MARIS: Marine Open-Vocabulary Instance Segmentation

Supplementary Material

A. Robustness Analysis of SAIM

Table 8. **Ablation study on TopN selection.** We report mAP, AP_{50}, and AP_{75} for different TopN values.

TopN	mAP	AP_{50}	AP_{75}
1	41.73	48.04	45.28
2	43.83	50.45	47.61
5	43.97	50.62	47.77
10	**44.37**	50.99	48.11
20	**44.37**	**51.41**	47.90
50	44.42	51.07	**48.18**
80	44.33	51.01	48.11

The SAIM module demonstrates strong robustness to the choice of TopN. Its template selection mechanism remains stable across different TopN settings, maintaining consistent segmentation performance. This insensitivity reduces the need for extensive hyperparameter tuning and ensures reliable performance.

B. Implementation Details

- **EOVSeg**: We set `NUM_STAGE` to 1, and adopted `ViT-B/16` as an auxiliary encoder. For CLIP pre-trained parameters, we experimented with both `ConvNeXt-B` and `ConvNeXt-L`.
- **FCCLIP**: The model was configured with `TRANSFORMER_ENC_LAYERS` = 6 and `DEC_LAYERS` = 10, and employed CLIP pre-trained weights from both `ConvNeXt-B` and `ConvNeXt-L`.
- **MAFT+**: We adopted the same transformer settings (`TRANSFORMER_ENC_LAYERS` = 6 and `DEC_LAYERS` = 10), with CLIP pre-training based on `ConvNeXt-B` and `ConvNeXt-L`.
- **MARIS**: We followed the same setting as FCCLIP and MAFT+, i.e., `TRANSFORMER_ENC_LAYERS` = 6 and `DEC_LAYERS` = 10, with CLIP pre-trained parameters from `ConvNeXt-B`and `ConvNeXt-L`.

For all other hyperparameters, we followed the original papers.

C. Code Release:

Full code and model weights are available at appendix. Includes: (1) Preprocessing scripts for MARIS dataset; (2) How to intall the environment to start the expriments. (3) How to run the code to reproduce our results.

D. Template Selection Strategy

I. Mixed-based Selection. Given the similarity tensor $\mathcal{S} \in \mathbb{R}^{B\times H\times W\times K\times T}$ between image patches and text templates, we compute the average score across spatial positions:

$$\bar{\mathcal{S}}_{b,k,t} = \frac{1}{H\cdot W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathcal{S}_{b,h,w,k,t}, \quad \bar{\mathcal{S}} \in \mathbb{R}^{B\times K\times T}. \tag{12}$$

For each category k, we rank the template indices t according to $\bar{\mathcal{S}}_{b,k,t}$ and select the top-N candidates. The corresponding embeddings are gathered and averaged across batches:

$$\mathbf{E}_k^{\text{top}} = \frac{1}{B\cdot N}\sum_{b=1}^{B}\sum_{t\in \text{TopN}(\bar{\mathcal{S}}_{b,k,:})}\mathbf{E}_{k,t}. \tag{13}$$

To balance global and local information, the final category embedding is obtained by interpolating between the aggregated top-N features and the overall average embedding:

$$\mathbf{E}_k = \lambda\cdot\mathbf{E}_k^{\text{top}} + (1-\lambda)\cdot\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}_{k,t}, \tag{14}$$

where λ controls the contribution of top-ranked templates. This strategy emphasizes the most discriminative templates while retaining global semantic consistency.

II. Weighted Top-N Enhancement. Alternatively, we introduce an adaptive weighting scheme to explicitly enhance the contribution of high-confidence templates. Based on the mean similarity $\bar{\mathcal{S}}_{b,k,t}$, we identify the top-$N$ templates per category k and construct a binary mask $\mathcal{M}_{b,k,t}$ where $\mathcal{M}_{b,k,t} = 1$ if t is in the top-N set and 0 otherwise. Each selected template is assigned an enhancement factor $\alpha > 1$:

$$W_{b,k,t} = \begin{cases}\alpha, & \text{if } \mathcal{M}_{b,k,t} = 1,\\ 1, & \text{otherwise.}\end{cases} \tag{15}$$

The weights are normalized across templates to form a probability distribution:

$$\tilde{W}_{b,k,t} = \frac{W_{b,k,t}}{\sum_{t=1}^{T} W_{b,k,t}}. \tag{16}$$

The final category embedding is then computed as the weighted sum of template features:

$$\mathbf{E}_k = \frac{1}{B}\sum_{b=1}^{B}\sum_{t=1}^{T}\tilde{W}_{b,k,t}\cdot\mathbf{E}_{k,t}. \tag{17}$$

This strategy adaptively emphasizes high-confidence templates without discarding others, leading to a more robust and discriminative representation.

Practical Consideration. To ensure efficient training and evaluation, we adopt a simplified yet effective strategy by performing template selection with only a single randomly sampled image per category. Although this reduces the computational cost substantially, our experiments demonstrate that even a single image provides sufficient discriminative signal to reliably identify informative templates.

E. Dataset Diversity Analysis

Instance Diversity. To provide a comprehensive understanding of category coverage in MARIS, we analyze the distribution of instances across the validation set, as illustrated in Fig. 8-Fig. 10. We visualize the relationship between instance counts and category IDs across different splits of the MARIS dataset. Fig. 8 reports the distribution of intersection classes shared between training and validation, revealing substantial imbalance where frequent species (e.g., common reef fish) dominate the samples, while rare species contain fewer than 60 instances. Fig. 9 focuses on the open-vocabulary (OV) classes that appear only in the validation set. Although MARIS contains 74 OV categories, their frequency varies significantly, indicating that models must handle long-tailed distributions when generalizing to unseen classes. Finally, Fig. 10 presents the overall class distribution, highlighting the combined imbalance across both seen and unseen categories.

This analysis demonstrates that MARIS is not only fine-grained but also diverse, covering a wide range of marine organisms, man-made objects, and substrates. At the same time, the inherent long-tailed distribution reflects real-world underwater environments, where rare species often occur sparsely. Thus, MARIS provides a challenging yet realistic benchmark for evaluating the generalization ability of open-vocabulary segmentation models.

