Exploring the Underwater World Segmentation without Extra Training

Supplementary Material

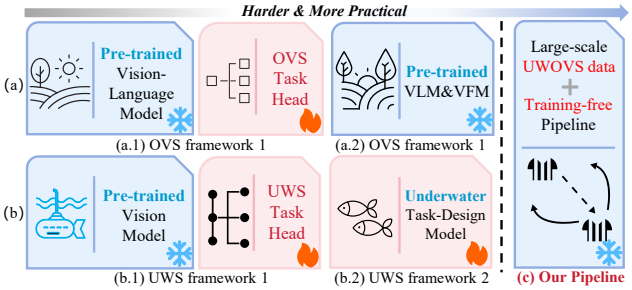

Figure 9. Comparison between our Earth2Ocean framework and existing approaches, highlighting the increased task complexity and practical applicability of our method.

A. Code Availability

The implementation code of our Earth2Ocean framework, including all core modules (Geometric-guided Visual Mask Generator and Category-visual Semantic Alignment) and experimental scripts, **is provided in the appendix**. This includes preprocessing pipelines, model configuration files, and inference demos to facilitate full reproducibility of our results.

B. Distinctive Framework of Earth2Ocean

Earth2Ocean adopts a more comprehensive and challenging task with practical implications, as illustrated in Figure 9. The framework aims to bridge the gap between terrestrial and underwater scenarios, offering a robust solution to transfer learning across domains. This approach not only tackles the inherent challenges of underwater environments but also facilitates the development of training-free frameworks, enabling efficient adaptation to aquatic contexts.

C. Numerical Analysis of AquaOV255

In this section, we provide a more detailed numerical analysis of the **AquaOV255** dataset, focusing on various aspects such as category, quantity, area, and brightness (see Figure 10). Panels (a.1–a.6) present the basic dataset analysis, including the distribution of image quantities across categories, as well as the area and brightness characteristics. Additionally, Panel (b) offers a fine-grained analysis: (b.1) shows split based on biological attributes, while (b.2) categorizes the species according to their commonness. These analyses offer deeper insights into the dataset's structure and diversity, further supporting the methodological choices made in our study.

D. Long Tail Analysis of AquaOV255

As shown in Fig. 11, the dataset exhibits a highly imbalanced class distribution, where a few dominant categories contain the majority of samples, while numerous rare classes have only limited instances. This imbalance poses challenges for feature work to learn robust representations.

E. AquaOV255 Category Taxonomy

The dataset comprises **254 unique underwater object categories**, as shown in Tab. 7, serving as a comprehensive resource for complex aquatic detection and recognition tasks.

E.1. Split Scheme I: Based on Biological and Object Type

According to object type, we propose the first categorization scheme (see Tab. 8). Specifically, the biological component of the ecosystem is dominated by **Fish** (154 classes, approximately 60.6%) and **Invertebrates** (48 classes), reflecting a strong emphasis on fine-grained species identification. The inclusion of **Artificial Objects** (32 classes)—covering marine debris (e.g., *PlasticBag*, *Tyre*) and underwater infrastructure (e.g., *AUV*, *Pipeline*)—further demonstrates the dataset's relevance to key application domains such as environmental monitoring and underwater robotics.

E.2. Split Scheme II: Based on Object Frequency (Commonality)

The second split scheme (see Tab. 9) categorizes objects according to their occurrence frequency or detectability into **Common** (47 classes), **General** (68 classes), and **Special** (139 classes) groups. With the **Special** category constituting the majority (approximately 54.7%), this taxonomy is deliberately designed to support research on **long-tailed recognition** and model robustness under data sparsity or challenging visual conditions.

E.3. Clarification on Grouped mIoU Metric

To analyze performance within semantically related categories, we report *Grouped mIoU* (e.g., **Fish mIoU**), computed as the arithmetic mean of per-class IoU scores for all fine-grained classes belonging to a macro-category. Specifically, IoU is first obtained for each of the 255 fine-grained classes, and the group-level value is calculated as:

$$\text{Fish mIoU} = \frac{1}{N_{\text{fish}}} \sum_{i \in \text{Group}_{\text{Fish}}} \text{IoU}_i, \tag{12}$$

where N_{fish} denotes the number of classes in the fish group. Importantly, intra-group misclassifications (e.g.,

