LoCoFuse: Locality-Constrained Dual-Branch Inference for Training-Free Underwater Open-Vocabulary Semantic Segmentation

Yuhang Zhang, Weidong Tang, Yinxue Shi*
College of Information and Electrical Engineering, China Agricultural University, Beijing, China
Emails: zyh140981@outlook.com, 2023308130229@cau.edu.cn, syx2821@cau.edu.cn

***Abstract*—Underwater open-vocabulary semantic segmentation (UOVS) is a key technology for underwater robotic perception and marine ecological monitoring. However, multimodal foundation models are trained mainly on terrestrial imagery, and the resulting distribution gap severely limits their generalization in underwater scenes; absorption and scattering further degrade images and exacerbate this challenge. We propose LoCoFuse, a training-free inference framework for UOVS that keeps both the vision–language model and the geometric encoder frozen, and improves robustness for UOVS under underwater image degradations through three complementary modules: LoCoMix introduces locality-constrained geometry-guided token mixing to curb long-range semantic leakage; Text-CPC mitigates domain bias via prompt contrast to calibrate category prototypes and fuses image-conditioned cues from a multimodal large language model (MLLM), using these MLLM cues to strengthen text representations; DualMix leverages token-level uncertainty and boundary cues to fuse global and local branches in probability space, balancing region consistency and boundary accuracy. Extensive experiments on UOVSBench demonstrate notable performance gains with efficient inference.**

***Index Terms*—open-vocabulary segmentation, underwater vision, training-free inference, vision–language models**

I. Introduction

As autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) become more widely deployed, robust underwater visual perception is essential for long-term ocean monitoring and intelligent operations [1]. These systems often need to segment marine organisms and man-made debris. However, dense underwater annotations are costly, and most existing underwater segmentation datasets and models remain closed-set with predefined categories [2], [3]. Recent underwater segmentation approaches still assume fixed taxonomies and task-specific training [4].

Large-scale vision–language pre-trained models (VLMs; e.g., CLIP [5]) enable open-vocabulary recognition by aligning visual and textual representations. Their performance often scales predictably with data and model capacity [6]. Recent training-free dense transfer approaches build on this alignment to produce pixel-level predictions with frozen VLMs, without task-specific fine-tuning [7]. More recent refinements further improve dense inference through representation decomposition and proxy attention [8], [9].

*Corresponding author.

We focus on training-free UOVS, leveraging frozen VLMs to segment unseen fine-grained categories specified by text prompts without additional training on target underwater datasets [10]. Direct transfer is complicated by severe domain shift: absorption and scattering cause haziness, low contrast, and color distortions [11], [12]. Such degradations weaken local texture cues and exacerbate semantic ambiguity in downstream perception [13]. This setting calls for inference that balances global semantic consistency with local boundary fidelity and for text cues that remain reliable under underwater degradation.

To address these challenges, we propose LoCoFuse, a training-free inference framework that refines VLM features with geometric cues from a frozen geometric encoder (a depth model) to balance global context with local constraints. LoCoMix introduces a locality constraint into geometry-guided token mixing to reduce long-range semantic leakage and preserve boundaries. DualMix adaptively fuses global and local branches in probability space using uncertainty and boundary cues. Text-CPC calibrates text prototypes by contrasting underwater-specific prompt templates with a generic template bank, improving robustness without fine-tuning.

In summary, our main contributions are threefold: (1) we introduce LoCoMix, a locality-constrained geometry mixer in LoCoFuse that limits geometry-guided token aggregation to reduce long-range semantic leakage and preserve fine boundaries; (2) we propose DualMix and Text-CPC, two training-free adaptation modules that respectively enable uncertainty-aware global–local fusion and prompt calibration in the presence of underwater domain shift; (3) on UOVSBench [10], we achieve state-of-the-art training-free UOVS performance, improving the average mIoU by 4.71–5.51% over the best prior training-free baseline across three backbones.

II. Related Work

Open-vocabulary semantic segmentation (OVS) extends dense prediction to text-defined classes, with paradigms spanning language-driven formulations and grouping-based designs [5], [14]–[16]. Many approaches further rely on prompt optimization or extra supervision; complementary work calibrates global knowledge or uses image-level labels to scale OVS [17]–[20]. In contrast, training-free OVS adapts frozen VLMs to dense prediction without parameter updates, using