Category Diversity. Following the parent category taxonomy defined in [13], we analyze the category diversity of our dataset, as summarized in Tables 9, 10, and 11. This analysis highlights the extensive coverage of both common and rare underwater object classes, illustrating the richness of our dataset. Compared to previous datasets such as WaterMask [24] and UWSAM [21], our dataset not only includes a broader set of categories but also demonstrates a more balanced and rational parent category organization. The breakdown into Intersection, OV, and Overall classes further supports the validity of our category design, emphasizing the dataset's potential for training robust models and evaluating generalization across diverse underwater scenarios.

F. Dataset Image Feature Analysis

The underwater validation set is analyzed across nine dimensions (in Fig. 7), spanning color space, perceptual quality, and geometric attributes. These distributions reveal characteristics highly adapted to underwater imaging conditions, providing crucial support for model evaluation in this domain. **Color space.** The RGB channels exhibit balanced distributions within the 0–250 intensity range, with frequencies concentrated in mid-level values (300–500 counts), mitigating bias from single-color dominance caused by light scattering. Hue follows a "middle-high, low-at-extremes" distribution with peaks around 400 counts, reflecting the prevalence of neutral tones consistent with water transparency and plankton density. Saturation is concentrated in the 40–120 range (500–600 counts), with low contributions at both extremes, aligning with the natural attenuation of vivid colors caused by underwater light refraction. **Perceptual quality.** Contrast shows a monotonic increase across the 0–100 range, peaking at 600 counts within 80–100, which counteracts blurring induced by turbidity. Brightness values are concentrated in the 100–200 range with probability density 0.015–0.0175, corresponding well to illumination variations across depths, thus ensuring visual clarity and feature discriminability. **Geometric attributes.** Image width (0–7000 pixels) and height (0–5000 pixels) are concentrated in mid-scales, with peaks in 2000–4000 (width, 3500 counts) and 2000–3000 (height, 2500 counts). Image sizes in the 2×10^6–6×10^6 pixel range dominate (7000 counts). Aspect ratios are primarily distributed between 1.0–2.0 (peak 3500 counts), which matches standard underwater camera formats while preserving object integrity for targets such as corals and fish. Overall, the validation set exhibits feature distributions that align closely with underwater optical characteristics, environmental conditions, and imaging requirements, thereby providing a reliable basis for assessing model generalization in tasks such as underwater object detection and scene segmentation.

G. Acknowledgement of Data Sources

We would like to formally acknowledge the contributions of the following datasets, which serve as the foundation for MARIS. The WaterMask [24] dataset provides richly annotated underwater imagery for diverse scene understanding tasks. Additionally, the recently released underwater datasets USIS16K [13], UWSAM [21], and the semantic segmentation dataset by [15] have been systematically re-annotated and extended to ensure consistency and comprehensive coverage. We are grateful for the efforts of the original dataset creators, whose careful data collection and an-

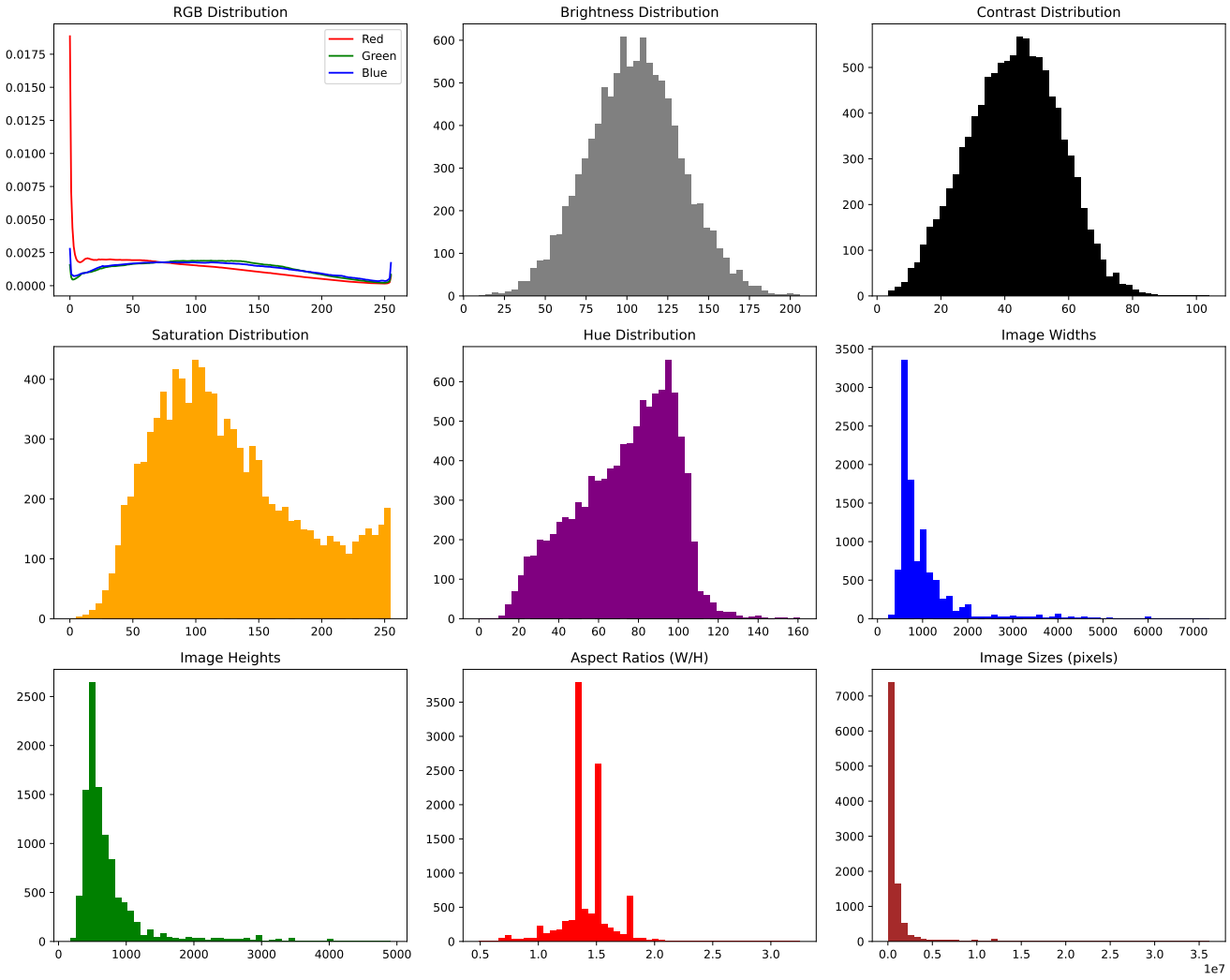

Figure 7. **Validation Set Image Feature Analysis.** Comprehensive analysis of the underwater validation set across nine dimensions, including *color space* (RGB distribution, hue, saturation), *perceptual quality* (contrast, brightness), and *geometric attributes* (width, height, resolution, aspect ratio).

notation make this work possible.