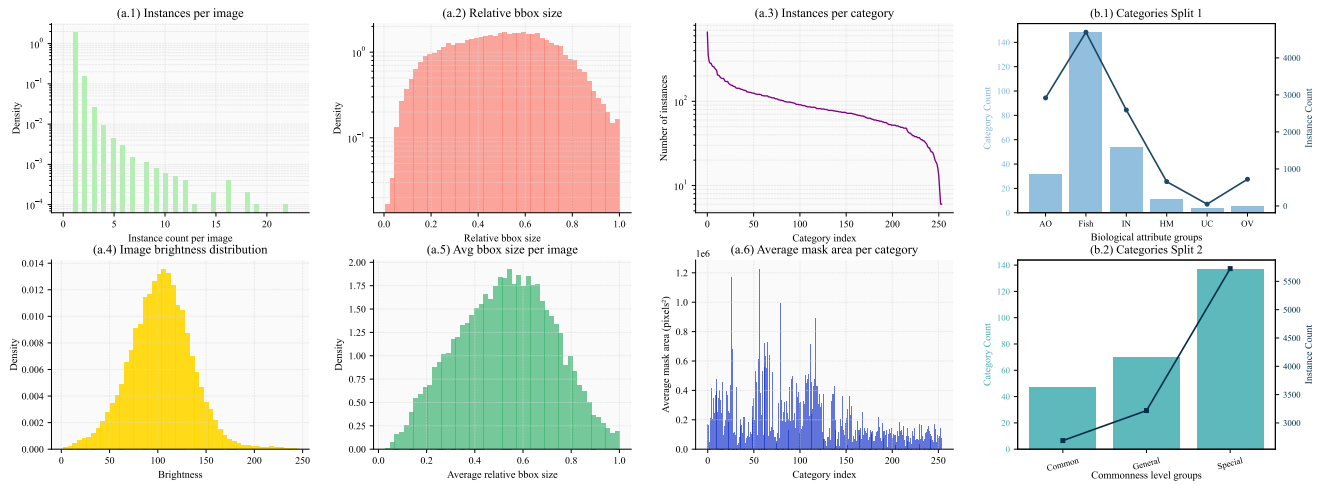

Figure 10. (a.1–a.6) Dataset analysis in terms of quantity, category, area, and brightness; and (b) fine-grained dataset analysis, where (b.1) shows split based on biological attributes and (b.2) shows split based on species commonness.

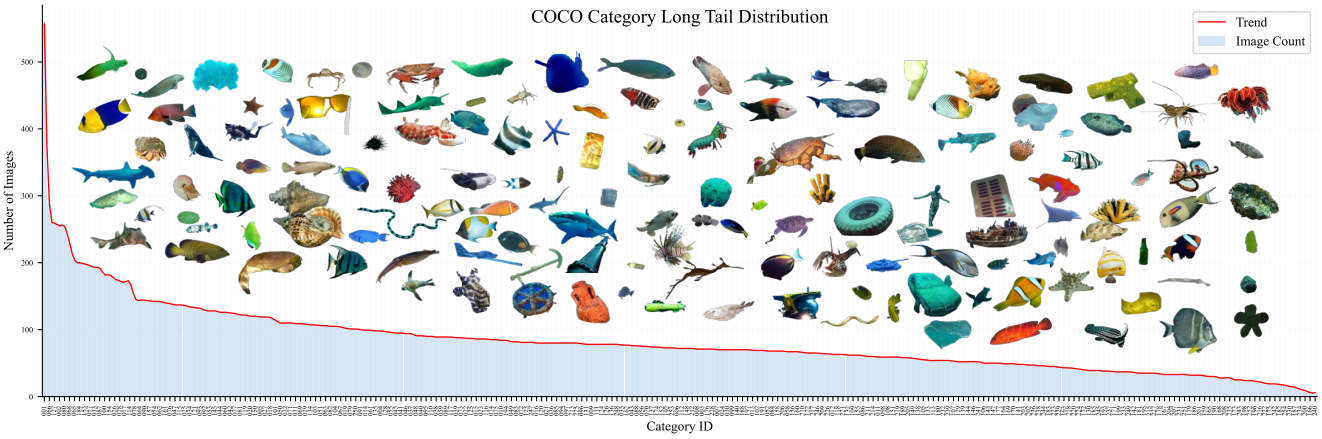

Figure 11. Long-tail distribution analysis of the AquaOV255 dataset.

predicting "clownfish" as "butterflyfish") remain penalized, preserving the fine-grained nature of the 255-class evaluation.

Unlike *merged mIoU*, which treats intra-group confusions as correct, our grouped formulation serves as a diagnostic measure of fine-grained discrimination within a semantic subset. It highlights relative task difficulty (e.g., distinguishing fish species vs. coral types) and maintains class-equal fairness by giving rare and common classes equal weight. This ensures that the metric faithfully reflects model robustness across both frequent and rare categories within each semantic group.

F. Visualization Validation of GMG and CSA

F.1. Effectiveness of GMG in Background Differentiation

To further validate the contribution of the GMG module, we visualize the segmentation results obtained from different methods, as shown in Fig. 12. Unlike conventional approaches that often struggle to separate objects from visually similar underwater backgrounds, our model achieves clearer boundaries and more consistent object localization. This improvement demonstrates that GMG effectively mitigates the ambiguity caused by underwater lighting variations, scattering, and background clutter, leading to more robust and accurate segmentation performance.

F.2. Effectiveness of CSA in Rare Underwater Organisms Pixel-level Classification

To assess the capability of the proposed CSA module in handling rare underwater categories, we conduct qualitative analyses focusing on pixel-level classification. As shown inFig. 12, our model exhibits stronger discrimination between rare object classes. The CSA module effectively aligns MLLM semantic cues, ensuring that both semantic information contribute to precise pixel-level predictions.

