 distinct tasks, compiled from a total of 700 independent trials, are summarized in Table IV. As observed, fine-tuning on the USIM dataset successfully unlocks the underwater execution capabilities of all three terrestrial VLA baselines, underscoring the dataset’s efficacy. Among the baselines, $\pi_{0.5}$ exhibits competitive performance in grasping-related tasks, while GR00T N1.5 excels in navigation missions. Conversely, OpenVLA underperforms across the board; this is primarily attributed to its high computational overhead and lower inference frequency, which incur noticeable control latency during complex 3D locomotion and fine-grained manipulation.

Benefiting from the integration of the CAP module, our proposed U0 achieves a SOTA overall SR of 43.1%, outperforming the strongest baseline by a margin of 5.5%. Fig. 4 breaks down the performance across all 20 individual tasks. Specifically, the model demonstrates exceptional proficiency in navigation tasks, achieving an 87.5% SR and an average SPL of 0.71. This high SPL indicates that the executed trajectories closely approximate the optimal geometric paths, reflecting superior path efficiency. In manipulation tasks, U0 also delivers robust performance, securing a 30.0% SR in grasping. Furthermore, in the long-horizon transport task, which demands a seamless integration of navigation and manipulation capabilities, U0 achieves a 25.0% SR—surpassing comparison methods by 5.0%. In dynamic target tracking, U0 boosts the SR by 10.0% and maintains a significantly tighter tracking loop, reducing the MTD to 3.61 m.

F. Case Studies and Robustness Verification

Fig. 5 provides snapshots of the robot’s performance during real-time execution. For grasping and transport tasks, the figures display two distinct visual fields: the ego view and the wrist view. Conversely, in navigation and tracking tasks, where the robot is not equipped with a manipulator, the dual

TABLE IV: Online Evaluation Performance Comparison

Model	Navigation		Grasping		Transporting		Tracking		Overall SR (succeed/total trials)
	SR \uparrow	SPL \uparrow	SR \uparrow	ASD \downarrow	SR \uparrow	SSR \uparrow	SR \uparrow	MTD \downarrow	
OpenVLA	25.6%	0.70 \pm 0.21	0.0%	–	0.0%	0.0%	5.0%	4.02 \pm 0.00 m	6.0% (42/700)
$\pi_{0.5}$	46.9%	0.87 \pm 0.18	37.1%	97.3 s	20.0%	45.0%	10.0%	4.52 \pm 0.06 m	37.6% (263/700)
GR00T N1.5	76.3%	0.69 \pm 0.21	24.0%	92.3 s	15.0%	30.0%	30.0%	4.16 \pm 0.14 m	35.6% (249/700)
U0 (Ours)	87.5%	0.71 \pm 0.23	30.0%	87.6 s	25.0%	45.0%	40.0%	3.61 \pm 0.35 m	43.1% (302/700)

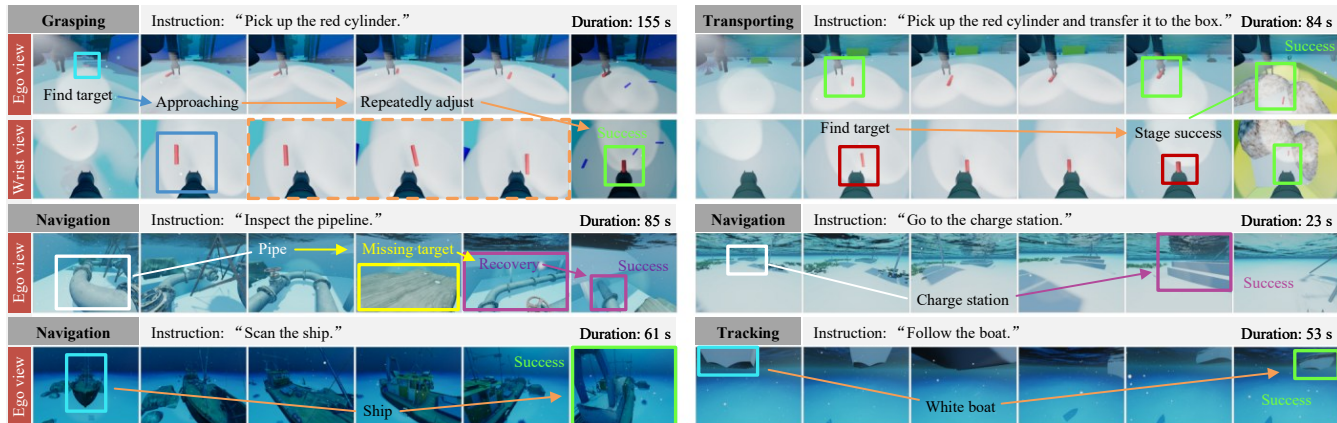


Fig. 5: Real-time execution traces of U0 across various tasks.

views are natively provided by a stereo camera setup on the robot body; thus, only one of these ego views is displayed.