H. Underwater Prompts

To effectively adapt text embeddings to underwater semantics, we design a comprehensive collection of domain-aware prompt templates. Beyond generic templates (e.g., "a photo of a {}"), our design incorporates five additional dimensions that explicitly capture the unique characteristics of underwater imagery: *environment*, *medium/visibility*, *lighting*, *depth*, and *scene interaction*, as summarized in Appendix Tab. 12-Tab. 14.

Environment-oriented prompts describe contextual backgrounds such as coral reefs, caves, or shipwrecks (e.g., "a {} near a coral reef"), which provide strong location priors. Medium/visibility prompts reflect variations in water clarity, ranging from crystal-clear tropical seas to turbid or plankton-rich conditions (e.g., "a {} in low visibility conditions"), thereby modeling visual degradations that frequently occur underwater. Lighting prompts capture distinct illumination conditions including bioluminescence, diver flashlights, or strong sunlight filtering through the water column (e.g., "a {} illuminated by artificial light underwater"), which are crucial for robust representation learning under diverse visual appearances. Depth-related prompts explicitly encode the ecological and physical differences across ocean layers, from shallow reefs to the hadal trenches (e.g., "a {} at mesopelagic depth"), helping the model disambiguate species that are depth-specific. Finally, scene/interaction prompts describe dynamic rela-

tionships such as co-occurrence, interactions with divers or vehicles, and natural behaviors (e.g., "a {} swimming with other fish underwater"), which improve context awareness.

By enriching textual representations with these underwater-specific prompts, our method bridges the semantic gap between terrestrial-pretrained vision–language models and the marine domain. Empirical results in Tab. 5 confirm that the combination of prompt engineering and adaptive template selection consistently improves both overall segmentation accuracy and open-vocabulary generalization, demonstrating the importance of underwater-aware textual priors in guiding vision–language alignment.

I. More Qualitative Results.

We present additional qualitative and visualization results (in Fig. 11 - Fig. 14), where the internal feature visualizations further support the effectiveness of our proposed method. The final segmentation map comparisons demonstrate improved model confidence and enhanced prediction capability.

J. More Per-Class Experiment Results.

We further present the per-class performance in Fig. 15, using category IDs on the x-axis for clearer visualization. We report results for the top-50 best- and worst-performing classes. Consistent with our earlier findings, the In-Domain setting generally outperforms the Cross-Domain setting, highlighting the importance of underwater scene adaptation to improve model performance and suggesting the need for more extensive underwater datasets. Notably, our model achieves superior Cross-Domain performance on certain categories, likely due to the broad coverage of the COCO dataset combined with the strong adaptability of our GPEM and SAIM methods to underwater scenarios.

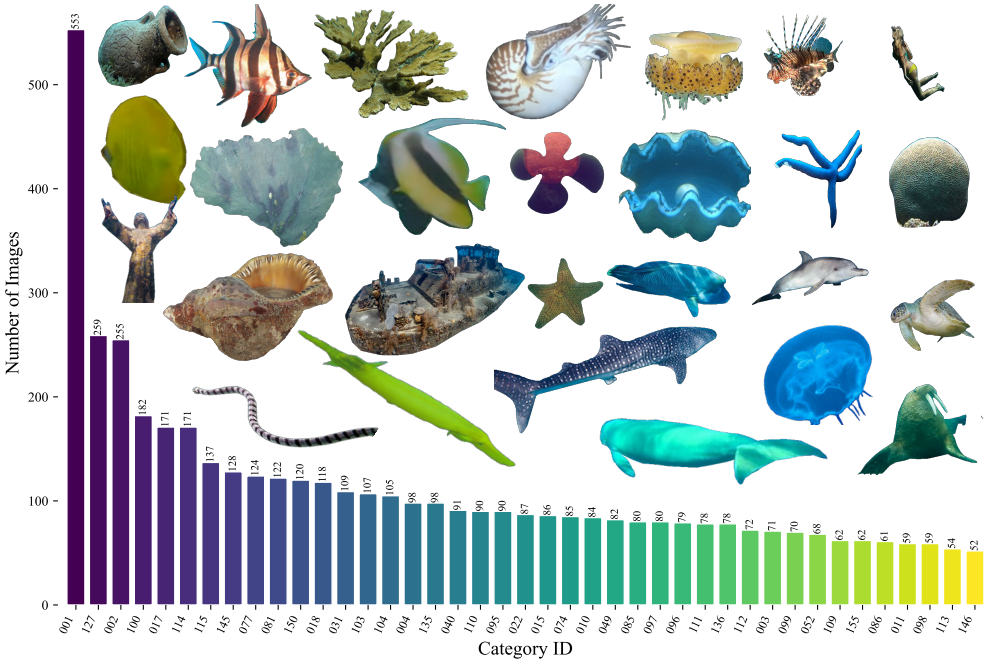

Figure 8. **Instance distribution of Intersection Classes in MARIS validation set.** Shows the number of instances for classes shared between training and validation sets.