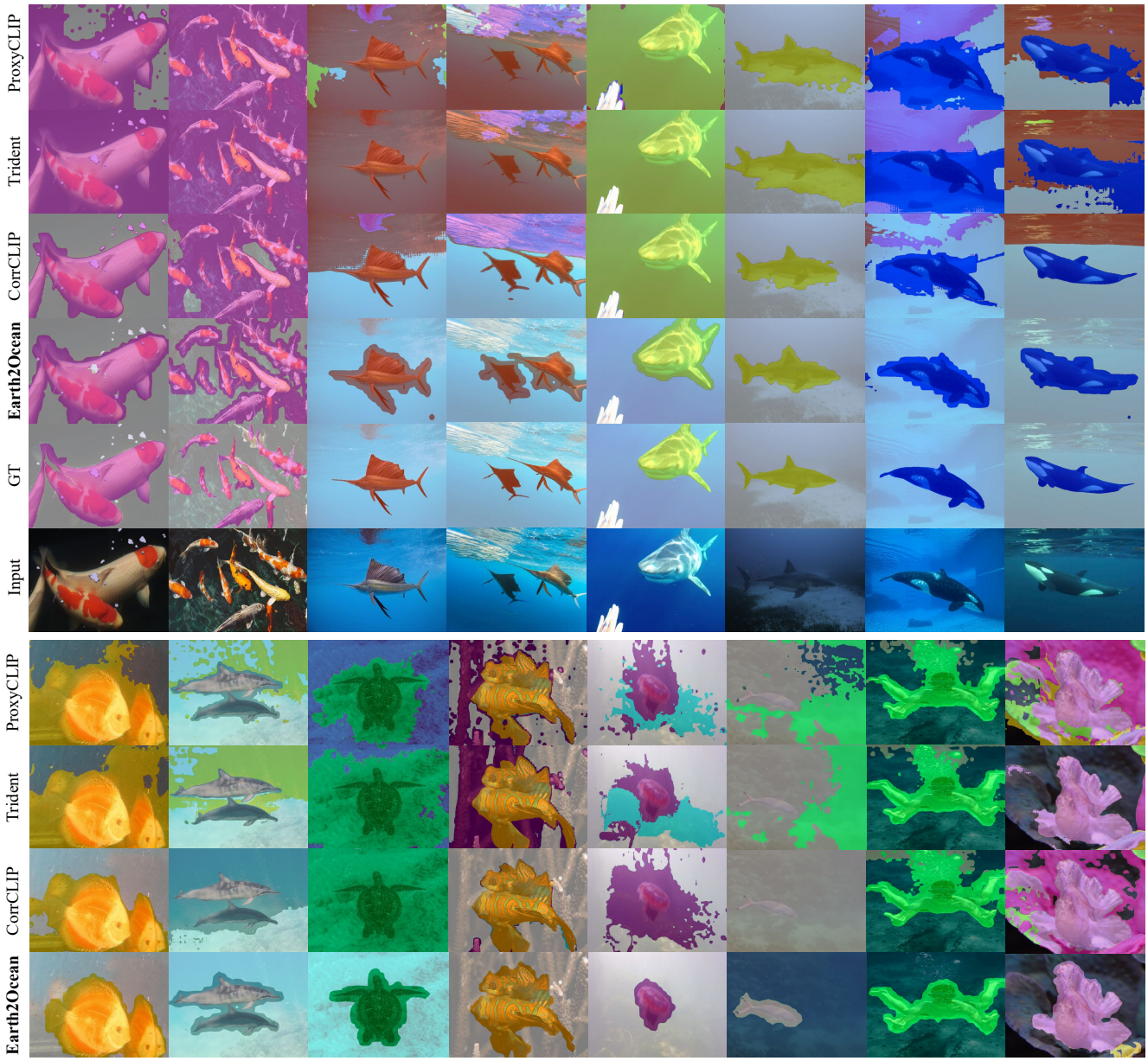

Figure 12. Visualization of segmentation results compared with other methods. Our model demonstrates superior capability in distinguishing background regions, particularly in underwater scenes, where the enhanced visual separation between objects and the background highlights the effectiveness of the GMG module. The background visualized as white

G. Reasoning Prompt for MLLM

We present the prompt design used for multimodal reasoning in our large multimodal model in Fig. 13.

H. Examples of multimodal reasoning outputs

We show some examples of multimodal reasoning outputs in Fig. 14.

