

Object-Centric State-Guided SAM for Temporally Coherent Multi-Object Segmentation in Complex Underwater Environments

Wenbin Feng¹, Yu Lu^{1,2,*}, Yuetong Lin¹, Meng Li¹, Qiang Li¹

Abstract—Robust perception in complex underwater environments is critical for marine resource exploration, ecological monitoring, and robotic operations, yet remains challenging due to low visibility, color distortion, dynamic backgrounds, and long-term occlusions. Although foundation models such as SAM 2 demonstrate strong segmentation performance on static imagery, their lack of explicit temporal modeling limits reliability in underwater video scenarios. To address this limitation, we propose AquaSlot-SAM, a hybrid object-centric temporal segmentation framework that integrates Slot-based State Space Models (SlotSSMs) with prompt-guided refinement in SAM 2. SlotSSMs explicitly maintain object-level temporal states, which are converted into adaptive prompts to guide SAM 2 toward spatially precise and temporally coherent segmentation. A sparse Slot Mixer enhances inter-object interaction, while confidence-aware mask fusion balances global stability and local boundary refinement. Extensive experiments on three public underwater benchmarks demonstrate that AquaSlot-SAM consistently outperforms strong contemporary baselines, achieving substantial improvements across multiple metrics and validating the effectiveness of structured memory-guided prompting for dynamic underwater segmentation.

I. INTRODUCTION

Robust perception in complex underwater environments is essential for marine engineering, ecological monitoring, and robotic operations. Accurate segmentation of seabed resources and marine organisms improves the efficiency and safety of underwater inspections, supporting applications such as resource localization [1], ecological monitoring [2], and equipment management [3]. Beyond ecological research, reliable scene understanding enables autonomous underwater vehicles (AUVs) to navigate, inspect, and interact safely in challenging and visually degraded conditions.

Early segmentation methods evolved from thresholding and graph-based models to deep learning architectures such as FCNs and U-Net, achieving pixel- and instance-level segmentation. However, underwater perception introduces unique difficulties, including light absorption, scattering, turbidity-induced blur, and irregular object motion [4], [5], [6], [7], as illustrated in Fig. 1. These degradations reduce visual contrast, distort color distribution, and cause temporal instability, thereby compromising the perception reliability of robotic systems operating in real-world underwater environments.

¹Wenbin Feng, Yu Lu, Yuetong Lin, Meng Li, Qiang Li are with the School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China.

²Yu Lu is also with the Department of Mechanical and Mechatronics Engineering, University of Auckland, Auckland, New Zealand.

*Corresponding author. Email: lvyu@sztu.edu.cn

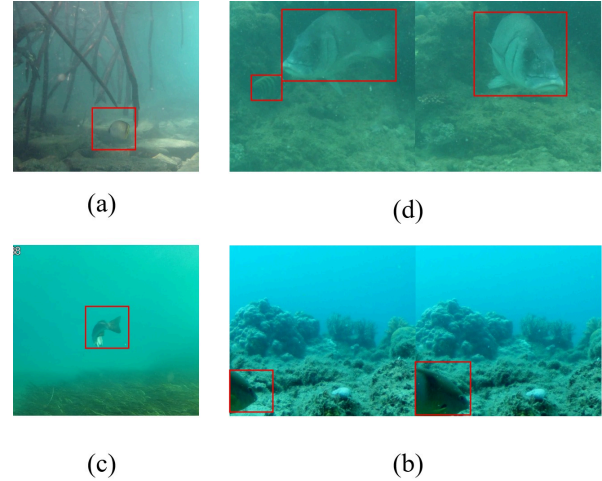


Fig. 1. Challenges in complex underwater environments: (a) heterogeneous backgrounds and illumination; (b–d) occlusion and reappearance of marine organisms; (c) irregular motion and turbidity effects. These factors lead to low contrast, color distortion, and reduced perception accuracy, posing difficulties for robotic sensing and segmentation.

Recent advances in video object segmentation (VOS) leverage memory-based propagation and feature association mechanisms to improve temporal consistency. Representative approaches such as STCN and XMem maintain appearance memories across frames to enhance mask continuity. While effective in relatively clean visual domains, these methods primarily rely on feature similarity and lack explicit object-centric state modeling, which limits robustness under severe occlusion, low contrast, and underwater degradations.

Foundation models such as the Segment Anything Model (SAM) and its variants demonstrate strong generalization ability in natural images. SAM 2 further improves segmentation quality in complex scenes through improved mask decoding and prompt interaction. However, in highly dynamic underwater environments, segmentation performance remains sensitive to temporal inconsistency, weak edges, and degraded visual cues. Recent adaptations for marine imagery, including WaterSAM [8], Dual-SAM [9], and MAS-SAM [10], extend SAM to underwater datasets but often rely on dataset-specific fine-tuning and remain largely oriented toward static image segmentation.

Real-world underwater environments therefore require a framework that jointly models object identity, temporal dynamics, and prompt-guided refinement. Frame-wise segmentation lacks temporal memory, while memory-based propagation alone may drift under occlusion or severe appearance

changes. An object-centric temporal modeling mechanism that can continuously update scene representations and guide segmentation prompts is therefore desirable.

To address these limitations, we propose **AquaSlot-SAM**, a hybrid object-centric temporal segmentation framework that integrates SAM 2 with enhanced Slot State Space Models (SlotSSMs). Unlike straightforward prompt-based refinement, AquaSlot-SAM introduces a structured temporal state modeling layer that decomposes scenes into object-centric slots and maintains dynamic slot states over time. These slot representations are converted into adaptive SAM-compatible prompts (points, boxes, or masks), enabling precise mask refinement while preserving temporal coherence. A slot memory mechanism supports trajectory recovery under occlusion, and sparse inter-slot interactions improve robustness to environmental variations. This integration enables stable, coherent, and accurate segmentation in complex underwater scenarios.