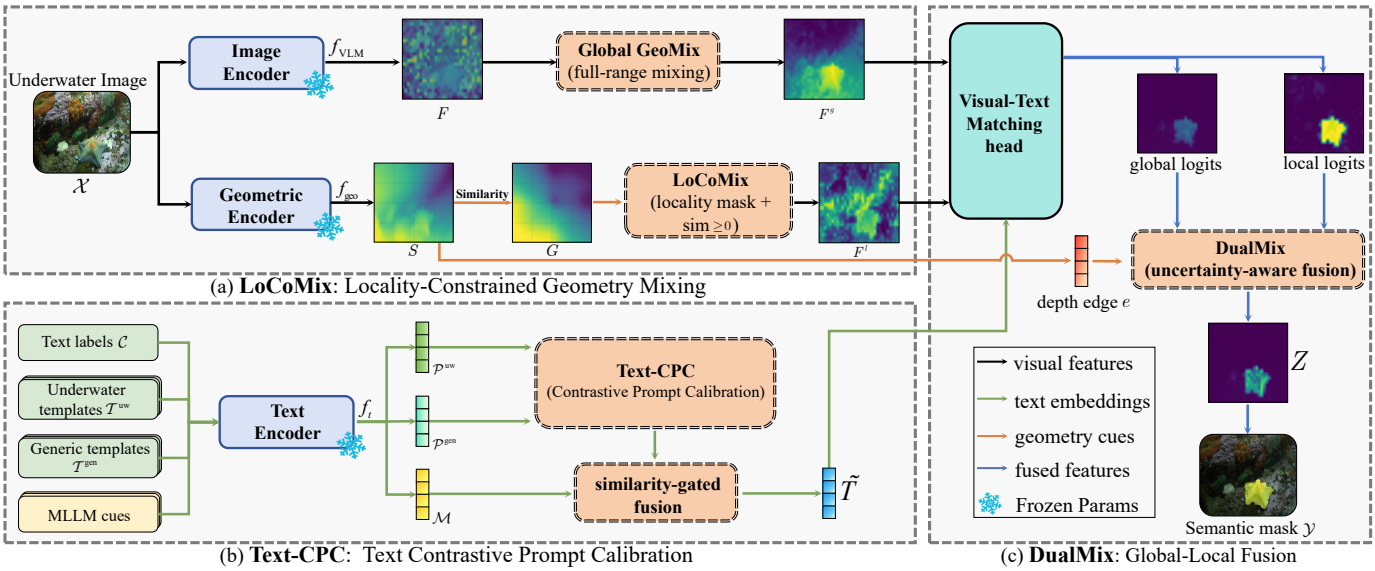

Fig. 1. Framework of LoCoFuse. With frozen vision–language and geometric encoders, LoCoFuse performs training-free inference via a global GeoMix branch, a locality-constrained LoCoMix branch, Text-CPC with image-conditioned MLLM cues, and DualMix probability-level fusion. Orange blocks denote training-free adaptation modules.

MaskCLIP-style dense transfer and structure-aware priors such as SAM masks to improve localization [7]–[9], [21], [22]. Most existing OVS methods are developed and benchmarked on terrestrial imagery, leaving robustness under severe underwater degradation less explored.

Underwater imagery suffers from absorption and scattering, which induce haze, low contrast, and color distortions and lead to pronounced domain shift [11], [12]. Underwater image enhancement and restoration have been widely studied, and benchmark datasets examine their impact on downstream vision tasks [13]. Representative learning-based enhancement methods include WaterGAN [23] and UGAN [24]. Fast enhancement models such as FUnIE-GAN target real-time perception [25]. To mitigate the resulting domain gap for downstream semantic segmentation, domain adaptation has also been explored [26], [27]. However, such approaches rely on additional training and are therefore incompatible with a strict training-free setting.

Underwater segmentation is often formulated as a closed-set task with fixed taxonomies and task-specific training [4], motivating open-vocabulary perception under domain shift. Recent efforts begin to incorporate stronger semantic priors for underwater understanding [28] and marine open-vocabulary instance segmentation [29]. Earth2Ocean establishes UOVS-Bench for systematic training-free evaluation of UOVS [10]. Nevertheless, directly transferring frozen VLMs underwater remains challenging, motivating robust inference-time strategies without extra training.

III. Method

A. Problem Definition

We address *training-free* UOVS. Given an underwater image $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$ and a text-defined class set $\mathcal{C} = \{c_1, \ldots, c_K\}$, the goal is to predict a dense semantic mask $\mathcal{Y} \in [K]^{H\times W}$. Here $[K] = \{1, \ldots, K\}$. In the training-free setting, we do not use underwater datasets to train model parameters; instead, we keep both the VLM and the geometric encoder frozen, and perform segmentation via training-free adaptation modules.

B. Overview

Fig. 1 illustrates the LoCoFuse inference pipeline for training-free UOVS under underwater domain shift. We keep the VLM and geometric encoder frozen, extracting patch-token embeddings F and aligned geometric cues G on the same token grid, and perform segmentation via training-free adaptation modules. LoCoFuse constructs two complementary visual representations, a global GeoMix branch F^g for large-scale consistency and a locality-constrained LoCoMix branch F^l for boundary and small-object preservation under underwater degradation.

LoCoMix/GeoMix performs geometry-guided token mixing driven by G, where LoCoMix restricts interactions within a Chebyshev-radius window r and blends mixed and identity features using α, while GeoMix uses full interactions to form F^g. For open-vocabulary semantics, Text-CPC builds class prototypes from underwater-specific and generic template banks and applies contrastive calibration (scale s), optionally injecting MLLM-conditioned cue embeddings $\mathcal{M}$ to obtain the final prototype set $\tilde{T}$. Token embeddings from F^g and F^l are matched against $\tilde{T}$ to obtain token-level logits, which are converted to per-token probabilities p_i^g and p_i^l using the logit scale λ. DualMix fuses p_i^g and p_i^l using uncertainty and the depth-edge prior e_i from G under safety constraints, performing probability-space fusion and decoding to the final semantic mask $\mathcal{Y}$.