I. Evaluation Metrics

For evaluating the semantic segmentation performance, three key metrics are adopted: Overall Pixel Accuracy (aAcc), Mean Intersection over Union (mIoU), and Mean Pixel Accuracy (mAcc). Their formulas are defined as follows:

Overall Pixel Accuracy (aAcc) measures the proportion

Please describe the content of the following image in English.
Your response must include:
(1) a concise English caption of the image;
(2) a list of objects present in the image (only from the given category list);
(3) for each object, describe its attributes (e.g., color, shape, material, size).
The output must be a valid JSON in this format:

```
{
  "Captioning": "A cozy living room...",
  "Objects": ["television", "cabinet", ...],
  "Attributes": {"television": ["black", "flat-screen"]}
}
```

Category list: {category_1, category_2, ...}

Figure 13. Reasoning Prompt for MLLM

of correctly classified pixels relative to all pixels:

$$\text{aAcc} = \frac{\sum_{i=0}^{K-1} TP_i}{\sum_{i=0}^{K-1} TP_i + FP_i + FN_i} \tag{13}$$

Mean Intersection over Union (mIoU) averages the intersection-over-union across all K classes, where IoU for class i is the ratio of overlapping pixels (intersection) to the total pixels in either the prediction or ground truth (union):

$$\text{mIoU} = \frac{1}{K} \sum_{i=0}^{K-1} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{14}$$

Mean Pixel Accuracy (mAcc) calculates the average of per-class accuracy, where per-class accuracy for class i is the ratio of correctly classified pixels of class i to the total pixels belonging to class i in the ground truth:

$$\text{mAcc} = \frac{1}{K} \sum_{i=0}^{K-1} \frac{TP_i}{TP_i + FN_i} \tag{15}$$

In these formulas, K denotes the number of classes; TP_i, FP_i, and FN_i represent true positives, false positives, and false negatives for class i, respectively.

J. Experimental Reproduction Details

This appendix provides comprehensive reproduction details for all evaluated models on the proposed **UOVSBench**, ensuring reproducibility and transparency of our results. All models follow a consistent *training-free paradigm*.

J.1. Common Experimental Setup

All experiments adhere to unified configurations to eliminate environmental biases. We employ **OpenCLIP** (ViT-B/16, ViT-L/14, ViT-H/14) pretrained on LAION-2B [39] as the base Vision-Language Model (VLM), initialized with official weights. The text prompt follows the standard ImageNet-style template: *"a photo of a* {*class_name*}*."* All experiments are implemented using MMSegmentation [9] with PyTorch 2.0 and conducted on NVIDIA RTX 4090 GPUs under FP16 precision for efficiency.

J.2. Model-Specific Reproduction Details

SCLIP. SCLIP replaces the last self-attention block of OpenCLIP's vision encoder with Correlative Self-Attention (CSA), which jointly applies query-query (**qq**) and key-key (**kk**) attention to enhance spatial covariance [45]. All other layers remain frozen during inference, ensuring full reproducibility without additional fine-tuning.

ClearCLIP. ClearCLIP modifies the final transformer layer of OpenCLIP to reduce segmentation noise [24]. Specifically, it removes the residual connection, employs **qq** self-attention as the primary attention mechanism, and discards the feed-forward network (FFN) to prevent feature distortion. The implementation follows the official configuration without further hyperparameter tuning.

ProxyCLIP. ProxyCLIP introduces a proxy attention mechanism using DINO ViT-B/8 [8] as a Vision Foundation Model (VFM) for improved spatial consistency due to its smaller patch size. DINO features serve as proxy attention with adaptive normalization and masking ($\beta = 1.2$, $\gamma = 3.0$) [25]. To align the feature space, CLIP's visual embeddings are interpolated to match DINO's output resolution. Both OpenCLIP and DINO backbones remain frozen throughout inference.

Trident. Trident adopts a *Splice-then-Segment* paradigm to handle high-resolution inputs efficiently [42]. It integrates three complementary models: OpenCLIP ViT-H/14 for semantic reasoning, DINO ViT-B/8 for sub-image spatial guidance, and SAM ViT-B/16 for global correlation modeling. The SAM refinement module is activated via the `--sam_refinement` flag and utilizes mask, point, and box prompts with a scaling factor $\alpha = 0.005$. The affinity matrix combines SAM's cosine similarity and attention weights for enhanced segmentation accuracy.

CorrCLIP. CorrCLIP reconstructs patch-level correlations through a two-stage process [54]. The scope reconstruction employs SAM2 with a Hiera-L backbone [40] and DBSCAN clustering (radius = 0.2, min_samples = 1) to generate coherent region masks. The value reconstruction step leverages DINO ViT-B/8's query and key embeddings ($\tau = 0.25$) to restore fine-grained similarity patterns. The final representation fuses a spatial branch ($\alpha = 1$) and a