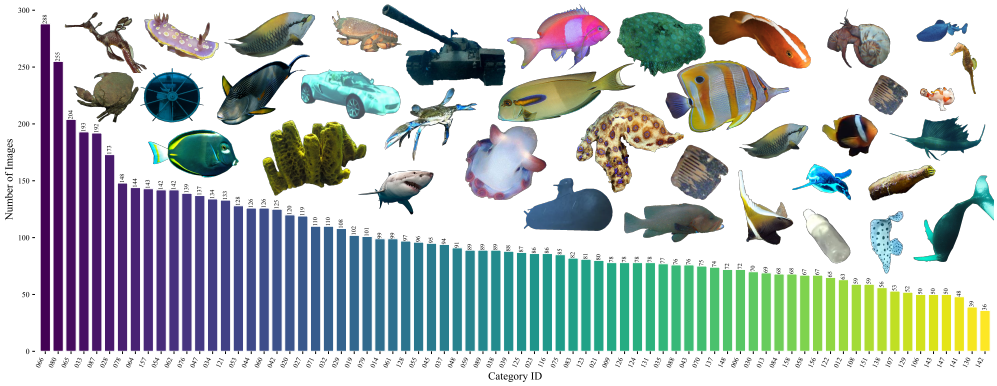

Figure 9. **Instance distribution of Open-Vocabulary (OV) Classes in MARIS validation set.** Shows the number of instances for classes that appear only in validation.

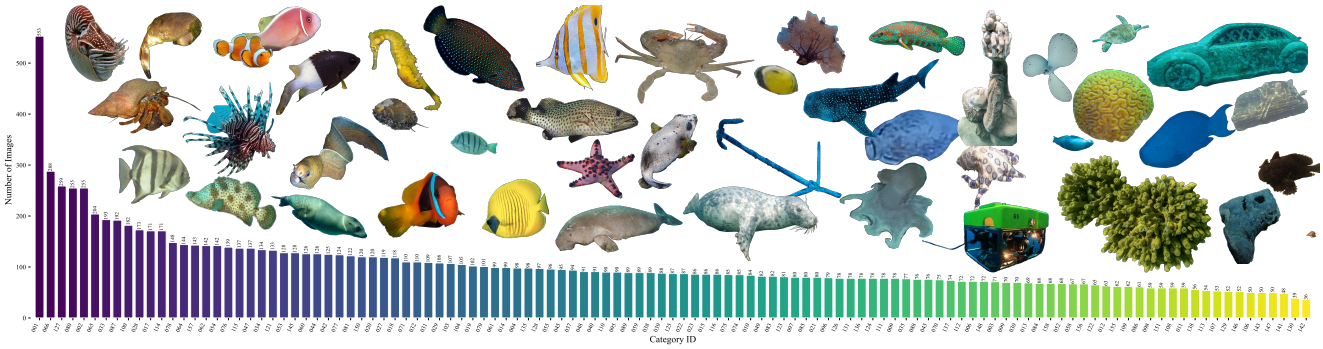

Figure 10. **Instance distribution of Overall Classes in MARIS validation set.** Provides the counts for all classes, giving an overall view of dataset composition and class imbalance.

Parent Category	Child Category (Train)
Human	Diver, Swimmer
Fish	Achilles Tang, Anampses Twistii, Bicolor Angelfish, Blue Parrotfish, Blue-spotted Wrasse, Bluecheek Butterflyfish, Bullhead Shark, Enoplosus Armatus, Giant Wrasse, Graysby, Hammerhead Shark, Lined Surgeonfish, Lionfish, Manta Ray, Mirror Butterflyfish, Mola, Moorish Idol, Moray Eel, Orbicular Batfish, Potato Grouper, Redsea Bannerfish, Regal Blue Tang, Saddle Butterflyfish, Sawfish, Spotted Wrasse, Stoplight Parrotfish, Threadfin Butterflyfish, Trumpetfish, Twinspot Goby, Whale Shark, Whitespotted Surgeonfish
Non fish	Brain Coral, Common Octopus, Common Prawn, Crinoid, Dolphin, Dugong, Elkhorn Coral, Fan Coral, Fried Egg Jellyfish, Geoduck, Giant Clams, Killer Whale, King Crab, Linckia Laevigata, Lion's Mane Jellyfish, Manatee, Mantis Shrimp, Moon Jellyfish, Nautilus, Oreaster Reticulatus, Protoreaster Nodosus, Scallop, Sea Anemone, Sea Cucumber, Sea Lion, Sea Urchin, Snake, Spiny Lobster, Squid, Triton's Trumpet, Turtle, Walrus
Marine Garbage	Can, Plastic Bag, Surgical Mask, Tyre
Wrecked Vehicle	Shipwreck, Wrecked Aircraft
Lost item	Gun, Phone
Archeology	Amphora, Coin, Statue
Underwater equipment	Autonomous Underwater Vehicle (AUV), Personal Submarines, Remotely Operated Vehicle (ROV)
Underwater operation	Over Board Valve, Propeller, Ship's Anode

Table 9. **Category Diversity Analysis for Train dataset.** This table presents a detailed breakdown of parent categories in the dataset, highlighting the diversity of objects in the training set.

Parent Category	Child Category(Only in Train)
Human	
Fish	Achilles Tang, Anampses Twistii, Bicolor Angelfish, Bullhead Shark, Graysby, Lined Surgeonfish, Manta Ray, Mirror Butterflyfish, Mola, Moorish Idol, Orbicular Batfish, Potato Grouper, Regal Blue Tang, Saddle Butterflyfish, Sawfish, Spotted Wrasse, Stoplight Parrotfish, Twin-spot Goby, Whitespotted Surgeonfish
Non fish	Common Octopus, Common Prawn, Crinoid, Killer Whale, King Crab, Lion's Mane Jellyfish, Mantis Shrimp, Scallop, Sea Anemone, Sea Cucumber, Spiny Lobster, Squid
Marine Garbage	Can, Surgical Mask, Tyre
Wrecked Vehicle	
Lost item	Gun, Phone
Archeology	Coin
Underwater equipment	Autonomous Underwater Vehicle (AUV), Personal Submarines
Underwater operation	Over Board Valve, Ship's Anode

Table 10. **Category Diversity Analysis for Class Only in Train dataset.** This table presents a detailed breakdown of parent categories in the dataset, highlighting the diversity of objects in the training set.