C. LoCoMix: Locality-Constrained Geometry Mixing

a) Geometry-guided similarity.: The frozen VLM encodes $\mathcal{X}$ into $L = H_t W_t$ patch tokens (excluding the CLS token), $F \in \mathbb{R}^{L\times d}$. The frozen geometric encoder outputs a

cue map aligned to the token grid, $G \in \mathbb{R}^{d_g \times H_t \times W_t}$. Denote the ℓ_2-normalized geometric feature at token i as $\hat{\mathbf{g}}_i \in \mathbb{R}^{d_g}$, and define $S_{ij} = \hat{\mathbf{g}}_i^\top \hat{\mathbf{g}}_j$. We apply centering-and-sharpening:

$$\tilde{S}_{ij} = \gamma \left(S_{ij} - \beta \bar{S} \right), \tag{1}$$

where $\bar{S}$ is the mean similarity over token pairs within a batch, and (β, γ) are hyperparameters controlling the centering strength and the sharpening scale.

b) Locality mask, hard pruning, and mixing.: LoCoMix restricts token interactions using a Chebyshev-radius window on the token grid and a hard similarity pruning:

$$\begin{aligned} M_{ij} &= \mathbb{I}[\max(|u_i - u_j|,\ |v_i - v_j|) \le r], \\ \tilde{S}_{ij}^{\text{loc}} &= \begin{cases} \tilde{S}_{ij}, & M_{ij} = 1 \text{ and } \tilde{S}_{ij} \ge 0, \\ -\infty, & \text{otherwise}, \end{cases} \\ A_{ij}^{\text{loc}} &= \text{softmax}_j(\tilde{S}_{ij}^{\text{loc}}), \\ F^l &= (1-\alpha)F + \alpha\, A^{\text{loc}} F, \end{aligned} \tag{2}$$

where (u_i, v_i) is token i's grid coordinate, r is the Chebyshev radius of the local window on the token grid, and α is the identity blending coefficient. The global GeoMix branch follows the same GeoMix formulation with full token interactions ($M_{ij} \equiv 1$), hard pruning at $\tilde{S}_{ij} < 0$, and $\alpha = 1$, producing F^g.

D. Text-CPC: Contrastive Prompt Calibration

a) Template-averaged prototypes.: We encode prompts using a frozen text encoder $f_t(\cdot)$ and two template banks: $\mathcal{T}^{\text{uw}}$ (underwater-specific) and $\mathcal{T}^{\text{gen}}$ (generic). We denote the instantiated prompt-string sets for class c as $\mathcal{Q}^{\text{uw}}(c)$ and $\mathcal{Q}^{\text{gen}}(c)$. For each class $c \in \mathcal{C}$, we compute underwater-specific and generic prototypes by averaging ℓ_2-normalized prompt embeddings and re-normalizing:

$$\begin{aligned} \bar{\mathbf{t}}_c^{\text{uw}} &= \frac{1}{|\mathcal{Q}^{\text{uw}}(c)|} \sum_{q \in \mathcal{Q}^{\text{uw}}(c)} \frac{f_t(q)}{\|f_t(q)\|_2}, \qquad \mathbf{t}_c^{\text{uw}} = \frac{\bar{\mathbf{t}}_c^{\text{uw}}}{\|\bar{\mathbf{t}}_c^{\text{uw}}\|_2}, \\ \bar{\mathbf{t}}_c^{\text{gen}} &= \frac{1}{|\mathcal{Q}^{\text{gen}}(c)|} \sum_{q \in \mathcal{Q}^{\text{gen}}(c)} \frac{f_t(q)}{\|f_t(q)\|_2}, \qquad \mathbf{t}_c^{\text{gen}} = \frac{\bar{\mathbf{t}}_c^{\text{gen}}}{\|\bar{\mathbf{t}}_c^{\text{gen}}\|_2}. \end{aligned} \tag{3}$$

We denote the template-averaged prototype sets as $\mathcal{P}^{\text{uw}} = \{\mathbf{t}_c^{\text{uw}}\}_{c \in \mathcal{C}}$ and $\mathcal{P}^{\text{gen}} = \{\mathbf{t}_c^{\text{gen}}\}_{c \in \mathcal{C}}$.

b) Contrastive calibration.: Given $\mathcal{P}^{\text{uw}}$ and $\mathcal{P}^{\text{gen}}$, Text-CPC calibrates text prototypes by contrasting underwater-specific and generic templates:

$$\mathbf{t}_c = \frac{\mathbf{t}_c^{\text{uw}} + s\left(\mathbf{t}_c^{\text{uw}} - \mathbf{t}_c^{\text{gen}}\right)}{\|\mathbf{t}_c^{\text{uw}} + s\left(\mathbf{t}_c^{\text{uw}} - \mathbf{t}_c^{\text{gen}}\right)\|_2}, \tag{4}$$

where $s \ge 0$ is the Text-CPC scale.