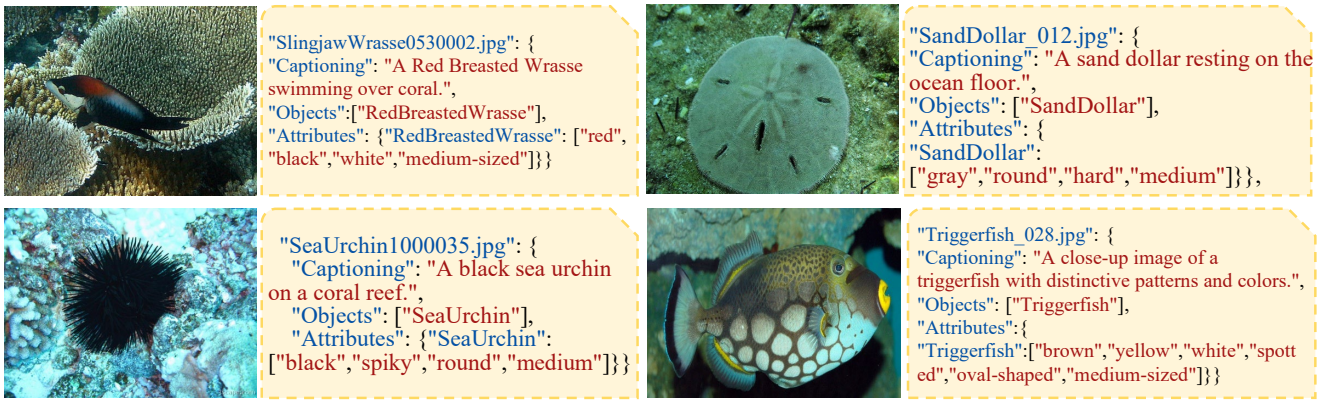

Figure 14. Examples of multimodal reasoning outputs generated by our large multimodal model (MLLM).

semantic branch ($\beta = 0.5$), followed by mode-based label correction for spatial consistency. All parameters follow the optimal configuration reported in [54].

K. Efficient MLLM Semantic Extraction Design

To enhance inference efficiency, we adopt a strategy combining offline MLLM feature extraction and CLIP encoding: we first use GPT-4o and Qwen2.5VL-3B/7B to extract semantic information (e.g., shape, color, habitat) for each category, which is stored in JSON format. During the inference initialization phase, these semantic details are encoded into fixed-dimensional text embeddings via the CLIP text encoder and cached. This approach eliminates direct MLLM calls during runtime, significantly accelerating inference speed.

L. Impact of MLLMs on Earth2Ocean

Table 10 reports the performance of Earth2Ocean across different vision backbones (ViT-B/16, ViT-L/14, ViT-H/14) and multimodal large language models (MLLMs) for inference. Overall, GPT-4o consistently outperforms the Qwen variants in terms of average mIoU and mAcc, highlighting its stronger multimodal understanding and alignment capabilities. Within the Qwen models, the larger 7B version shows modest improvements over the 3B version, suggesting that model scale contributes to performance but is less influential than pretraining quality. These results demonstrate that the choice of MLLM for inference significantly affects Earth2Ocean's segmentation accuracy and generalization.

M. Underwater Image Description Templates

The following templates are designed to generate descriptive prompts for underwater images. Each template provides different perspectives, scene settings, lighting variations, dynamic interactions, and appearance traits for various underwater scenarios. The descriptions can be used for tasks like data annotation or image generation in underwater research.

M.1. Basic Visual Description

- A photo of a `class` underwater.
- An underwater photo of a `class`.
- A close-up photo of a `class` underwater.
- A side view of a `class` underwater.
- A top-down view of a `class` underwater.
- A clear underwater view of a `class`.
- An underwater snapshot of a `class`.
- A natural underwater photo of a `class`.
- A detailed underwater picture of a `class`.
- An underwater macro photo of a `class`.

M.2. Scene Semantics

- A `class` swimming in the ocean.
- A `class` resting on the seabed.
- A `class` near a coral reef.
- A `class` among rocks underwater.
- A `class` surrounded by marine plants.
- A `class` gliding through the sea.
- A `class` moving in shallow water.
- A `class` in the deep ocean.
- A `class` floating near the surface.
- A `class` hiding in coral structures.
- A `class` exploring the ocean floor.
- A `class` captured in a marine ecosystem.
- A `class` near underwater vegetation.
- A `class` surrounded by small fish.
- A `class` swimming close to a diver.

M.3. Lighting and Imaging Variations

- A `class` in turbid underwater conditions.

- A `class` in clear blue water.
- A `class` in greenish water with particles.
- A `class` in low-light underwater conditions.
- A `class` illuminated by sunlight through the water.
- A `class` in bright tropical water.
- A `class` under weak underwater lighting.
- A `class` in dark deep-sea conditions.
- A `class` seen through murky water.
- A `class` under artificial underwater lighting.
- A `class` glowing under bioluminescent light.
- A `class` in a color-distorted underwater image.
- A `class` with reflections on its body underwater.
- A `class` in a hazy underwater view.
- A `class` captured with a waterproof camera.
- A `class` viewed through air bubbles.
- A `class` affected by light scattering underwater.
- A `class` partially blurred by motion underwater.
- A `class` in a high-contrast underwater shot.
- A `class` captured in a long-exposure underwater photo.

M.4. Interaction and Dynamic Scenes

- A `class` interacting with coral.
- A `class` chasing small fish.
- A `class` near a rock formation.
- A `class` partially hidden behind seaweed.
- A `class` resting under coral branches.
- A `class` swimming with other sea creatures.
- A `class` near bubbles and particles.
- A `class` hunting underwater.
- A `class` feeding near the seabed.
- A `class` hiding inside a reef cave.
- A `class` floating above sand.
- A `class` following the water current.
- A `class` playing with another `class`.
- A `class` entangled in marine plants.
- A `class` moving across a coral ridge.
- A `class` resting quietly underwater.
- A `class` escaping a predator.
- A `class` in a calm underwater scene.
- A `class` captured during motion underwater.
- A `class` facing the camera underwater.