The contributions of this work are summarized as follows:

- We propose **AquaSlot-SAM**, a hybrid object-centric temporal segmentation framework that integrates Slot State Space Models with SAM 2 through structured prompt interaction for dynamic underwater environments.
- We introduce an adaptive prompt generation mechanism that converts slot-level representations into interactive prompts (points and boxes), enabling precise and temporally consistent segmentation.
- We design a sparse Slot Mixer to model long-term object trajectories and support re-identification under occlusion and turbidity.
- We validate our approach on three public underwater benchmarks, demonstrating consistent improvements over classical, segmentation-based, and contemporary VOS methods in both spatial accuracy and temporal robustness.

II. RELATED WORK

A. SAM-Based Segmentation

The Segment Anything Model (SAM) has demonstrated strong zero-shot generalization across natural and domain-specific datasets. Its extension to video segmentation introduces challenges related to temporal consistency and cross-frame object association. Recent video object segmentation (VOS) methods such as STCN [11] and XMem [12] employ memory-based feature propagation to maintain mask continuity across frames, while AOT [13] explores object association through transformer-based temporal aggregation. SAM 2 further enhances prompt-guided segmentation for dynamic scenes via improved mask decoding and temporal refinement.

Several works adapt SAM to specialized domains. VideoSAM [14] integrates temporal propagation mechanisms, while medical adaptations such as H-SAM [15] and MemSAM [16] introduce LoRA tuning and memory-enhanced prompts for domain-specific robustness. In ma-

rine scenarios, WaterSAM [8], Dual-SAM [9], and MAS-SAM [10] extend SAM to underwater imagery but primarily focus on static or frame-wise segmentation. Compared with memory-based propagation approaches, these methods lack explicit object-centric temporal state modeling for dynamic underwater environments.

B. Underwater Segmentation

Underwater segmentation faces challenges including light scattering, motion blur, turbidity, and complex backgrounds. Existing approaches can be broadly categorized into motion-based, appearance-driven, and enhancement-assisted methods. Motion-aware models such as MSGNet [17] and GraphMOS-U [18] integrate optical flow and saliency cues for dynamic object extraction, but performance remains sensitive to flow estimation quality under heavy degradation. Appearance-driven segmentation networks improve robustness through architectural design and feature refinement, including adaptive convolutional modules and contrast-aware learning. Enhancement-based pipelines attempt to restore visibility prior to segmentation, yet preprocessing often fails to recover fine structures and may introduce artifacts.

While these approaches improve underwater perception, most rely on either frame-wise processing or short-term feature propagation, limiting robustness under long occlusions and severe environmental disturbances.

C. SSM-Based Segmentation

State space models (SSMs) provide a principled framework for modeling temporal dynamics through latent state evolution [19]. Classical filtering methods such as Kalman filters [20] and HMMs [21] enable sequential reasoning but struggle with high-dimensional visual inputs. Recent neural state space models introduce structured linear recurrence to efficiently capture long-range dependencies, enabling scalable spatiotemporal modeling in vision tasks. Object-centric extensions further decompose scenes into slot-based representations, supporting multi-object reasoning and persistent identity tracking across frames.

In contrast to memory-based feature propagation alone, object-centric state modeling explicitly maintains dynamic object representations over time, providing stronger resilience to occlusion, motion irregularity, and environmental noise. This perspective motivates the integration of SlotSSMs [22] with prompt-guided segmentation in dynamic underwater environments.

III. METHOD

We propose **AquaSlot-SAM**, a hybrid object-centric temporal segmentation framework for dynamic underwater perception. Rather than performing frame-wise refinement or memory-based feature propagation alone, AquaSlot-SAM introduces structured temporal state modeling to explicitly maintain object identities and guide prompt-based segmentation within SAM 2.

The framework integrates Slot State Space Models (SlotSSMs) with SAM 2 through structured prompt interaction and confidence-aware fusion. As illustrated in Fig. 2,

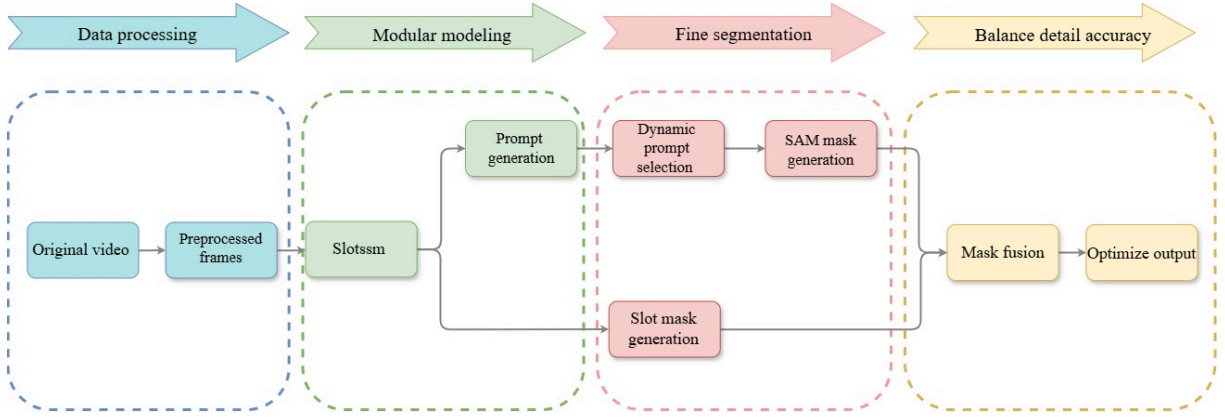


Fig. 2. Pipeline of AquaSlot-SAM: (1) preprocessing; (2) SlotSSM-based decomposition and prompt generation; (3) SAM 2 refinement; (4) mask fusion balancing global stability and local precision.

the pipeline consists of four stages: (1) feature extraction and preprocessing, (2) slot-based temporal state modeling, (3) slot-guided SAM 2 refinement, and (4) confidence-aware mask fusion.