c) MLLM-conditioned prototype fusion.: We prompt an image-conditioned MLLM to generate (1) a concise image description, (2) an object list drawn from the class set $\mathcal{C}$, and (3) attribute descriptions for each listed object. We encode these cues with the same frozen text encoder, obtaining a set of normalized embeddings $\mathcal{M} = \{\mathbf{m}_n\}$. Each $\mathbf{m}_n$ is assigned to its closest class and injected into that class text prototype with a bounded, similarity-gated weight:

$$\begin{aligned} \pi(n) &= \arg\max_{c \in \mathcal{C}} \cos(\mathbf{m}_n, \mathbf{t}_c), \\ w_{cn} &= \min(w_{\max},\ \cos(\mathbf{m}_n, \mathbf{t}_{\pi(n)})) \cdot \mathbb{I}[c = \pi(n)] \\ &\quad \cdot \mathbb{I}\big[\cos(\mathbf{m}_n, \mathbf{t}_{\pi(n)}) \ge \delta\big], \\ \tilde{\mathbf{t}}_c &= \frac{\mathbf{t}_c + \sum_n w_{cn} \mathbf{m}_n}{\|\mathbf{t}_c + \sum_n w_{cn} \mathbf{m}_n\|_2}, \end{aligned} \tag{5}$$

where δ is the similarity threshold for accepting a cue and $w_{\max}$ is the maximum fusion weight. We denote the final prototype set as $\tilde{T} = \{\tilde{\mathbf{t}}_c\}_{c \in \mathcal{C}}$.

E. DualMix: Uncertainty-Aware Global–Local Fusion

a) Token–text similarity.: Let $\mathbf{f}_i^g, \mathbf{f}_i^l \in \mathbb{R}^d$ be token embeddings from F^g and F^l. We ℓ_2-normalize token embeddings and compute cosine similarity against text prototypes $\tilde{T}$:

$$\begin{aligned} z_{i,c}^g &= \Big(\frac{\mathbf{f}_i^g}{\|\mathbf{f}_i^g\|_2}\Big)^\top \tilde{\mathbf{t}}_c, \qquad z_{i,c}^l = \Big(\frac{\mathbf{f}_i^l}{\|\mathbf{f}_i^l\|_2}\Big)^\top \tilde{\mathbf{t}}_c, \\ p_i^g &= \text{softmax}(\lambda z_i^g), \qquad p_i^l = \text{softmax}(\lambda z_i^l), \end{aligned} \tag{6}$$

where $z_i^g, z_i^l \in \mathbb{R}^K$ are token-level logits and λ is the VLM logit scale.

b) Uncertainty-aware gating weights.: We measure uncertainty by normalized entropy:

$$\mathcal{H}(p) = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log p_k. \tag{7}$$

DualMix computes per-token gating weights between the global GeoMix and locality-constrained LoCoMix branches by combining (i) relative confidence gating, (ii) geometric and uncertainty priors, and (iii) safety constraints. The geometric prior $e_i \in [0, 1]$ is computed from the depth-edge magnitude on the token grid. We first obtain pseudo-label predictions from both branches:

$$\hat{c}_i^g = \arg\max_{c \in \mathcal{C}} p_{i,c}^g, \qquad \hat{c}_i^l = \arg\max_{c \in \mathcal{C}} p_{i,c}^l. \tag{8}$$

The gating weight is computed as:

$$\begin{aligned} w_i^{\text{conf}} &= \sigma\big(\kappa\left(\mathcal{H}(p_i^g) - \mathcal{H}(p_i^l)\right)\big), \\ w_i^{\text{reg}} &= 0.5\,\mathcal{H}(p_i^g) + 0.5\,e_i, \qquad w_i = w_i^{\text{conf}} \cdot w_i^{\text{reg}}, \\ \tilde{w}_i &= w_i \cdot \mathbb{I}\big[\hat{c}_i^g \ne \hat{c}_i^l\big] \cdot \mathbb{I}\Big[p_{i,\hat{c}_i^l}^g \ge \tau\Big], \end{aligned} \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function, κ controls sharpness, and τ is a small threshold. Here w_i^{conf} measures relative confidence via the entropy gap, and w_i^{reg} leverages the geometric prior e_i and entropy to emphasize local cues around object boundaries and high-uncertainty regions. The final $\tilde{w}_i$ enforces safety mechanisms: fusion is performed only when the two branches disagree ($\hat{c}_i^g \ne \hat{c}_i^l$) and the local suggestion is sufficiently compatible with the global distribution ($p_{i,\hat{c}_i^l}^g \ge \tau$).

TABLE I

COMPARISON WITH PRIOR TRAINING-FREE OVS METHODS ON UOVSBENCH IN TERMS OF AACC/MIOU/MACC. FOR EACH BACKBONE AND METRIC, BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED, RESPECTIVELY.