M.5. Appearance and Scale Diversity

- A small `class` underwater.
- A large `class` underwater.
- A distant view of a `class` underwater.
- A close view of a `class` underwater.
- A group of `class` underwater.
- A single `class` underwater.
- A colorful `class` underwater.
- A pale `class` in dim water.
- A `class` with a patterned texture underwater.
- A `class` covered in sand underwater.
- A transparent `class` underwater.
- A `class` with vivid stripes underwater.
- A metallic-looking `class` underwater.
- A camouflaged `class` underwater.
- A shadowy silhouette of a `class` underwater.
- A partially visible `class` underwater.
- A detailed close-up of the `class` skin underwater.
- A `class` with motion blur underwater.
- A glowing `class` underwater.
- A dark-colored `class` underwater.

M.6. Environmental and Background Variations

- A `class` near underwater rocks.
- A `class` above sandy seabed.
- A `class` in a coral garden.
- A `class` near a sunken ship.
- A `class` swimming in open sea.
- A `class` near a deep trench.
- A `class` in a lagoon.
- A `class` in shallow tropical water.
- A `class` near underwater volcanic vents.
- A `class` surrounded by bubbles.
- A `class` next to an underwater cave.
- A `class` near marine debris.
- A `class` in a rocky underwater canyon.
- A `class` among sea sponges.
- A `class` swimming through kelp.

Table 7. ID Name Mapping

ID	Name	ID	Name	ID	Name	ID	Name
1	Diver	64	OrangeBandSurgeonfish	127	PlasticBag	190	Fangtooth
2	Swimmer	65	ConvictSurgeonfish	128	PlasticBottle	191	Filefish
3	Geoduck	66	SohalSurgeonfish	129	PlasticCup	192	Flamingotonguesnail
4	LinckiaLaevigata	67	RegalBlueTang	130	PlasticBox	193	FlashlightFish
5	MantaRay	68	LinedSurgeonfish	131	GlassBottle	194	Flatworm
6	ElectricRay	69	AchillesTang	132	Mask	195	FrilledShark
7	Sawfish	70	PowderBlueTang	133	Tyre	196	GardenEel
8	BullheadShark	71	WhitecheekSurgeonfish	134	Can	197	GiantGourami
9	GreatWhiteShark	72	SaddleButterflyfish	135	Shipwreck	198	Goblinshark
10	WhaleShark	73	MirrorButterflyfish	136	WreckedAircraft	199	Goldfish
11	HammerheadShark	74	BluecheekButterflyfish	137	WreckedCar	200	GrassCarp
12	ThresherShark	75	BlacktailButterflyfish	138	WreckedTank	201	Grayling
113	WeedySeaDragon	76	RaccoonButterflyfish	139	Gun	202	Guppy
14	Hippocampus	77	ThreadfinButterflyfish	140	Phone	203	HorseshoeCrab
15	MorayEel	78	EritreanButterflyfish	141	Ring	204	Killifish
16	OrbicularBatfish	79	PyramidButterflyfish	142	Boots	205	Koi
17	Lionfish	80	CopperbandButterflyfish	143	Glasses	206	KuhliLoach
18	Trumpetfish	81	GiantClams	144	Coin	207	Lanternfish
19	Flounder	82	Scallop	145	Statue	208	LargemouthBass
20	Frogfish	83	Abalone	146	Amphora	209	LeafScorpionfish
21	Sailfish	84	QueenConch	147	Anchor	210	Leafyseadragon
22	EnoplosusArmatus	85	Nautilus	148	ShipsWheel	211	MandarinFish
23	PseudanthiasPleurotaenia	86	TritonsTrumpet	149	AUV	212	MarineIguana
24	Mola	87	SeaSlug	150	ROV	213	MimicOctopus
25	MoorishIdol	88	DumboOctopus	151	MilitarySubmarines	214	Mudskipper
26	BicolorAngelfish	89	BlueRingedOctopus	152	PersonalSubmarines	215	NeonTetra
27	AtlanticSpadefish	90	CommonOctopus	153	ShipsAnode	216	Oarfish
28	SpottedDrum	91	Squid	154	OverBoardValve	217	OscarFish
29	ThreespotAngelfish	92	Cuttlefish	155	Propeller	218	Paddlefish
30	ChromisDimidiata	93	SeaAnemone	156	SeaChestGrating	219	PearlGourami
31	RedseaBannerfish	94	LionsManeJellyfish	157	SubmarinePipeline	220	Perch
32	HeniochusVarius	95	MoonJellyfish	158	PipelinesAnode	221	Pike
33	MaldivesDamselfish	96	FriedEggJellyfish	159	AlligatorGar	222	PilotFish
34	ScissortailSergeant	97	FanCoral	160	Archerfish	223	PineconeFish
35	FireGoby	98	ElkhornCoral	161	Arowana	224	PomPomCrab
36	TwinSpotGoby	99	BrainCoral	162	BanggaiCardinalfish	225	PomacanthusFish
37	Porcupinefish	100	SeaUrchin	163	BarreleyeFish	226	PygmySeahorse
38	YellowBoxfish	101	SeaCucumber	164	BaskingShark	227	Remora
39	BlackspottedPuffer	102	Crinoid	165	BigheadCarp	228	RibbonEel
40	BlueParrotfish	103	OreasterReticulatus	166	BlackCarp	229	RosyBarb
41	StoplightParrotfish	104	ProtoreasterNodosus	167	BlanketOctopus	230	Salmon
42	PomacentrusSulfureus	105	KillerWhale	168	Bluegill	231	SandDollar
43	LunarFusilier	106	SpermWhale	169	BubbleCoral	232	SeaAngel
44	OcellarisClownfish	107	HumpbackWhale	170	Burbot	233	SeaApple
45	CinnamonClownfish	108	Seal	171	CarpSucker	234	SeaPig
46	RedSeaClownfish	109	Manatee	172	Catfish	235	SeaSpider
47	PinkAnemonefish	110	SeaLion	173	Chimaera	236	SeaSquirt
48	OrangeSkunkClownfish	111	Dolphin	174	ChristmasTreeWorm	237	SilverCarp
49	GiantWrasse	112	Walrus	175	CleanerShrimp	238	SmallmouthBass
50	SpottedWrasse	113	Dugong	176	ClownLoach	239	SnakeheadFish
51	AnampsesTwistii	114	Turtle	177	CoconutCrab	240	SnowCrab
52	BlueSpottedWrasse	115	Snake	178	CommonCarp	241	SpanishDancerNudibranch
53	SlingjawWrasse	116	Homarus	179	ConeSnail	242	SpiderCrab
54	RedBreastedWrasse	117	SpinyLobster	180	ConvictCichlid	243	SpottedGar
55	PeacockGrouper	118	CommonPrawn	181	Copepod	244	Sturgeon
56	PotatoGrouper	119	MantisShrimp	182	CoralShrimp	245	Swordtail
57	Graysby	120	KingCrab	183	Crappie	246	TigerBarb
58	RedmouthGrouper	121	HermitCrab	184	Crocodile&Alligator	247	Tilapia
59	HumpbackGrouper	122	CancerPagurus	185	CrucianCarp	248	Triggerfish
60	CoralHind	123	SwimmingCrab	186	CushionStar	249	TripodSpiderfish
61	Porkfish	124	SpannerCrab	187	DeepSeaHatchetfish	250	Trout
62	AnyperodonLeucogrammicus	125	Penguin	188	DiscusFish	251	VelvetBellyLanternshark
63	WhitespottedSurgeonfish	126	Sponge	189	Fangblenny	252	WeatherLoach
						253	Wobbegong
						254	Zebrafish