A. Slot-Based Temporal State Modeling

As shown in Fig. 3, SlotSSMs maintain object-level temporal states and interact with SAM 2 through structured prompts and mask fusion.

To explicitly model dynamic marine objects, we decompose frame features into K object-centric slots and evolve each slot through a structured state space model.

Slot Initialization. Let $F_t \in \mathbb{R}^{N \times D}$ denote frame features extracted at time t . We initialize K learnable slot embeddings $S_0 \in \mathbb{R}^{K \times D}$. Slot assignment is performed via attention:

$$\begin{aligned} Q &= W_Q S_{t-1}, \\ K_t &= W_K F_t, \\ V_t &= W_V F_t, \\ \text{Slot}_t^k &= \text{Softmax}\left(\frac{Q K_t^\top}{\sqrt{D}}\right) V_t. \end{aligned} \quad (1)$$

This attention mechanism distributes spatial features to object-centric slots, enabling spatial disentanglement and multi-object reasoning.

Structured State Evolution. Each slot evolves independently via a linear state space model:

$$h_t^k = A_k h_{t-1}^k + B_k \text{Slot}_t^k, \quad y_t^k = C_k h_t^k, \quad (2)$$

where (A_k, B_k, C_k) are learnable transition matrices parameterized to ensure stable long-term dynamics. This formulation explicitly maintains object-level temporal states, providing resilience to occlusion and motion irregularity.

Sparse Slot Interaction. To capture inter-object relationships, we apply sparse self-attention across slots:

$$\text{Slot}_t^{1:K} = \text{SelfAttn}(\text{LayerNorm}(y_t^{1:K})). \quad (3)$$

This sparse interaction preserves modularity while enabling contextual exchange across objects.

Mask Decoding. A broadcast decoder reconstructs slot-level masks M_t^k . Pixel-wise normalization is performed via Softmax with temperature α :

$$M_t(i, j) = \sum_{k=1}^K \frac{\exp(\alpha M_t^k(i, j))}{\sum_{l=1}^K \exp(\alpha M_t^l(i, j))}. \quad (4)$$

This formulation ensures mutually exclusive and globally consistent object masks.

B. Slot-Guided Prompting for SAM 2

While SAM 2 provides strong prompt-guided segmentation, it lacks explicit temporal identity preservation. We therefore convert slot representations into structured prompts to guide refinement.

For each slot k , geometric centers and extrema of M_t^k are extracted to generate point and box prompts. The SAM 2 refinement is defined as:

$$M_{\text{sam}}^k, c_{\text{sam}}^k = \text{SAM2}(I_t; \text{box} = b_k, \text{points} = (x_c^k, y_c^k)), \quad (5)$$

where b_k denotes the bounding box derived from slot extrema.

Adaptive Prompt Selection. Prompt types are dynamically selected using slot confidence and temporal stability metrics. Slot confidence is defined as the mean mask probability, while temporal stability is measured by inter-frame IoU between consecutive slot masks. This adaptive mechanism enables refinement strategies that are robust to occlusion, blur, and low-contrast boundaries.

C. Confidence-Aware Mask Fusion

SlotSSM masks provide temporal consistency, while SAM 2 masks offer refined boundaries. We combine them using normalized confidence weights:

$$M_{\text{fused}}^k = \frac{c_{\text{slot}}^k}{c_{\text{slot}}^k + c_{\text{sam}}^k} M_{\text{slot}}^k + \frac{c_{\text{sam}}^k}{c_{\text{slot}}^k + c_{\text{sam}}^k} M_{\text{sam}}^k, \quad (6)$$

where c_{slot}^k and c_{sam}^k are normalized confidence scores. This convex fusion balances global temporal stability and local boundary precision.

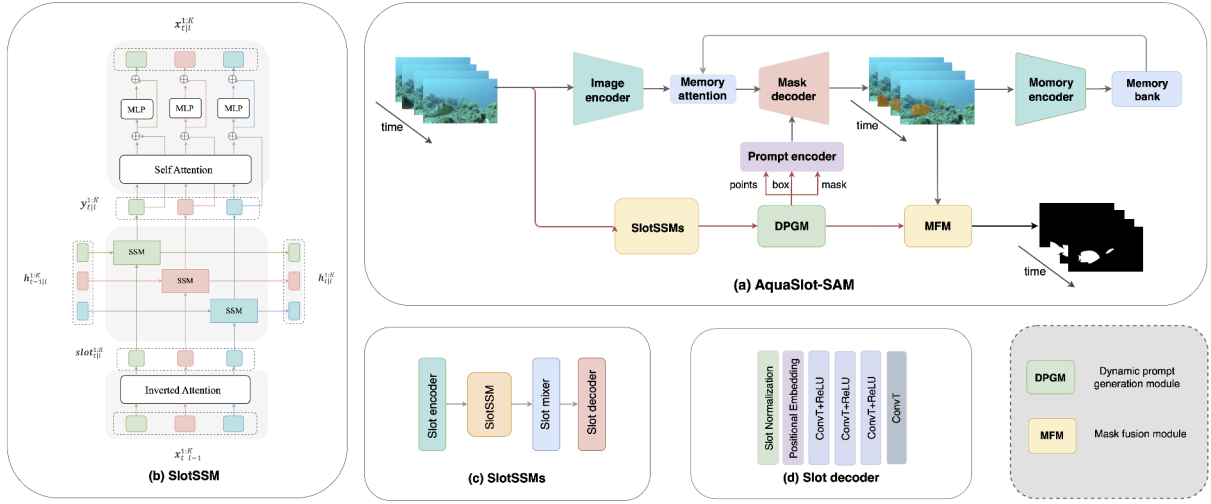


Fig. 3. Overview of AquaSlot-SAM. (a) Integration of Slot-based State Space Models (SlotSSMs) and SAM 2 through dynamic prompt generation (DPGM) and mask fusion (MFM). (b) Internal SlotSSM structure with slot encoding, state evolution, and sparse slot interaction. (c) Temporal state propagation across frames. (d) Slot decoder for object-level mask reconstruction.