Backbone	Method	DUT-USEG			MAS3K			SUIM			USIS10K			USIS16K			AquaOV255			Avg.		
		aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc
ViT-B/16	ClearCLIP [8]	18.10	5.55	25.25	43.54	23.11	31.84	62.57	26.08	42.98	54.01	26.59	37.04	29.31	16.78	26.29	20.93	11.29	18.16	38.08	18.23	30.26
	SCLIP [30]	34.90	18.88	45.42	37.66	17.55	29.48	64.34	37.60	61.51	63.34	33.09	47.57	27.26	14.26	25.69	18.12	8.03	15.61	40.94	21.57	37.55
	ProxyCLIP [9]	24.74	13.24	35.74	44.76	24.05	36.65	73.07	48.47	63.40	68.86	39.10	51.08	35.34	20.45	32.66	25.00	12.44	21.60	45.30	26.29	40.19
	Trident [22]	22.15	12.29	33.81	45.85	25.76	38.03	73.35	48.72	62.39	68.07	39.05	50.26	37.88	22.86	35.06	26.88	14.13	23.57	45.70	27.14	40.52
	CorrCLIP [31]	27.24	16.81	36.41	48.26	27.19	41.55	**73.87**	51.00	68.05	67.56	40.39	53.15	42.44	25.95	39.54	30.57	16.02	26.54	48.32	29.56	44.21
	Earth2Ocean [10]	52.69	34.07	53.64	50.42	28.41	42.34	73.06	51.97	72.73	71.85	44.63	59.24	45.42	29.02	43.03	32.61	17.81	28.06	54.34	34.32	49.84
	LoCoFuse	**67.25**	**45.16**	**59.19**	**52.11**	**31.03**	**44.88**	73.27	**52.37**	**72.75**	**72.39**	**45.76**	**60.46**	**53.31**	**36.85**	**50.29**	**39.27**	**22.98**	**34.50**	**59.60**	**39.03**	**53.68**
ViT-L/14	ClearCLIP [8]	29.31	14.90	44.36	51.22	32.09	48.98	65.88	32.27	45.43	50.06	28.15	39.50	39.96	25.46	36.52	32.52	18.76	29.05	44.83	25.27	40.64
	SCLIP [30]	55.66	33.38	58.80	46.50	25.33	41.66	56.72	34.20	60.27	62.19	33.04	46.98	33.84	20.46	32.71	26.21	14.56	23.75	46.85	26.83	44.03
	ProxyCLIP [9]	58.05	32.16	58.83	54.79	33.30	53.10	72.37	51.65	66.01	65.43	41.99	54.29	45.23	30.55	42.35	37.28	22.29	33.03	55.53	35.32	51.27
	Trident [22]	54.10	29.41	56.54	54.96	34.25	54.53	72.28	51.47	64.66	65.56	41.68	53.42	47.22	32.13	43.86	38.70	23.44	34.29	55.47	35.40	51.22
	CorrCLIP [31]	62.93	37.47	58.53	56.97	34.57	54.01	74.01	54.95	71.38	64.37	42.81	57.38	46.99	31.32	43.59	38.54	22.85	34.12	57.30	37.33	53.17
	Earth2Ocean [10]	68.28	41.32	61.86	63.99	40.94	61.87	74.27	55.17	75.81	70.36	46.90	61.45	59.97	45.13	58.04	52.59	34.53	47.74	64.91	44.00	61.13
	LoCoFuse	**74.50**	**51.00**	**68.39**	**68.39**	**46.02**	**65.35**	**74.40**	**55.33**	**75.83**	**70.39**	**47.12**	**61.68**	**67.04**	**51.69**	**64.96**	**61.62**	**43.22**	**57.25**	**69.39**	**49.06**	**65.58**
ViT-H/14	ClearCLIP [8]	22.69	13.23	41.83	60.53	39.26	57.64	65.21	28.41	43.88	47.83	22.80	31.80	51.34	36.08	49.07	40.75	26.42	37.93	48.06	27.70	43.69
	SCLIP [30]	53.60	30.80	57.53	47.95	26.14	43.11	60.57	34.43	59.04	56.09	33.03	49.84	38.90	25.32	39.92	28.72	16.58	27.14	47.64	27.72	46.10
	ProxyCLIP [9]	50.86	31.29	55.21	64.81	37.91	56.33	75.11	52.34	67.62	70.93	41.26	50.43	53.12	37.89	52.42	41.32	26.41	38.88	59.36	37.85	53.48
	Trident [22]	42.73	27.15	51.87	64.08	40.43	58.34	75.76	52.83	66.39	71.05	40.96	49.56	57.94	42.43	56.46	45.61	30.17	43.04	59.53	39.00	54.28
	CorrCLIP [31]	54.15	36.10	55.44	62.43	40.34	57.73	76.20	54.34	70.84	**75.04**	44.88	54.27	57.82	42.35	56.01	44.86	29.57	41.56	61.75	41.26	55.98
	Earth2Ocean [10]	74.37	55.24	67.66	67.26	47.15	67.40	78.34	61.04	78.85	72.62	49.12	62.04	60.45	45.68	59.99	55.98	39.76	53.37	68.17	49.67	64.89
	LoCoFuse	**77.69**	**60.45**	**72.78**	**70.70**	**51.47**	**70.00**	**78.63**	**61.51**	**79.02**	72.68	**49.49**	**62.21**	**72.19**	**59.24**	**71.23**	**64.47**	**48.93**	**62.47**	**72.73**	**55.18**	**69.62**