Table 8. Categories Split 1

Category Name	Count	ID
ArtificialObjects	32	127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158
Fish	154	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 159, 160, 161, 162, 163, 164, 165, 166, 168, 170, 171, 172, 173, 176, 178, 180, 183, 185, 187, 188, 189, 190, 191, 193, 195, 196, 197, 198, 199, 200, 201, 202, 204, 205, 206, 207, 208, 209, 210, 211, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 225, 226, 227, 228, 229, 230, 237, 238, 239, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254
Invertebrates	48	3, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 100, 101, 102, 103, 104, 116, 117, 118, 119, 120, 121, 122, 123, 124, 126, 167, 174, 175, 177, 179, 181, 182, 186, 192, 194, 203, 213, 224, 231, 232, 233, 234, 235, 236, 240, 241, 242
Humans&LargeMammals	11	1, 2, 105, 106, 107, 108, 109, 110, 111, 112, 113
UnderwaterPlants&Corals	4	97, 98, 99, 169
OtherVertebrates	5	114, 115, 125, 184, 212

Table 9. Categories Split 2 (Commonality-based)

Category Name	Count	ID
Common	47	1, 2, 9, 10, 15, 17, 25, 37, 44, 45, 46, 47, 48, 67, 76, 81, 88, 90, 93, 95, 105, 107, 108, 110, 111, 114, 116, 117, 120, 121, 122, 126, 127, 128, 129, 130, 131, 132, 133, 134, 144, 150, 178, 199, 205, 230, 247
General	68	3, 5, 6, 8, 16, 18, 19, 21, 22, 26, 27, 29, 31, 34, 40, 41, 49, 50, 55, 56, 57, 61, 63, 65, 68, 72, 77, 82, 83, 84, 91, 92, 97, 98, 99, 100, 101, 103, 104, 106, 109, 112, 113, 118, 123, 125, 135, 139, 147, 148, 149, 151, 152, 153, 154, 155, 156, 157, 158, 172, 184, 203, 208, 220, 221, 240, 242, 244, 248, 250
Special	139	4, 7, 11, 12, 13, 14, 20, 23, 24, 28, 30, 32, 33, 35, 36, 38, 39, 42, 43, 51, 52, 53, 54, 58, 59, 60, 62, 64, 66, 69, 70, 71, 73, 74, 75, 78, 79, 80, 85, 86, 87, 89, 94, 96, 102, 115, 119, 124, 136, 137, 138, 140, 141, 142, 143, 145, 146, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 173, 174, 175, 176, 177, 179, 180, 181, 182, 183, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 200, 201, 202, 204, 206, 207, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 222, 223, 224, 225, 226, 227, 228, 229, 231, 232, 233, 234, 235, 236, 237, 238, 239, 241, 243, 245, 246, 249, 251, 252, 253, 254

Table 10. Performance of different Earth2Ocean variants across multiple datasets. The average values are highlighted.