Parent Category	Child Category (Intersection)	Child Category (OV)	Child Category (Overall)
Human	Diver, Swimmer		Diver, Swimmer
Fish	Blue Parrotfish, Blue-spotted Wrasse, Bluecheek Butterflyfish, Enoplosus Armatus, Giant Wrasse, Hammerhead Shark, Lionfish, Moray Eel, Redsea Bannerfish, Threadfin Butterflyfish, Trumpetfish, Whale Shark	Anyperodon Leucogrammicus, Atlantic Spadefish, Blackspotted Puffer, Blacktail Butterflyfish, Chromis Dimidiata, Cinnamon Clownfish, Convict Surgeonfish, Copperband Butterflyfish, Coral Hind, Electric Ray, Eritrean Butterflyfish, Fire Goby, Flounder, Frogfish, Great White Shark, Heniochus Varius, Hippocampus, Humpback Grouper, Lunar Fusilier, Maldives Damselfish, Ocellaris Clownfish, Orange Skunk Clownfish, Orange-band Surgeonfish, Peacock Grouper, Pink Anemonefish, Pomacentrus Sulfureus, Porcupinefish, Porkfish, Powder Blue Tang, Pseudanthias Pleurotaenia, Pyramid Butterflyfish, Raccoon Butterflyfish, Red-breasted Wrasse, Redmouth Grouper, Sailfish, Scissortail Sergeant, Sea Dragon, Slingjaw Wrasse, Sohal Surgeonfish, Spotted Drum, Threespot Angelfish, Thresher Shark, Whitecheek Surgeonfish, Yellow Boxfish	Anyperodon Leucogrammicus, Atlantic Spadefish, Blackspotted Puffer, Blacktail Butterflyfish, Blue Parrotfish, Blue-spotted Wrasse, Bluecheek Butterflyfish, Chromis Dimidiata, Cinnamon Clownfish, Convict Surgeonfish, Copperband Butterflyfish, Coral Hind, Electric Ray, Enoplosus Armatus, Eritrean Butterflyfish, Fire Goby, Flounder, Frogfish, Giant Wrasse, Great White Shark, Hammerhead Shark, Heniochus Varius, Hippocampus, Humpback Grouper, Lionfish, Lunar Fusilier, Maldives Damselfish, Moray Eel, Ocellaris Clownfish, Orange Skunk Clownfish, Orange-band Surgeonfish, Peacock Grouper, Pink Anemonefish, Pomacentrus Sulfureus, Porcupinefish, Porkfish, Powder Blue Tang, Pseudanthias Pleurotaenia, Pyramid Butterflyfish, Raccoon Butterflyfish, Red-breasted Wrasse, Redmouth Grouper, Redsea Bannerfish, Sailfish, Scissortail Sergeant, Sea Dragon, Slingjaw Wrasse, Sohal Surgeonfish, Spotted Drum, Threadfin Butterflyfish, Threespot Angelfish, Thresher Shark, Trumpetfish, Whale Shark, Whitecheek Surgeonfish, Yellow Boxfish
Non fish	Brain Coral, Dolphin, Dugong, Elkhorn Coral, Fan Coral, Fried Egg Jellyfish, Geoduck, Giant Clams, Linckia Laevigata, Manatee, Moon Jellyfish, Nautilus, Oreaster Reticulatus, Protoreaster Nodosus, Sea Lion, Sea Urchin, Snake, Triton's Trumpet, Turtle, Walrus	Abalone, Blue-ringed Octopus, Cancer Pagurus, Dumbo Octopus, Hermit Crab, Homarus, Humpback Whale, Penguin, Queen Conch, Sea Slug, Seal, Spanner Crab, Sperm Whale, Sponge, Swimming Crab	Abalone, Blue-ringed Octopus, Brain Coral, Cancer Pagurus, Dolphin, Dugong, Dumbo Octopus, Elkhorn Coral, Fan Coral, Fried Egg Jellyfish, Geoduck, Giant Clams, Hermit Crab, Homarus, Humpback Whale, Linckia Laevigata, Manatee, Moon Jellyfish, Nautilus, Oreaster Reticulatus, Penguin, Protoreaster Nodosus, Queen Conch, Sea Lion, Sea Slug, Sea Urchin, Seal, Snake, Spanner Crab, Sperm Whale, Sponge, Swimming Crab, Triton's Trumpet, Turtle, Walrus
Marine Garbage	Plastic Bag	Glass Bottle, Plastic Bottle, Plastic Box, Plastic Cup	Glass Bottle, Plastic Bag, Plastic Bottle, Plastic Box, Plastic Cup
Wrecked Vehicle	Shipwreck, Wrecked Aircraft	Wrecked Car, Wrecked Tank	Shipwreck, Wrecked Aircraft, Wrecked Car, Wrecked Tank
Lost item		Boots, Glasses, Ring	Boots, Glasses, Ring
Archeology	Amphora, Statue	Anchor, Ship's Wheel	Amphora, Anchor, Ship's Wheel, Statue
Underwater equipment	Remotely Operated Vehicle (ROV)	Military Submarines	Military Submarines, Remotely Operated Vehicle (ROV)
Underwater operation	Propeller	Pipeline's Anode, Sea Chest Grating, Submarine Pipeline	Pipeline's Anode, Propeller, Sea Chest Grating, Submarine Pipeline

Table 11. **Combined Category Diversity for Validation Dataset.** This table integrates Intersection Class, OV Class, Overall Class for each parent category. It provides a comprehensive overview of category coverage and diversity, highlighting both shared and unique classes.

Generic Prompt	Environment / Background
a photo of a {}	a {} underwater
This is a photo of a {}	a {} in the ocean
There is a {} in the underwater scene	a {} in the deep sea
a photo of a {} in {}	a {} near a coral reef
a photo of a small {}	a {} in murky underwater conditions
a photo of a medium {}	a {} in a tropical sea
a photo of a large {}	a {} in a freshwater lake
This is a photo of a small {}	a {} in brackish water
This is a photo of a medium {}	a {} in shallow coastal water
This is a photo of a large {}	a {} in open ocean water

Table 12. Prompt templates for **Generic** and **Environment/Background** categories.