c) Probability-level fusion and output.: We fuse predictions in the probability space and recover pre-scale logits, obtaining token-level logits $Z = \{z_i\}_{i=1}^{L} \in \mathbb{R}^{L\times K}$. Finally, we reshape Z to the $H_t \times W_t$ grid and upsample it to the original image resolution $H \times W$, outputting the final semantic mask:

$$p_i = \tilde{w}_i p_i^l + (1-\tilde{w}_i)p_i^g, \qquad z_i = \frac{1}{\lambda}\log(p_i),$$
$$\mathcal{Y} = \arg\max_{k\in\{1,\dots,K\}} \mathrm{Up}(Z)_k. \tag{10}$$

IV. EXPERIMENTS

A. Experimental Details

We conduct experiments on the UOVSBench benchmark [10] (DUT-USEG [3], MAS3K [4], SUIM [2], USIS10K [32], USIS16K [33], and AquaOV255 [10]) and follow its official evaluation protocol. We use OpenCLIP backbones [6] as the frozen VLM and a frozen geometric encoder (DepthAnythingV2 [34]) to provide geometric cues. For high-resolution images, we perform sliding-window inference. Unless stated otherwise, we set a LoCoMix radius of r=2, a blending coefficient α=0.6 (the global GeoMix branch uses α=1), s=0.5, β=1.2, γ=3.0, λ=40, and κ=20. We enable image-conditioned MLLM cue embeddings by default and use GPT-4o as the MLLM in our evaluation. All results are obtained without fine-tuning and are computed on a single NVIDIA GeForce RTX 4090 GPU.

B. Comparisons with State-of-the-Art Methods

Quantitative comparisons. Table I summarizes UOVSBench results across three OpenCLIP backbones. LoCoFuse achieves the best average performance for all backbones, outperforming prior training-free OVS baselines. The improvements tend to be more pronounced on instance-dense datasets (DUT-USEG, MAS3K, USIS16K, and AquaOV255), suggesting that enforcing locality helps suppress cross-class leakage and preserve fine structures. Notably, the gains generally increase with larger backbones, indicating that LoCoFuse can better exploit richer vision–language representations under underwater domain shift.

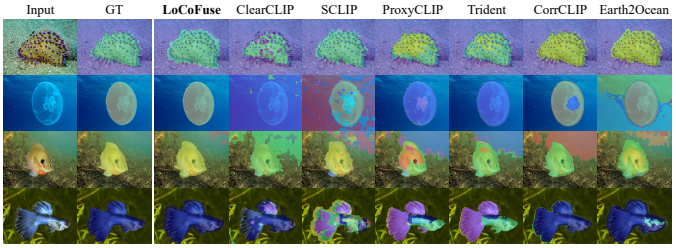

Fig. 2. Visualization of segmentation results on UOVSBench. Columns show the input image, ground truth, and predictions by LoCoFuse and training-free baselines.

Qualitative comparisons. Fig. 2 provides qualitative comparisons with representative training-free baselines. LoCoFuse tends to produce cleaner object extents and boundaries with fewer cross-class artifacts in cluttered scenes. Compared with baselines that may produce plausible masks but confuse categories, LoCoFuse generally yields more semantically consistent predictions.

C. Ablation Studies and Analyses

Each component contributes to LoCoFuse under ViT-L/14. Table II shows that each component contributes measurably, with Text-CPC and image-conditioned MLLM cue embeddings providing particularly strong gains under underwater

TABLE II
COMPONENT ABLATION OF LOCOFUSE ON DUT-USEG AND MAS3K (VIT-L/14) IN TERMS OF AACC/MIOU/MACC.

Variant	DUT-USEG			MAS3K		
	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc
w/o MLLM	68.84	43.59	62.52	56.01	34.39	56.64
w/o Text-CPC	73.16	47.28	66.25	64.70	42.16	62.07
w/o DualMix	73.53	50.30	67.99	64.90	42.57	62.31
w/o LoCoMix	73.81	49.99	67.59	68.33	45.85	65.23
LoCoFuse	74.50	51.00	68.39	68.39	46.02	65.35

TABLE III
IMPACT OF DIFFERENT MLLMS ON LOCOFUSE (VIT-L/14).

Method	DUT-USEG			MAS3K		
	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc
LoCoFuse (GPT-4o [35])	74.50	51.00	68.39	68.39	46.02	65.35
LoCoFuse (Qwen2.5VL-3B [36])	70.20	49.72	68.07	61.37	39.27	59.68
LoCoFuse (Qwen2.5VL-7B [36])	70.47	47.54	66.44	61.77	40.78	63.67

domain shift. Disabling either LoCoMix or DualMix also reduces accuracy, suggesting complementary roles: locality-constrained mixing mitigates cross-class leakage in cluttered regions, while the gating mechanism restores global consistency when locality is unreliable.

The MLLM choice affects cue quality and segmentation performance. We keep the backbone and all hyperparameters fixed, and only change the MLLM used to generate image-conditioned MLLM cue embeddings. Table III reports results with different MLLMs under the same setting, suggesting that richer and more reliable descriptions can provide stronger semantic grounding for training-free UOVS.