The grasping sequences reveal that the dynamic nature of the robot—requiring simultaneous station-keeping and manipulation—imposes rigorous demands on the model’s motion control. Notably, the trajectories demonstrate that U0 can autonomously initiate re-grasping attempts following an initial failure, signifying the emergence of self-recovery behaviors. The average duration for a successful grasp during online testing is 87.6 s, which, though the fastest among all evaluated models, is still over twice that of the expert strategy (37.5 s), highlighting potential for further policy optimization.

G. Ablation Studies

To further validate the efficacy of our framework, we conducted extensive ablation studies. The variant trained without the CAP module are analyzed using the same protocol of 700 trials. Table V presents detailed ablation results on 20 tasks. The data indicates that the CAP-based auxiliary training branch provides substantial performance gains across all task types, improving the overall success rate by 7.5%. This suggests that providing explicit guidance regarding the model’s points of interest during training significantly enhances the model’s task comprehension.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a unified VLA framework to enhance the multi-task autonomy of underwater robots. At first, we establish USIM, a large-scale simulation-based dataset generated through an automated synthesis pipeline, which

TABLE V: Ablation Study on the CAP Module of U0

Navigation			Navigation		
Model	SR \uparrow	SPL \uparrow	Model	SR \uparrow	SPL \uparrow
<i>Goto charge station</i>			<i>Goto water tower</i>		
w/	97.5% (39/40)	0.611	w/	75.0% (30/40)	0.822
w/o	72.5% (29/40)	0.576	w/o	75.0% (30/40)	0.859
<i>Inspect pipe (Pool)</i>			<i>Inspect pipe (Sea)</i>		
w/	80.0% (16/20)	0.714	w/	100.0% (20/20)	0.547
w/o	80.0% (16/20)	0.789	w/o	100.0% (20/20)	0.556
<i>Scan ship (Ancient)</i>			<i>Scan ship (Modern)</i>		
w/	90.0% (18/20)	0.978	w/	85.0% (17/20)	0.647
w/o	40.0% (8/20)	0.760	w/o	95.0% (19/20)	0.658
Grasping			Grasping		
Model	SR \uparrow	ASD \downarrow	Model	SR \uparrow	ASD \downarrow
<i>Pick blue (Factory)</i>			<i>Pick blue (Shallow)</i>		
w/	32.5% (13/40)	113.35	w/	40.0% (16/40)	78.50
w/o	22.5% (9/40)	83.13	w/o	20.0% (8/40)	71.45
<i>Pick blueX (Factory)</i>			<i>Pick blueX (Shallow)</i>		
w/	22.5% (9/40)	78.79	w/	35.0% (14/40)	101.89
w/o	17.5% (7/40)	109.80	w/o	27.5% (11/40)	63.08
<i>Pick pipe0 (Factory)</i>			<i>Pick pipe0 (Shallow)</i>		
w/	52.5% (21/40)	85.66	w/	22.5% (9/40)	79.53
w/o	35.0% (14/40)	97.23	w/o	27.5% (11/40)	103.89
<i>Pick pipe1 (Factory)</i>			<i>Pick pipe1 (Shallow)</i>		
w/	30.0% (12/40)	90.88	w/	30.0% (12/40)	86.56
w/o	17.5% (7/40)	100.38	w/o	30.0% (12/40)	89.81
<i>Pick red (Factory)</i>			<i>Pick red (Shallow)</i>		
w/	30.0% (12/40)	91.73	w/	20.0% (8/40)	48.81
w/o	27.5% (11/40)	111.53	w/o	15.0% (6/40)	90.54
<i>Pick redX (Factory)</i>			<i>Pick redX (Shallow)</i>		
w/	20.0% (8/40)	75.55	w/	25.0% (10/40)	100.92
w/o	20.0% (8/40)	73.80	w/o	27.5% (11/40)	107.64
Transporting			Tracking		
Model	SR \uparrow	SSR \uparrow	Model	SR \uparrow	MTD \downarrow
<i>Transfer red (Shallow)</i>			<i>Follow boat</i>		
w/	25.0% (10/40)	45.0%	w/	40.0% (8/20)	3.61
w/o	15.0% (6/40)	30.0%	w/o	30.0% (6/20)	4.16
Overall Success Rate (succeed/total)			Overall Success Rate (succeed/total)		
w/					43.1% (302/700)
w/o					35.6% (249/700)