D. Loss Function

The training objective enforces spatial accuracy, boundary sharpness, and slot disentanglement:

$$L_{IoU} = 1 - \frac{1}{K} \sum_{k=1}^K \frac{|M_{fused}^k \cap M_{gt}^k|}{|M_{fused}^k \cup M_{gt}^k|}, \quad (7)$$

$$L_{edge} = \frac{1}{K} \sum_{k=1}^K \|\nabla M_{fused}^k - \nabla M_{gt}^k\|_1, \quad (8)$$

$$L_{consist} = \frac{1}{K^2} \sum_{k \neq l} (\text{Cov}_t(h^k, h^l))^2, \quad (9)$$

$$L = L_{IoU} + \lambda_1 L_{edge} + \lambda_2 L_{consist}. \quad (10)$$

Here $\text{Cov}_t(\cdot, \cdot)$ denotes channel-wise covariance computed over temporal steps. The consistency term encourages slot independence and prevents representation collapse.

IV. EXPERIMENT

A. Datasets

We evaluate AquaSlot-SAM on three public underwater benchmarks: DeepFish [23], Seagrass [24], and Fish4Knowledge [25]. These datasets cover diverse underwater environments characterized by low contrast, motion blur, occlusion, and complex backgrounds (Fig. 4).

DeepFish contains 39,766 high-resolution frames (1920×1080) spanning 15 fish species across 20 tropical habitats, with pixel-level annotations for segmentation.

Seagrass includes 4,281 annotated images captured by 18 underwater cameras at varying depths and viewpoints, reflecting realistic ecological interactions between fish and vegetation.

Fish4Knowledge is a large-scale reef video dataset; following prior segmentation works, we use the annotated subsets with frame-level ground truth for quantitative evaluation.

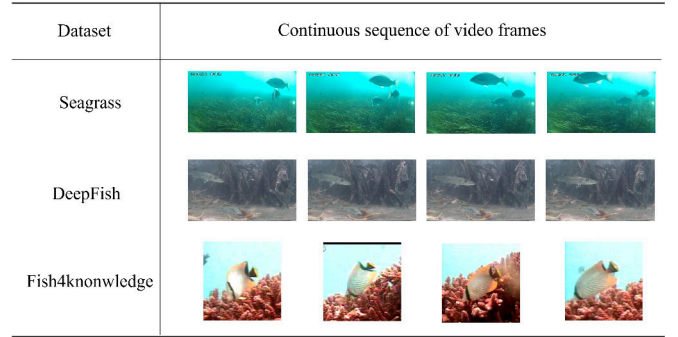


Fig. 4. Sample frames from three underwater datasets: Seagrass (top), DeepFish (middle), and Fish4Knowledge (bottom). They illustrate the environmental complexity and appearance variations faced in underwater segmentation.

B. Implementation Details

Experiments were conducted on a Xeon 8255C CPU and an NVIDIA Tesla T4 GPU using PyTorch. Input frames were resized to 512×512. We follow a 7:1:2 train/validation/test split unless official splits are provided.

SAM 2 was initialized from publicly available pretrained weights and fine-tuned jointly with SlotSSM components. The model was optimized using AdamW (learning rate $1e^{-5}$, weight decay $4e^{-5}$) with batch size 8 for 30 epochs. Data augmentation including random horizontal flips and color jitter was applied during training. Automatic mixed precision (AMP) was used to reduce memory consumption and accelerate training.

C. Evaluation Metrics

We adopt standard segmentation metrics following prior work.

J-Mean (JM) measures the region similarity (IoU) between prediction $\hat{M}$ and ground truth M :

$$\text{JM} = \frac{|\hat{M} \cap M|}{|\hat{M} \cup M|}. \quad (11)$$

F-Measure (FM) evaluates contour accuracy as the harmonic mean of boundary precision (P_c) and recall (R_c):

$$\text{FM} = \frac{2P_c R_c}{P_c + R_c}. \quad (12)$$

J&F denotes the average of JM and FM. **mPA** and **mIoU** are computed over $k + 1$ classes following standard definitions.

D. Task Definition and Evaluation Protocol

Although the three datasets originate from different underwater benchmarks, we unify the task as video object segmentation (VOS) with frame-level mask prediction. For DeepFish and Seagrass, we follow the official semantic segmentation annotations and treat each frame independently for region-based evaluation (mIoU, mPA). For Fish4K, which contains continuous video sequences, we adopt the standard video object segmentation protocol and report J&F metrics. To ensure consistency, all ablation studies on DeepFish are evaluated using the same protocol as the main comparison (mIoU and mPA), while J&F and mAP(0.5) are additionally reported only for sequence-level temporal analysis.

E. Performance Comparison

We evaluate AquaSlot-SAM against classical background subtraction methods, modern segmentation backbones, and contemporary video object segmentation (VOS) models on the **Seagrass**, **DeepFish**, and **Fish4K** datasets. Results are summarized in Tables I–III and Figures 5–7. In all tables, the best results are shown in **bold**, and the second-best results are underlined.

TABLE I
SEGMENTATION RESULTS ON THE **SEAGRASS** DATASET (HIGHER IS BETTER).