Performance is stable across r and α, while s is more sensitive. Fig. 3 suggests that LoCoFuse is relatively insensitive to the locality radius within $r \in [1, 4]$ and remains stable across a reasonable range of the mixing coefficient α. Fig. 4 further shows that the Text-CPC scale has a more pronounced and sometimes non-monotonic effect: performance typically improves with a moderate s, while overly large s can be detrimental. This trend reflects a trade-off between calibrating domain bias and over-amplifying the contrastive shift. We set

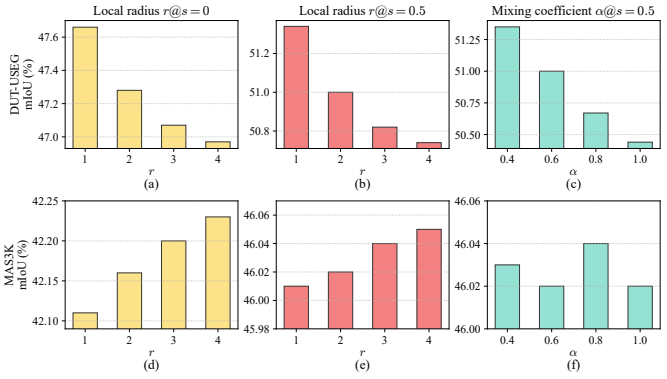

Fig. 3. Sensitivity to the locality radius r and the mixing coefficient α (ViT-L/14). Rows show results on DUT-USEG and MAS3K. The three columns correspond to varying r at s=0 (r@s=0), varying r at s=0.5 (r@s=0.5), and varying α at s=0.5 (α@s=0.5).

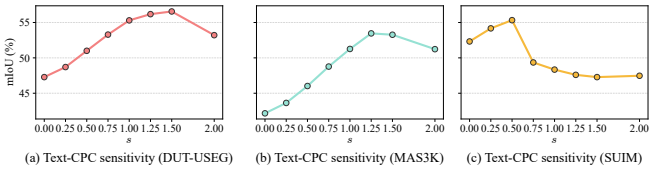

Fig. 4. Text-CPC scale sensitivity (ViT-L/14) on DUT-USEG, MAS3K, and SUIM.

TABLE IV
EFFICIENCY ON UOVSBENCH (VIT-L/14). PEAK MEMORY (MB), FLOPS (G; VIA THOP), AND THROUGHPUT (FPS) ARE MEASURED ON A SINGLE NVIDIA GEFORCE RTX 4090 WITH BATCH SIZE 1 UNDER EACH METHOD'S DEFAULT SLIDING-WINDOW INFERENCE.

Method	Memory (MB)	FLOPs (G)	FPS	DUT mIoU (%)	MAS mIoU (%)
ClearCLIP	988.16	465.85	30.91	14.90	32.09
SCLIP	987.51	726.53	6.81	33.38	25.33
ProxyCLIP	1335.78	1571.86	5.12	32.16	33.30
Trident	2521.01	1565.06	2.50	29.41	34.25
CorrCLIP	3155.06	1602.11	1.27	37.47	34.57
Earth2Ocean	1393.97	1072.69	23.30	41.32	40.94
LoCoFuse	1437.23	1742.72	13.91	51.00	46.02

s=0.5 as a conservative choice across datasets.

Efficiency comparisons. Table IV reports peak memory, FLOPs, and throughput under identical ViT-L/14 settings on an NVIDIA GeForce RTX 4090. Under batched sliding-window inference, LoCoFuse improves accuracy over prior training-free baselines while maintaining a favorable balance between accuracy and efficiency.

D. DualMix Gating Visualization

Fig. 5 shows the input image and the corresponding gate heatmap. Warmer colors indicate higher weight assigned to the locality-constrained branch (LoCoMix), while cooler colors indicate preference for the global GeoMix branch. The gate assigns higher weights to the locality-constrained (LoCoMix) branch around uncertain regions and object boundaries, and relies more on the global GeoMix branch in confident interior regions.

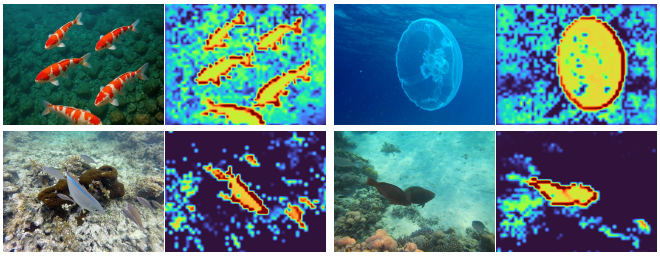

Fig. 5. Visualization of DualMix gating weights $\tilde{w}_i \in [0, 1]$ in LoCoFuse (warm = higher weight on LoCoMix). The gate concentrates on boundaries and uncertain regions, enabling training-free global–local fusion without injecting local noise into confident areas.