provides a diverse foundation for robust policy learning in dynamic aquatic environments. Widespread fine-tuning experiments demonstrated that USIM successfully unlocks the zero-to-one capability of existing terrestrial VLA models in executing underwater tasks, underscores its profound efficacy and cross-domain value. Building upon this dataset, we develop U0, a 3B-parameter VLA model specifically optimized for the unique embodiment and perception challenges of underwater robots. By incorporating a CAP module, U0 achieves enhanced spatial awareness, effectively bridging the gap between high-level instructions and low-level 6-DoF actuation. Our extensive closed-loop evaluations demonstrate that U0 significantly outperforms existing terrestrial-based VLA models, improving the overall success rate by 5.5% and achieving an exceptional 87.5% success rate in navigation, while exhibiting superior efficiency and robustness in long-horizon tasks and dynamic environments.

Future work will focus on incorporating supplementary modalities, such as sonar, to improve perception in low-visibility environments, eventually moving toward real-world deployment and field validation.

REFERENCES

- [1] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, “Aqualoc: An underwater dataset for visual-inertial–pressure localization,” *Int. J. Robot. Res.*, vol. 38, no. 14, pp. 1549–1559, Dec. 2019.
- [2] O. Álvarez-Tuñón *et al.*, “MIMIR-UW: A multipurpose synthetic dataset for underwater navigation and inspection,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Detroit, MI, USA, Oct. 2023, pp. 6141–6148.
- [3] O. Álvarez-Tuñón, L. R. Marnet, M. Aubard, L. Antal, M. Costa, and Y. Brodskiy, “SubPipe: A submarine pipeline inspection dataset for segmentation and visual-inertial localization,” in *OCEANS 2024 - Singapore*, Apr. 2024, pp. 1–7.
- [4] X. Chu *et al.*, “A multimodal optical dataset for underwater image enhancement, detection, segmentation, and reconstruction,” *Sci. Data*, vol. 12, no. 1, p. 1554, Sep. 2025.
- [5] T. Liu *et al.*, “A bioinspired multimotion modality underwater micro-robot,” *Sci. Adv.*, vol. 11, no. 19, May 2025, Art. no. eadu2527.
- [6] R. K. Katzschmann, J. DelPreto, R. MacCurdy, and D. Rus, “Exploration of underwater life with an acoustically controlled soft robotic fish,” *Sci. Robot.*, vol. 3, no. 16, Mar. 2018, Art. no. eaar3449.
- [7] A. Phung, G. Billings, A. F. Daniele, M. R. Walter, and R. Camilli, “A shared autonomy system for precise and efficient remote underwater manipulation,” *IEEE Trans. Robot.*, vol. 40, pp. 4147–4159, 2024.
- [8] C. Hu, S. Zhu, Y. Liang, and W. Song, “Tightly-coupled visual-inertial-pressure fusion using forward and backward IMU preintegration,” *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6790–6797, Jul. 2022.
- [9] E. Palmer, C. Holm, and G. Hollinger, “Angler: An autonomy framework for intervention tasks with lightweight underwater vehicle manipulator systems,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Yokohama, Japan, May 2024, pp. 6126–6132.
- [10] F. Bai, H. Zhang, T. Tao, Z. Wu, Y. Wang, and B. Xu, “Picor: Multi-task deep reinforcement learning with policy correction,” in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, Jun. 2023, pp. 6728–6736.
- [11] NVIDIA *et al.*, “GR00T N1: An open foundation model for generalist humanoid robots,” *arXiv preprint arXiv:2503.14734*, Mar. 2025.
- [12] Physical Intelligence *et al.*, “ $\pi_{0.5}$: A vision-language-action model with open-world generalization,” *arXiv preprint arXiv:2504.16054*, Apr. 2025.
- [13] M. J. Kim *et al.*, “OpenVLA: An open-source vision–language–action model,” *arXiv preprint arXiv:2406.09246*, Sep. 2024.
- [14] X. Tong *et al.*, “Quart–online: Latency-free multimodal large language model for quadruped robot learning,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Atlanta, GA, USA, May 2025, pp. 9533–9539.
- [15] W. Xu *et al.*, “NAUTILUS: A large multimodal model for underwater scene understanding,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, San Diego, USA, Dec. 2025.
- [16] R. Liu, H. Ha, M. Hou, S. Song, and C. Vondrick, “Self-improving autonomous underwater manipulation,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2025, pp. 16915–16922.