Medium / Visibility	Lighting / Visual
a {} in turbid blue-green water	a {} illuminated by artificial light underwater
a {} in crystal-clear water	a {} glowing in bioluminescent light
a {} in highly murky water	a {} under dim moonlight underwater
a {} in hazy underwater environment	a {} highlighted by a diver's flashlight
a {} in water filled with plankton	a {} glowing faintly in darkness
a {} in low visibility conditions	a {} in high-contrast underwater light
a {} in silted water	a {} in strong sunlight filtering from above
a {} in cloudy water	a {} in shimmering caustics underwater
a {} in algae-rich water	a {} under soft ambient blue light
a {} in dark underwater conditions	a {} in backlit silhouette underwater

Table 13. Prompt templates for **Medium/Visibility** and **Lighting/Visual** categories.

Depth / Distance	Scene / Interaction
a {} at shallow depth near surface	a {} surrounded by bubbles
a {} at mesopelagic depth	a {} swimming with other fish underwater
a {} at bathypelagic depth	a {} near a diver underwater
a {} in the hadal zone trench	a {} next to an underwater vehicle
close-up of the {} underwater	a {} entangled in fishing net underwater
a {} seen from a distance underwater	a {} resting near coral
a {} disappearing into darkness	a {} hiding under rocks
a {} approaching the camera underwater	a {} camouflaged in sand
a {} drifting into the distance	a {} gliding through seaweed
a {} hovering at seabed depth	a {} chasing prey underwater

Table 14. Prompt templates for **Depth/Distance** and **Scene/Interaction** categories.

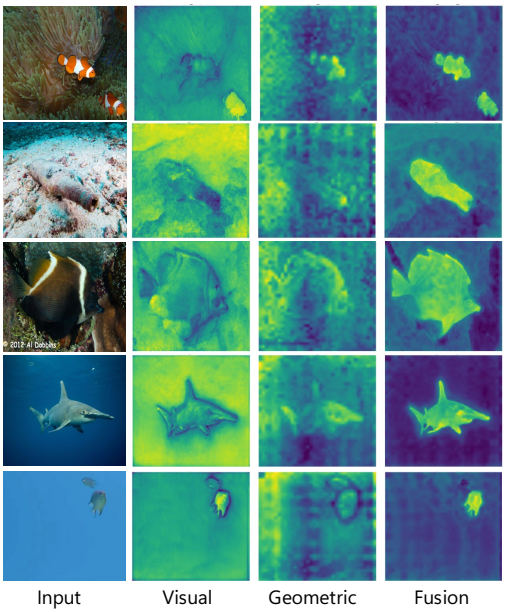

Figure 11. **Additional Qualitative Results on geometric-enhanced fusion features**

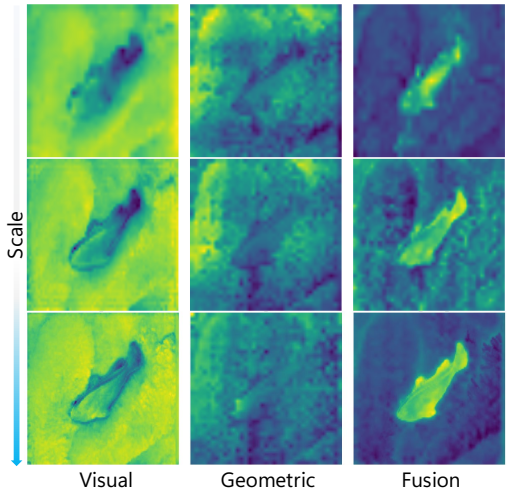

Figure 13. **Additional Qualitative Results on geometric-enhanced fusion features**

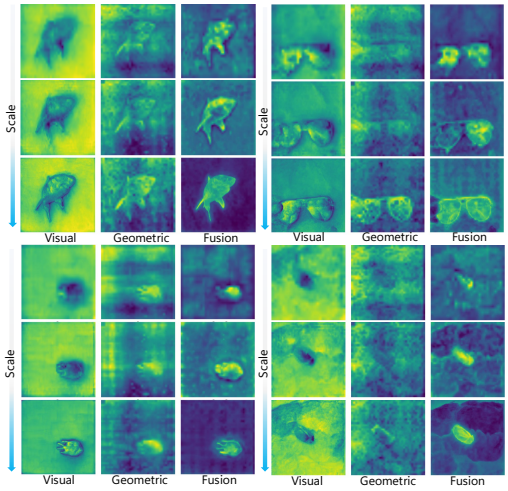

Figure 12. **Additional Qualitative Results on geometric-enhanced fusion features**

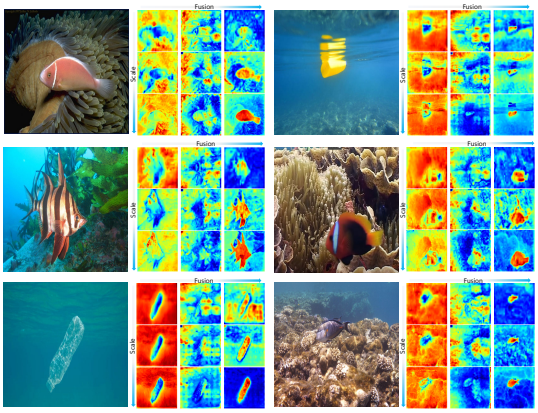

Figure 14. **Additional Qualitative Results on geometric-enhanced fusion features**