V. CONCLUSION

We present LoCoFuse, a training-free inference framework for underwater open-vocabulary semantic segmentation. LoCoFuse reduces long-range semantic leakage via geometry-guided mixing (LoCoMix), balances global consistency and boundary fidelity through uncertainty-aware DualMix fusion, and mitigates domain shift in text guidance using Text-CPC.

Without any fine-tuning, LoCoFuse improves performance on challenging underwater benchmarks. LoCoFuse is still constrained by the knowledge coverage of the underlying foundation models and may struggle with rare, heavily occluded, or strongly camouflaged organisms. Future work will explore parameter-efficient fine-tuning techniques to inject underwater biological knowledge into multimodal models, thereby further enhancing their performance in complex underwater scenarios.

ACKNOWLEDGMENT

This research was funded by the Yunnan Province 2025 Central Government-Guided Local Science and Technology Development Fund, "Research and Demonstration of the Lancang-Mekong Digital Coffee Triple-Chain Coupling Technology Innovation System", grant number 202507AD040001.

REFERENCES

[1] J. Xu, X. Irigoien, and M.-S. Alouini, "State-of-the-art underwater vehicles and technologies enabling smart ocean: Survey and classifications," *arXiv preprint arXiv:2412.18667*, 2024.

[2] M. J. Islam, C. Edge, Y. Xiao, P. Luo *et al.*, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE Press, 2020, pp. 1769–1776.

[3] Z. Ma, H. Li, Z. Wang, D. Yu *et al.*, "An underwater image semantic segmentation method focusing on boundaries and a real underwater scene semantic segmentation dataset," *arXiv preprint arXiv:2108.11727*, 2021.

[4] L. Li, B. Dong, E. Rigall, T. Zhou *et al.*, "Marine animal segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2303–2314, 2022.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[6] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman *et al.*, "Reproducible scaling laws for contrastive language-image learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 2818–2829.

[7] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2022, pp. 696–712.

[8] M. Lan, C. Chen, Y. Ke, X. Wang *et al.*, "Clearclip: Decomposing clip representations for dense vision-language inference," in *European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 143–160.

[9] M. Lan, C. Chen, Y. Ke, X. Wang *et al.*, "ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation," in *European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2025, pp. 70–88.

[10] B. Li, T. Huo, D. Zhang, Z. Zhao *et al.*, "Exploring the underwater world segmentation without extra training," *arXiv preprint arXiv:2511.07923*, 2025.

[11] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1682–1691.

[12] Y.-W. Chen and S.-C. Pei, "Domain adaptation for underwater image enhancement via content and style separation," *IEEE Access*, vol. 10, pp. 90 523–90 534, 2022.

[13] C. Li, C. Guo, W. Ren, R. Cong *et al.*, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020.

[14] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun *et al.*, "Language-driven semantic segmentation," in *International Conference on Learning Representations (ICLR)*, 2022.

[15] J. Xu, S. De Mello, S. Liu, W. Byeon *et al.*, "Groupvit: Semantic segmentation emerges from text supervision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 134–18 144.

[16] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 583–11 592.

[17] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[18] K. Han, Y. Liu, J. H. Liew, H. Ding *et al.*, "Global knowledge calibration for fast open-vocabulary segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 797–807.

[19] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2022, pp. 540–557.

[20] X. Zou, J. Yang, H. Zhang, F. Li *et al.*, "Segment everything everywhere all at once," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao *et al.*, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4015–4026.

[22] Y. Shi, M. Dong, and C. Xu, "Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 23 487–23 497.

[23] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 387–394, 2018.

[24] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7159–7165.

[25] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[26] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn *et al.*, "Learning to adapt structured output space for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7472–7481.

[27] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 9924–9935.

[28] W. Xu, C. Wang, D. Liang, Z. Zhao *et al.*, "Nautilus: A large multimodal model for underwater scene understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[29] B. Li, F. Wang, D. Zhang, Z. Zhao *et al.*, "Maris: Marine open-vocabulary instance segmentation with geometric enhancement and semantic alignment," *arXiv preprint arXiv:2510.15398*, 2025.

[30] F. Wang, J. Mei, and A. Yuille, "SCLIP: Rethinking self-attention for dense vision-language inference," in *European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2025, pp. 315–332.

[31] D. Zhang, F. Liu, and Q. Tang, "Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 24 677–24 687.

[32] S. Lian, Z. Zhang, H. Li, W. Li *et al.*, "Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 21–27 Jul 2024, pp. 29 545–29 559. [Online]. Available: https://proceedings.mlr.press/v235/lian24c.html

[33] L. Hong, X. Wang, Y. Li, and X. Wang, "Usis16k: High-quality dataset for underwater salient instance segmentation," *arXiv preprint arXiv:2506.19472*, 2025.

[34] L. Yang, B. Kang, Z. Huang, Z. Zhao *et al.*, "Depth anything v2," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37. Curran Associates, Inc., 2024, pp. 21 875–21 911.

[35] OpenAI, "GPT-4o system card," 2024. [Online]. Available: https://arxiv.org/abs/2410.21276

[36] S. Bai, K. Chen, X. Liu, J. Wang *et al.*, "Qwen2.5-VL technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923