- [17] J. Yang, J. Ni, M. Xi, J. Wen, and Y. Li, “Intelligent path planning of underwater robot based on reinforcement learning,” *IEEE Trans. Automat. Sci. Eng.*, vol. 20, no. 3, pp. 1983–1996, Jul. 2023.
- [18] Z. Fang, T. Chen, T. Shen, D. Jiang, Z. Zhang, and G. Li, “Multi-agent generative adversarial interactive self-imitation learning for AUV formation control and obstacle avoidance,” *IEEE Robot. Autom. Lett.*, vol. 10, no. 5, pp. 4356–4363, May 2025.
- [19] Y. Wang *et al.*, “Target tracking control of a biomimetic underwater vehicle through deep reinforcement learning,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 33, no. 8, pp. 3741–3752, Aug. 2022.
- [20] I. Masmitija *et al.*, “Dynamic robotic tracking of underwater targets using reinforcement learning,” *Sci. Robot.*, vol. 8, no. 80, Jul. 2023, Art. no. eade7811.
- [21] X. Lin *et al.*, “Uivnav: Underwater information-driven vision-based navigation via imitation learning,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Yokohama, Japan, May 2024, pp. 5250–5256.
- [22] J. Gao, Y. Li, Y. Chen, Y. He, and J. Guo, “An improved SAC-based deep reinforcement learning framework for collaborative pushing and grasping in underwater environments,” *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–14, 2024.
- [23] M. Buchholz, I. Carlucho, M. Grimaldi, and Y. R. Petillot, “Distributed AI agents for cognitive underwater robot autonomy,” *arXiv preprint arXiv:2507.23735*, Aug. 2025.
- [24] A. Gomez Chavez, A. Ranieri, D. Chiarella, and A. Birk, “Underwater vision-based gesture recognition: A robustness validation for safe human–robot interaction,” *IEEE Robot. Automat. Mag.*, vol. 28, no. 3, pp. 67–78, Sep. 2021.
- [25] B. Zitkovich *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” in *Proc. 7th Conf. Robot Learning (CoRL)*, vol. 229, Atlanta, GA, USA, Nov. 2023, pp. 2165–2183.
- [26] K. Black *et al.*, “ π_0 : A vision–language–action flow model for general robot control,” in *Proc. Robot.: Sci. Syst. (RSS)*, Delft, Netherlands, Nov. 2024.
- [27] X. Sun *et al.*, “Autofly: Vision-language-action model for UAV autonomous navigation in the wild,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, Rio de Janeiro, Brazil, Feb. 2026.
- [28] Y. Wu *et al.*, “VLA-AN: An efficient and onboard vision-language-action framework for aerial navigation in complex environments,” *arXiv preprint arXiv:2512.15258*, Dec. 2025.
- [29] A. Mallios, E. Vidal, R. Campos, and M. Carreras, “Underwater caves sonar data set,” *Int. J. Robot. Res.*, vol. 36, no. 12, pp. 1247–1251, Oct. 2017.
- [30] Y. Peng, Y. Hong, Z. Hong, A. P.-Y. Chui, and J. Wu, “Aquaticvision: Benchmarking visual SLAM in underwater environment with events and frames,” *arXiv preprint arXiv:2505.03448*, Jun. 2025.
- [31] C. Li *et al.*, “An underwater image enhancement benchmark dataset and beyond,” *IEEE Trans. on Image Process.*, vol. 29, pp. 4376–4389, 2020.
- [32] D. Singh and M. Valdenegro-Toro, “The marine debris dataset for forward-looking sonar semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCV)*, Montreal, Canada, Oct. 2021, pp. 3741–3749.
- [33] P. Georg Olofsson Zwilmeyer, M. Yip, A. Langeland Teigen, R. Mester, and A. Stahl, “The VAROS synthetic underwater data set: Towards realistic multi-sensor underwater data with ground truth,” in *Proc. 2021 IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCV)*, Montreal, Canada, Oct. 2021, pp. 3715–3723.
- [34] A. O’Neill *et al.*, “Open X-embodiment: Robotic learning datasets and RT-X models,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Yokohama, Japan, May 2024, pp. 6892–6903.
- [35] M. Grimaldi *et al.*, “Stonefish: Supporting machine learning research in marine robotics,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Atlanta, GA, USA, May 2025, pp. 13 605–13 611.
- [36] P. Cieslak, “Stonefish: An advanced open-source simulation tool designed for marine robotics, with a ROS interface,” in *Proc. OCEANS 2019 - Marseille*, Jun. 2019, pp. 1–6.
- [37] P. Anderson *et al.*, “On evaluation of embodied navigation agents,” *arXiv preprint arXiv:1807.06757*, Jul. 2018.