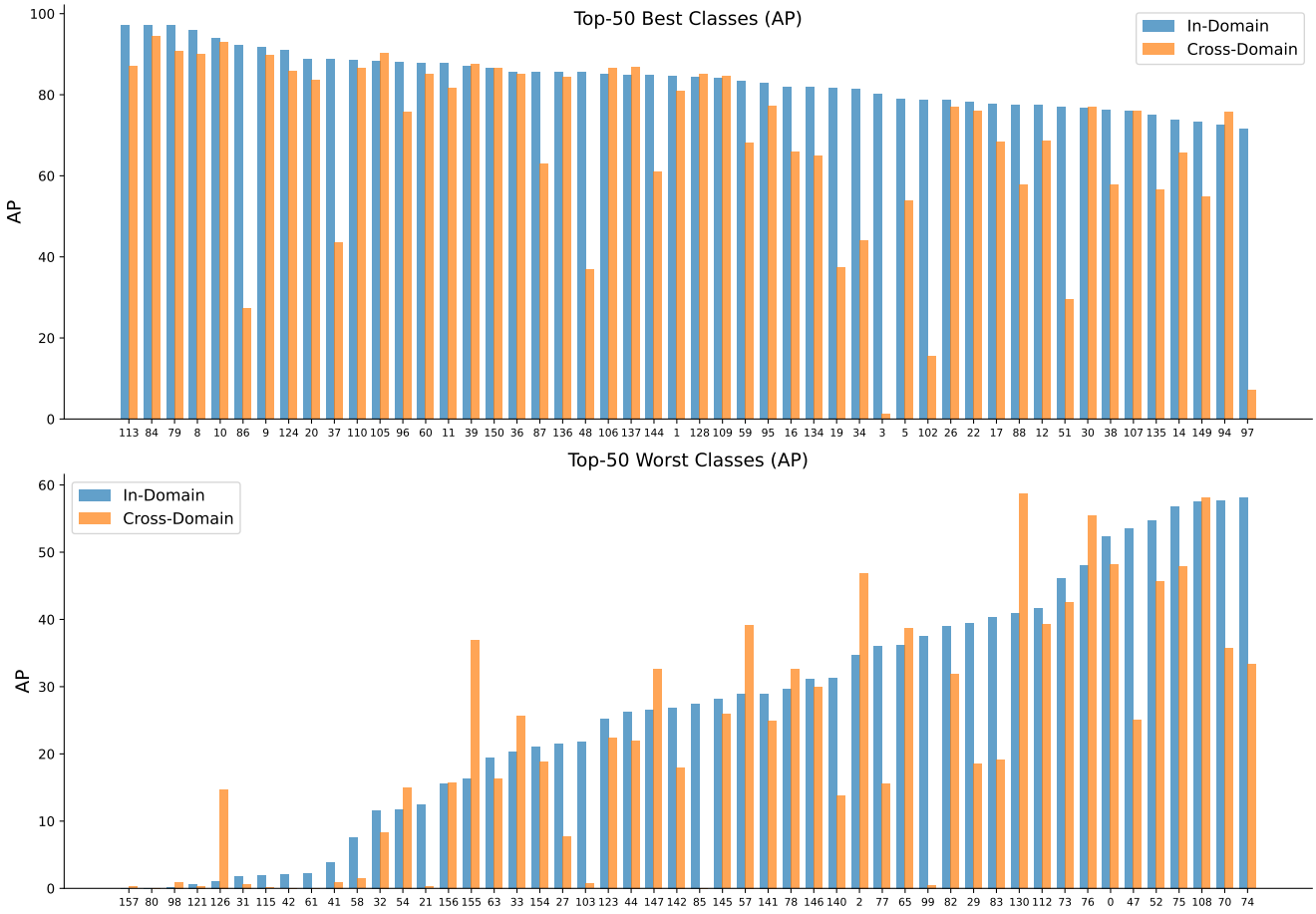

Figure 15. **Per-class performance comparison under In-Domain and Cross-Domain settings.** Shows how domain shifts affect AP across different classes.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grants 62306241 and U62576284.

References

[1] Adnan Abdullah, Titon Barua, Reagan Tibbetts, Zijie Chen, Md Jahidul Islam, and Ioannis Rekleitis. Caveseg: Deep semantic segmentation and scene parsing for autonomous underwater cave exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3781–3788. IEEE, 2024. 1, 2

[2] Lin Chen, Qi Yang, Kun Ding, Zhihao Li, Gang Shen, Fei Li, Qiyuan Cao, and Shiming Xiang. Efficient redundancy reduction for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2501.17642*, 2025. 1, 3

[3] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 2

[4] Seunghyun Cho, Hyunjung Shin, Seunghoon Hong, Anurag Arnab, Paul H. Seo, and Seon Joo Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 1, 3

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 5

[6] Zhengyuan Ding, Jingdong Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation maskclip. *arXiv preprint arXiv:2208.01343*, 2022. 2

[7] Antun DJuravs, Ben J Wolf, Athina Ilioudi, Ivana Palunko, and Bart De Schutter. A dataset for detection and segmentation of underwater marine debris in shallow waters. *Scientific data*, 11(1):921, 2024. 2

[8] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10995–11005, 2023. 3

[9] Songcheng Du, Yang Zou, Zixu Wang, Xingyuan Li, Ying Li, Changjing Shang, and Qiang Shen. Unsupervised hyperspectral image super-resolution via self-supervised modality decoupling. *International Journal of Computer Vision*, 2026. 2

[10] Zhenqi Fu, Ruizhe Chen, Yue Huang, En Cheng, Xinghao Ding, and Kai-Kuang Ma. Masnet: A robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*, 49(3):1104–1115, 2023. 2

[11] Huilin Ge and Jiali Ouyang. Underwater image segmentation via the progressive network of dual iterative complement enhancement. *Expert Systems with Applications*, 266:126049, 2025. 1, 2

[12] ZhiQian He, LiJie Cao, JiaLu Luo, XiaoQing Xu, JiaYi Tang, JianHao Xu, GengYan Xu, and ZiWen Chen. Uissnet: Underwater image semantic segmentation network for improving boundary segmentation accuracy of underwater images. *Aquaculture International*, 32(5):5625–5638, 2024. 2

[13] Lin Hong, Xin Wang, Yihao Li, and Xia Wang. Usis16k: High-quality dataset for underwater salient instance segmentation. *arXiv preprint arXiv:2506.19472*, 2025. 2, 3

[14] Yang Hong, Xiaowei Zhou, Ruzhuang Hua, Qingxuan Lv, and Junyu Dong. Watersam: Adapting sam for underwater object segmentation. *Journal of Marine Science and Engineering*, 12(9):1616, 2024. 1, 2

[15] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1769–1776. IEEE, 2020. 2, 3

[16] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Shi Humphrey. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, 2024. 3, 6

[17] Bowen Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[18] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Fgaseg: Fine-grained pixel-text alignment for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2501.00877*, 2025. 3

[19] Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. U3m: Unbiased multiscale modal fusion model for multimodal semantic segmentation. *Pattern Recognition*, page 111801, 2025. 1

[20] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. 2

[21] Hua Li, Shijie Lian, Zhiyuan Li, Runmin Cong, and Sam Kwong. Uwsam: Segment anything model guided underwater instance segmentation and a large-scale benchmark dataset. *arXiv preprint arXiv:2505.15581*, 2025. 1, 2, 3