Method	J&F $\uparrow$	J-M $\uparrow$	J-R $\uparrow$	F-M $\uparrow$	F-R $\uparrow$
<i>Self-supervised methods</i>					
CoaT-Transformer [26]	50.0	41.5	43.3	58.1	65.4
DenseFlow [27]	45.5	40.2	41.0	50.7	54.7
CRW [28]	43.2	38.9	40.4	46.2	50.8
MAST [29]	40.3	37.1	38.7	43.8	48.5
<i>Fully-supervised / VOS methods</i>					
FCN-8s [30]	64.7	52.4	55.7	71.4	<u>79.1</u>
AOT (2021)	67.3	61.2	69.5	73.4	74.2
STCN (2021)	69.8	63.7	71.6	75.9	76.8
XMem (2022)	71.4	65.8	73.2	77.6	78.1
SAM2 (fine-tuned)	<u>72.0</u>	<u>66.1</u>	<u>75.0</u>	<u>77.8</u>	78.6
AquaSlot-SAM	74.1	68.8	79.7	79.3	79.7

Seagrass. As shown in Table I, AquaSlot-SAM achieves the best performance across all five metrics, reaching a J&F score of **74.1%**. It consistently outperforms strong contemporary VOS baselines such as STCN (69.8%), XMem (71.4%), and fine-tuned SAM2 (72.0%), with notable gains

in both region similarity (J-M) and contour accuracy (F-M). These results indicate that integrating object-centric temporal modeling with adaptive SAM refinement improves segmentation robustness and temporal consistency in complex underwater scenes.

TABLE II
SEGMENTATION RESULTS ON THE **DEEPFISH** DATASET (HIGHER IS BETTER).

Method	mPA (%) $\uparrow$	mIoU (%) $\uparrow$
SUIM-Net [31]	89.89	83.27
WaterSNet [32]	91.14	86.83
FSNet [33]	86.77	82.84
AMC-Net [34]	91.55	87.61
MSGNet [17]	91.89	88.91
DeepLabV3+[35]	93.02	90.47
SegFormer-B2[36]	94.15	91.83
Mask2Former[37]	95.21	92.74
SAM2 (fine-tuned)	<u>95.88</u>	<u>93.61</u>
AquaSlot-SAM	96.88	94.93

DeepFish. In Table II, AquaSlot-SAM achieves **96.88%** mPA and **94.93%** mIoU, outperforming strong contemporary segmentation models such as Mask2Former (92.74%) and fine-tuned SAM2 (93.61%). Compared with these high-capacity backbones, our method provides consistent improvements in object-background separation and boundary precision. Fig. 6 further demonstrates more stable contours and finer structural details in challenging underwater conditions.

TABLE III
SEGMENTATION RESULTS ON THE **FISH4K** DATASET (F-MEASURE, HIGHER IS BETTER).

Method	F-Measure (%) $\uparrow$				
	Dynamic	Hybrid	Light changes	Standard	Overall
GMG [38]	16.26	21.60	33.04	19.67	22.39
MC-BGS [39]	22.46	81.05	60.22	49.89	50.90
ML-BGS [40]	43.16	70.37	66.74	49.39	57.41
SubSENSE [41]	70.81	70.90	67.73	57.32	66.69
GraphMOS-U [18]	84.03	86.02	81.58	83.27	83.72
AOT [13]	85.11	88.54	82.06	84.71	85.11
STCN [11]	86.47	90.03	83.72	85.91	86.53
XMem [12]	87.02	92.18	<u>84.06</u>	86.32	87.12
SAM2 (fine-tuned)	87.34	92.71	83.95	<u>86.84</u>	87.49
AquaSlot-SAM	88.23	94.12	84.18	87.43	87.88

Fish4K. Table III shows that AquaSlot-SAM consistently achieves the best performance, reaching **88.23%** in dynamic scenes and **94.12%** in hybrid scenes. It outperforms strong contemporary VOS methods such as STCN (86.53%), XMem (87.12%), and fine-tuned SAM2 (87.49%) in overall F-measure. As illustrated in Fig. 7, AquaSlot-SAM maintains more coherent masks under strong background motion and occlusion, demonstrating improved temporal stability over both classical subtraction approaches and recent learning-based models.

Overall, these results demonstrate that AquaSlot-SAM consistently achieves superior segmentation accuracy and improved temporal stability compared with strong contemporary baselines across diverse underwater environments.

Qualitative comparisons (Figures 5–7) further highlight the advantages of AquaSlot-SAM, showing that it produces more temporally consistent and spatially precise masks than competing methods, particularly under low-contrast, occluded, and dynamic underwater conditions.

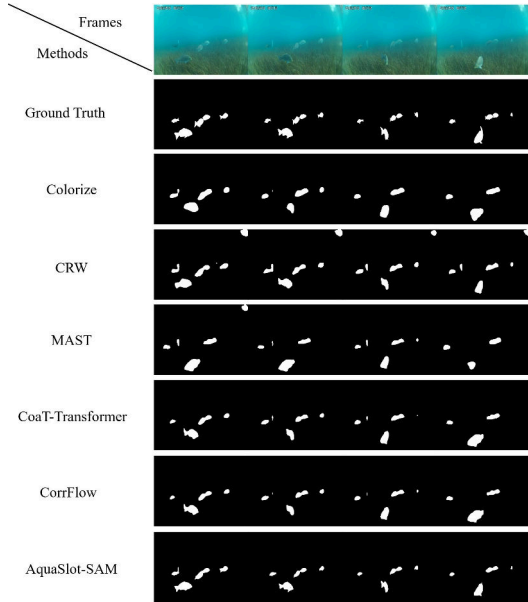


Fig. 5. Qualitative comparison on Seagrass low-contrast scenes. From top to bottom: input frames, representative self-supervised methods (Colorize, CRW, MAST, CoaT-Transformer, CorrFlow), and AquaSlot-SAM.

On the **Seagrass** dataset (Fig. 5), representative self-supervised methods struggle to preserve object integrity under low contrast and background clutter, often producing fragmented or noisy masks. In contrast, AquaSlot-SAM maintains more coherent object boundaries and consistent mask propagation across frames. These qualitative observations align with the quantitative improvements over strong contemporary VOS baselines reported in Table I.

For the **DeepFish** dataset (Fig. 6), representative underwater segmentation models such as MSGNet and WaterSNet produce generally accurate masks but exhibit boundary roughness and temporal instability under cluttered backgrounds, while AMC-Net and FSNet struggle with noise suppression and object continuity. In contrast, AquaSlot-SAM captures fine structures such as fish tails and thin contours while maintaining consistent mask propagation across frames. These qualitative improvements are consistent with the quantitative gains over strong contemporary segmentation backbones reported in Table II.

On the **Fish4K** dataset (Fig. 7), classical background subtraction methods (GMG, MC-BGS, ML-BGS, SubSENSE) exhibit artifacts, leakage, and instability under strong background motion. GraphMOS-U improves object completeness but remains sensitive to occlusion and blur. In contrast, AquaSlot-SAM maintains more coherent mask propagation and clearer object boundaries in dynamic scenes. These qualitative observations are consistent with the quantitative improvements over strong contemporary VOS baselines re-

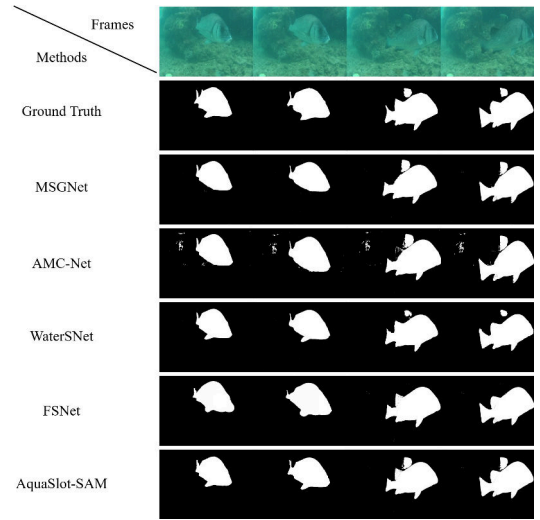


Fig. 6. Qualitative comparison on DeepFish. From top to bottom: input frames, Ground Truth, MSGNet, AMC-Net, WaterSNet, FSNet, and AquaSlot-SAM.

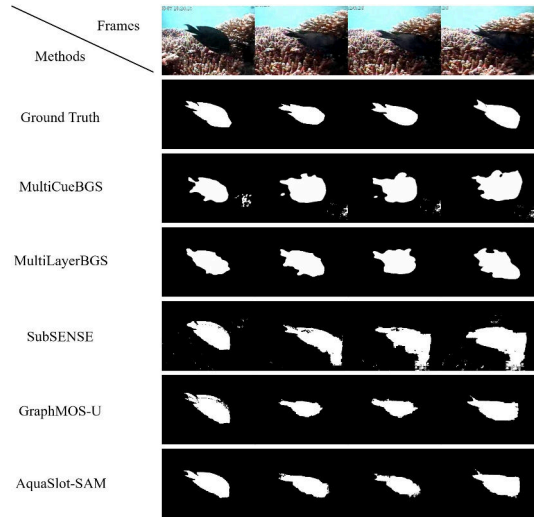


Fig. 7. Qualitative comparison on Fish4K dynamic scenes. From top to bottom: input frames, GMG, MC-BGS, ML-BGS, SubSENSE, GraphMOS-U, and AquaSlot-SAM.

ported in Table III.

F. Computational Efficiency Analysis

Although AquaSlot-SAM introduces additional temporal modeling and prompt refinement modules, we conduct a detailed computational analysis to quantify its runtime and memory overhead compared to representative baselines.

Metrics. We evaluate: (i) number of parameters (Params), (ii) FLOPs per forward pass, (iii) average inference time (ms), (iv) frames per second (FPS), (v) peak GPU memory consumption.

As shown in Table IV, AquaSlot-SAM introduces a moderate increase in parameters (+16M) and FLOPs (+16G) compared to SAM 2 due to the SlotSSM temporal modeling and mask fusion module. At 512×512 resolution, inference time increases from 72ms to 83ms (approximately +15%),

TABLE IV

COMPUTATIONAL EFFICIENCY COMPARISON UNDER IDENTICAL HARDWARE (TESLA T4, BATCH SIZE 1). LOWER VALUES ARE BETTER FOR PARAMS, FLOPS, AND LATENCY, WHILE HIGHER VALUES ARE BETTER FOR FPS.

Method	Params (M)↓	FLOPs (G)↓	512×512		1024×1024	
			Latency (ms)↓	FPS↑	Latency (ms)↓	FPS↑
FCN-8s [30]	134	58	24	41.7	78	12.8
SAM2 (baseline)	632	185	72	13.9	214	4.7
GraphMOS-U [18]	58	41	31	32.3	96	10.4
AquaSlot-SAM	648	201	83	12.0	247	4.0

corresponding to a runtime of 12.0 FPS. At 1024×1024 , the runtime remains practical at 4.0 FPS.

Importantly, the performance gains reported in Tables I–III come at a relatively small computational overhead relative to the SAM 2 backbone. Compared with classical fully convolutional models (e.g., FCN-8s), AquaSlot-SAM trades additional computation for significantly improved temporal coherence and segmentation accuracy.

Overall, the results indicate that AquaSlot-SAM achieves a favorable accuracy–efficiency trade-off, maintaining real-time feasibility at moderate resolutions while substantially improving robustness in complex underwater environments.

G. Ablation Study

To ensure evaluation consistency, all ablation experiments on the **DeepFish** dataset follow the same protocol as the main comparison in Table II, where segmentation performance is primarily evaluated using **mPA** and **mIoU**. To additionally analyze temporal coherence, we report J&F as an auxiliary sequence-level metric. No detection-based metrics are used in this section.

Table V summarizes the contribution of each module in AquaSlot-SAM.

TABLE V
ABLATION STUDY OF AQUASLOT-SAM COMPONENTS ON THE **DEEPFISH** DATASET (HIGHER IS BETTER).

Modules	mPA (%) ↑	mIoU (%) ↑	J&F (%) ↑
SAM 2 (baseline)	95.88	93.61	81.24
w/o PromptSelect	95.34	92.72	83.06
w/o MFM	96.02	93.48	85.19
AquaSlot-SAM	96.88	94.93	87.71

Effect of SlotSSM (Temporal Modeling). Removing the SlotSSM module (i.e., using only SAM 2) leads to a decrease of 1.32% in mIoU and 6.47% in J&F. While per-frame segmentation accuracy remains relatively strong, temporal consistency degrades significantly, demonstrating the necessity of object-centric temporal modeling in underwater video sequences.