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5

[24] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1305–1315, 2023. 2, 3

[25] Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Laurence Tianruo Yang, Sam Kwong, and Runmin Cong. Diving into

underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. *arXiv preprint arXiv:2406.06039*, 2024. 2, 3

[26] Feng Liang, Baitao Wu, Xinyu Dai, Kuan Li, Yue Zhao, Han Zhang, Peng Zhang, Peter Vajda, and Daniel Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 6

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[28] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 3

[29] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026. 2

[30] Weijian Ma, Shizhao Sun, Tianyu Yu, Ruiyu Wang, Tat-Seng Chua, and Jiang Bian. Thinking with blueprints: Assisting vision-language models in spatial reasoning via structured object representation, 2026. 2

[31] Zhiwei Ma, Haojie Li, Zhihui Wang, Dan Yu, Tianyi Wang, Yingshuang Gu, Xin Fan, and Zhongxuan Luo. An underwater image semantic segmentation method focusing on boundaries and a real underwater scene semantic segmentation dataset. *arXiv preprint arXiv:2108.11727*, 2021. 2

[32] Hongwei Niu, Jie Hu, Jianghang Lin, Guannan Jiang, and Shengchuan Zhang. Eov-seg: Efficient open-vocabulary panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6254–6262, 2025. 2, 6

[33] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19446–19455, 2023. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 5

[35] Xudong Shan, Di Wu, Guorong Zhu, Yong Shao, Nong Sang, and Changxin Gao. Open - vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28412–28421, 2024. 3

[36] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Matthew Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *arXiv preprint arXiv:2506.21656*, 2025. 2

[37] Pengfei Shi, Shen Shao, Yueyue Liu, Xinnan Fan, and Yuanxue Xin. Crackinst: a real-time instance segmentation method for underwater dam cracks. *IEEE Transactions on Instrumentation and Measurement*, 2024. 2

[38] Xiu Su, Qinghua Mao, Zhongze Wu, Xi Lin, Shan You, Yue Liao, and Chang Xu. Large language models driven neural architecture search for universal and lightweight disease diagnosis on histopathology slide images. *npj Digital Medicine*, 8(1):682, 2025. 2

[39] Quang Trung Truong, Wong Yuk Kwan, Duc Thanh Nguyen, Binh-Son Hua, and Sai-Kit Yeung. Autv: Creating underwater video datasets with pixel-wise annotations. *arXiv preprint arXiv:2503.12828*, 2025. 2

[40] Zhongze Wu, Hongyan Xu, Yitian Long, Shan You, Xiu Su, Jun Long, Yueyi Luo, and Chang Xu. Detecting any instruction-to-answer interaction relationship: Universal instruction-to-answer navigator for med-vqa. In *Forty-first International Conference on Machine Learning*, 2024. 2

[41] Bo Xie, Jie Cao, Jing Xie, Fahad Shahbaz Khan, and Youtao Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3426–3436, 2024. 3

[42] Yuechen Xie, Jie Song, Huiqiong Wang, and Mingli Song. Training data provenance verification: Did your model use synthetic data from my generative model for training? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23817–23827, 2025. 2

[43] Yuechen Xie, Xiaoyan Zhang, Yicheng Shan, Hao Zhu, Rui Tang, Rong Wei, Mingli Song, Yuanyu Wan, and Jie Song. Spatialqa: A benchmark for evaluating spatial logical reasoning in vision-language models. *arXiv preprint arXiv:2602.20901*, 2026. 2

[44] Feng Xu, Guangyao Zhai, Xin Kong, Tingzhong Fu, Daniel FN Gordon, Xueli An, and Benjamin Busam. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. *arXiv preprint arXiv:2512.05107*, 2025. 2

[45] Jingyi Xu, Shu Liu, Arash Vahdat, Woojin Byeon, Xinyong Wang, and Stefano De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3, 6

[46] Ming Xu, Zhen Zhang, Feng Wei, Yixuan Lin, Yukun Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[47] Ming Xu, Zhen Zhang, Feng Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 3, 6

[48] Wenhan Xu, Chen Wang, Xin Feng, Runze Xu, Lei Huang, Zhen Zhang, Lei Guo, and Shuaicheng Xu. Generalization

boosted adapter for open - vocabulary segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 3

[49] Xizhe Xue, Yang Zhou, Dawei Yan, Ying Li, Haokui Zhang, and Rong Xiao. Uvlm: Benchmarking video language model for underwater world understanding. *arXiv preprint arXiv:2507.02373*, 2025. 2

[50] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 4

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 4

[52] Dewei Yi, Hasan Bayarov Ahmedov, Shouyong Jiang, Yiren Li, Sean Joseph Flinn, and Paul G Fernandes. Coordinate-aware mask r-cnn with group normalization: A underwater marine animal instance segmentation framework. *Neurocomputing*, 583:127488, 2024. 2

[53] Qingyi Yu, Jiahao He, Xin Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 1, 3, 5, 6

[54] Pingrui Zhang, Yifei Su, Pengyuan Wu, Dong An, Li Zhang, Zhigang Wang, Dong Wang, Yan Ding, Bin Zhao, and Xuelong Li. Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation. *arXiv preprint arXiv:2505.20897*, 2025. 2

[55] Ruilin Zhang, Haiyang Zheng, and Hongpeng Wang. Cnmbi: Determining the number of clusters using center pairwise matching and boundary filtering. In *International Conference on Advanced Data Mining and Applications*, pages 262–277. Springer, 2023. 2

[56] Ruilin Zhang, Haiyang Zheng, and Hongpeng Wang. Tdec: Deep embedded image clustering with transformer and distribution information. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 280–288, 2023.

[57] Haiyang Zheng, Ruilin Zhang, and Hongpeng Wang. Deep image clustering based on curriculum learning and density information. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 330–338, 2024. 2

[58] Xin Zuo, Jiaran Jiang, Jifeng Shen, and Wankou Yang. Improving underwater semantic segmentation with underwater image quality attention and muti-scale aggregation attention. *Pattern Analysis and Applications*, 28(2):1–12, 2025. 2