Effect of Dynamic Prompt Selection. Replacing adaptive slot-guided prompts with fixed bounding boxes results in a 2.21% drop in mIoU and a 4.65% decrease in J&F. This confirms that confidence-aware prompt generation improves robustness under occlusion and background interference.

Effect of Mask Fusion Module (MFM). Removing the confidence-weighted mask fusion module reduces mIoU by 1.45% and J&F by 2.52%, indicating that fusion effectively balances temporal stability with fine-grained boundary refinement.

Full Model. The complete AquaSlot-SAM model achieves the best performance across all metrics (mIoU: 94.93%, J&F: 87.71%), demonstrating the complementary benefits of temporal slot modeling, adaptive prompting, and mask fusion.

Overall, the ablation results verify that each component contributes consistently to both spatial segmentation accuracy and temporal coherence, without altering the evaluation protocol used in the main experiments.

V. DISCUSSION

A. Overall Insights

Experiments on three diverse underwater benchmarks indicate that AquaSlot-SAM achieves robust segmentation under low contrast, dynamic backgrounds, and prolonged occlusions. Beyond empirical gains, a key insight is that structured object-centric state modeling complements prompt-based foundation segmentation. SlotSSM explicitly maintains temporal object identities through state evolution, while SAM 2 contributes high-resolution boundary refinement. Rather than treating video segmentation as independent frame-wise prediction or feature propagation, AquaSlot-SAM formulates it as state-guided prompt refinement. This paradigm bridges structured temporal modeling and foundation model adaptation, yielding consistent spatial accuracy and temporal stability across challenging underwater conditions.

B. Limitations

Despite its effectiveness, several limitations remain. **(1) Computational Cost:** The dual-path architecture increases inference latency, which may limit deployment on resource-constrained underwater robots. **(2) Heuristic Prompt Selection:** Although adaptive, the current prompt strategy relies on threshold-based confidence and stability metrics, which may require dataset-specific tuning. **(3) Extreme Visual Degradation:** In scenarios with severe turbidity, glare, or strong motion blur, segmentation performance deteriorates due to insufficient visual cues, suggesting that purely RGB-based modeling has inherent limitations.

C. Future Directions

Future work will focus on improving efficiency, adaptability, and cross-domain robustness. Lightweight SlotSSM variants and parameter-efficient prompt tuning could reduce computational overhead. Learning-based prompt controllers may replace heuristic thresholding, enabling fully adaptive refinement. In addition, incorporating multimodal sensing (e.g., sonar or depth) and domain-adaptive training strategies may enhance robustness under extreme distortions. Finally, extending AquaSlot-SAM to unified multi-object tracking and multi-class segmentation represents a natural progression, as the slot-based state formulation inherently supports trajectory-aware reasoning.

VI. CONCLUSION

We presented AquaSlot-SAM, a hybrid object-centric temporal segmentation framework that integrates Slot-based state modeling with prompt-guided refinement in SAM 2 for robust underwater perception. By explicitly maintaining object-level temporal states and leveraging adaptive prompt interaction, the proposed approach achieves improved spatial precision and temporal coherence under challenging conditions such as low contrast, dynamic backgrounds, and occlusions. Extensive evaluations across multiple underwater benchmarks validate its effectiveness over strong contemporary baselines. This work highlights the potential of structured memory-guided prompting as a principled paradigm for dynamic segmentation, with future extensions toward efficiency optimization and broader real-world deployment.

REFERENCES

- [1] T. Zheng, J. Wu, H. Kong, H. Zhao, B. Qu, L. Liu, H. Yu, and C. Zhou, "A video object segmentation-based fish individual recognition method for underwater complex environments," *Ecological Informatics*, vol. 82, p. 102689, 2024.
- [2] J.-I. Watanabe, Y. Shao, and N. Miura, "Underwater and Airborne Monitoring of Marine Ecosystems and Debris," *Journal of Applied Remote Sensing*, vol. 13, no. 4, p. 044509, 2019.
- [3] J. Zhang, F. Han, D. Han, J. Yang, W. Zhao, and H. Li, "Advanced Underwater Measurement System for ROVs: Integrating Sonar and Stereo Vision for Enhanced Subsea Infrastructure Maintenance," *Journal of Marine Science and Engineering*, vol. 12, no. 2, p. 306, 2024.
- [4] D. Zhang, C. Wu, J. Zhou, W. Zhang, Z. Lin, K. Polat, and F. Alenezi, "Robust underwater image enhancement with cascaded multi-level sub-networks and triple attention mechanism," *Neural Networks*, vol. 169, pp. 685–697, 2024.
- [5] L. Chang, Y. Wang, B. Du, and C. Xu, "Rectangling and enhancing underwater stitched image via content-aware warping and perception balancing," *Neural Networks*, vol. 181, p. 106809, 2025.
- [6] D. K. Rout, M. Kapoor, B. N. Subudhi, V. Thangaraj, V. Jakhethiya, and A. Bansal, "Underwater visual surveillance: A comprehensive survey," *Ocean Engineering*, vol. 309, p. 118367, 2024.
- [7] L. Chen, F. Zhou, S. Wang, J. Dong, N. Li, H. Ma, X. Wang, and H. Zhou, "SWIPENET: Object detection in noisy underwater scenes," *Pattern Recognition*, vol. 132, p. 108926, 2022.
- [8] Y. Hong, X. Zhou, R. Hua, Q. Lv, and J. Dong, "WaterSAM: Adapting SAM for Underwater Object Segmentation," *Journal of Marine Science and Engineering*, vol. 12, no. 9, p. 1616, 2024.
- [9] P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic Animals and Where to Find Them: Segment Any Marine Animal with Dual SAM," in *CVPR*, 2024, pp. 2578–2587.
- [10] T. Yan, Z. Wan, X. Deng, P. Zhang, Y. Liu, and H. Lu, "MAS-SAM: segment any marine animal with aggregated features," in *IJCAI*, 2024, pp. 6886–6894.
- [11] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Space-Time Memory Networks for Video Object Segmentation With User Guidance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 442–455, 2020.
- [12] H. K. Cheng and A. G. Schwing, "XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model," in *ECCV*. Springer, 2022, pp. 640–658.
- [13] Z. Yang, Y. Wei, and Y. Yang, "Associating Objects with Transformers for Video Object Segmentation," *NeurIPS*, vol. 34, pp. 2491–2502, 2021.
- [14] P. Guo, Z. Zhao, J. Gao, C. Wu, T. He, Z. Zhang, T. Xiao, and W. Zhang, "VideoSAM: Open-World Video Segmentation," *arXiv preprint arXiv:2410.08781*, 2024.
- [15] Z. Cheng, Q. Wei, H. Zhu, Y. Wang, L. Qu, W. Shao, and Y. Zhou, "Unleashing the potential of SAM for medical adaptation via hierarchical decoding," in *CVPR*, 2024, pp. 3511–3522.
- [16] X. Deng, H. Wu, R. Zeng, and J. Qin, "MemSAM: taming segment anything model for echocardiography video segmentation," in *CVPR*, 2024, pp. 9622–9631.
- [17] P. Zhang, H. Yu, H. Li, X. Zhang, S. Wei, W. Tu, Z. Yang, J. Wu, and Y. Lin, "MSGNet: multi-source guidance network for fish segmentation in underwater videos," *Frontiers in Marine Science*, vol. 10, p. 1256594, 2023.
- [18] M. Kapoor, W. Prummel, J. H. Giraldo, B. N. Subudhi, A. Zakharova, T. Bouwmans, and A. Bansal, "Graph-based Moving Object Segmentation for underwater videos using semi-supervised learning," *Computer Vision and Image Understanding*, p. 104290, 2025.
- [19] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari, "S4: Distributed Stream Computing Platform," in *ICDM Workshops*. IEEE, 2010, pp. 170–177.
- [20] G. Bishop and G. Welch, "An Introduction to the Kalman Filter," in *Proceedings of SIGGRAPH, Course 8*, 2001, p. 41.
- [21] S. R. Eddy, "Accelerated Profile HMM Searches," *PLoS Computational Biology*, vol. 7, no. 10, p. e1002195, 2011.
- [22] J. Jiang, F. Deng, G. Singh, M. Lee, and S. Ahn, "Slot State Space Models," in *NeurIPS*, 2010, pp. 1–32.
- [23] A. Saleh, I. H. Laradji, D. A. Konovalov, O. Strauss, M. A. Westenberg, and J. D. Lawrey, "A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis," *Scientific Reports*, vol. 10, p. 14671, 2020.
- [24] E. M. Ditría, R. M. Connolly, E. L. Jinks, and S. Lopez-Marciano, "Annotated Video Footage for Automated Identification and Counting of Fish in Unconstrained Seagrass Habitats," *Frontiers in Marine Science*, vol. 8, 2021.
- [25] R. B. Fisher, Y. H. Chen-Burger, D. Giordano, L. Hardman, and F. P. Lin, *Fish4Knowledge: Collecting and Analyzing Massive Coral Reef Fish Video Data*. Berlin/Heidelberg, Germany: Springer, 2016, vol. 104.
- [26] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-Scale Conv-Attentional Image Transformers," in *ICCV*, 2021, pp. 9981–9990.
- [27] N. Araslanov, S. Schaub-Meyer, and S. Roth, "Dense Unsupervised Learning for Video Segmentation," *NeurIPS*, vol. 34, pp. 25 308–25 319, 2021.
- [28] A. Jabri, A. Owens, and A. Efros, "Space-Time Correspondence as a Contrastive Random Walk," *NeurIPS*, vol. 33, pp. 19 545–19 560, 2020.
- [29] Z. Lai, E. Lu, and W. Xie, "MAST: A Memory-Augmented Self-Supervised Tracker," in *CVPR*, 2020, pp. 6479–6488.
- [30] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2016.
- [31] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," in *IROS*, 2020, pp. 1769–1776.
- [32] R. Chen, Z. Fu, Y. Huang, E. Cheng, and X. Ding, "A Robust Object Segmentation Network for Underwater Scenes," in *ICASSP*, 2022, pp. 2629–2633.
- [33] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao, "Full-Duplex Strategy for Video Object Segmentation," in *ICCV*, 2021, pp. 4902–4913.
- [34] S. Yang, L. Zhang, J. Qi, H. Lu, S. Wang, and X. Zhang, "Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation," in *ICCV*, 2021, pp. 1544–1553.
- [35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *ECCV*, 2018, pp. 801–818.
- [36] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *NeurIPS*, vol. 34, pp. 12 077–12 090, 2021.
- [37] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-Attention Mask Transformer for Universal Image Segmentation," in *CVPR*, 2022, pp. 1290–1299.
- [38] A. B. Godbehare, A. Matsukawa, and K. Goldberg, "Visual Tracking of Human Visitors under Variable-Lighting Conditions for a Responsive Audio Art Installation," in *ACC*, 2012, pp. 4305–4312.
- [39] S. Noh and M. Jeon, "A New Framework for Background Subtraction Using Multiple Cues," in *ACCV*. Berlin, Heidelberg: Springer, 2013, pp. 493–506.
- [40] J. Yao and J.-M. Odobez, "Multi-Layer Background Subtraction Based on Color and Texture," in *CVPR*, 2007, pp. 1–8.
- [41] P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin, "Flexible Background Subtraction with Self-Balanced Local Sensitivity," in *CVPRW*, 2014, pp. 414–419.