	Method	DUT-Seg			MAS3K			SUIM			USIS10K			USIS16K			AquaOV255			Average		
		aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc
ViT-B/16	Earth2Ocean(GPT-4o)	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**
	Earth2Ocean(qwen2.5VL-3B)	43.33	28.57	49.97	48.61	27.27	41.12	73.28	51.53	71.53	71.6	44.63	59.44	43.68	27.20	41.25	30.50	16.65	26.46	51.83	32.64	48.30
	Earth2Ocean(qwen2.5VL-7B)	45.82	29.69	40.64	50.49	29.16	43.87	73.02	50.98	71.79	71.93	45.14	59.64	44.43	27.97	42.03	31.19	17.13	26.93	52.81	33.35	47.48
ViT-L/14	Earth2Ocean(GPT-4o)	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**
	Earth2Ocean(qwen2.5VL-3B)	66.72	40.42	61.51	60.67	37.98	58.88	74.40	53.35	73.41	71.29	46.23	60.28	51.20	36.44	49.84	42.21	26.37	38.56	61.08	40.13	57.08
	Earth2Ocean(qwen2.5VL-7B)	66.80	40.33	61.38	61.06	38.57	61.73	73.36	52.94	74.15	71.82	47.40	61.94	53.54	38.26	52.11	46.06	29.25	41.79	62.11	41.13	58.85
ViT-H/14	Earth2Ocean(GPT-4o)	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**
	Earth2Ocean(qwen2.5VL-3B)	69.69	51.79	65.19	62.44	40.79	60.78	76.89	58.64	76.52	73.42	48.38	59.88	59.69	44.30	59.14	47.36	32.14	45.33	64.92	46.01	61.14
	Earth2Ocean(qwen2.5VL-7B)	70.95	52.15	64.80	63.28	41.95	63.62	76.19	57.70	76.20	73.42	48.41	59.95	60.72	45.84	60.43	49.71	33.91	47.29	65.71	46.66	62.05

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

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Salim Al-Amri, Jie Yang, and Xin Wang. Ulrs: A large-scale dataset for underwater image segmentation. *Sensors*, 21(14): 4675, 2021. 2

[3] M. Arda Aydın, Efe Mert Çırpar, Elvin Abdinli, Gozde Unal, and Yusuf H. Sahin. Itaclip: Boosting training-free semantic segmentation with image, text, and architectural enhancements. *arXiv preprint arXiv:2402.12345*, 2024. 3

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[5] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2403.23456*, 2024. 3

[6] Luca Barsellotti, Roberto Amoroso, Lorenzo Baraldi, and Rita Cucchiara. Fossil: Free open-vocabulary semantic segmentation through synthetic references retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 3

[7] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 4

[9] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 4

[10] 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

[11] Shuang Fu, Heng Zhang, Yifan Wang, and Chao Ma. Masnet: A multi-scale adaptive segmentation network for underwater imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1, 2

[12] Liang Haixin, Zheng Ziqiang, Ma Zeyu, and Sai-Kit Yeung. Marinedet: Towards open-marine object detection. *arXiv preprint arXiv:2310.01931*, 2023. 1

[13] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2403.01234*, 2024. 3

[14] 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

[15] 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. 2

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[17] Md Jahidul Islam, Caleb Edge, and Chen Xiao. Semantic segmentation of underwater imagery: Dataset and benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1, 2

[18] Shuwei Ji and Hongyuan Zhang. ISAT with Segment Anything: An Interactive Semi-Automatic Annotation Tool, 2024. Updated on 2025-02-07. 2

[19] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2403.45678*, 2024. 3

[20] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. *arXiv preprint arXiv:2403.54321*, 2024. 3

[21] Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. *arXiv preprint arXiv:2404.67890*, 2024. 3

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[24] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 4

[25] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 3, 4

[26] Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring efficient open-vocabulary segmentation in the remote sensing. *arXiv preprint arXiv:2509.12040*, 2025. 1

[27] Bingyu Li, Feiyu Wang, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Maris: Marine open-vocabulary in-

stance segmentation with geometric enhancement and semantic alignment. *arXiv preprint arXiv:2510.15398*, 2025. 1, 2

[28] Chao Li, Jun Xu, Zhanpeng Cui, Yimin Yang, and Chang Wen Chen. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. In *IEEE Transactions on Robotics*, 2017. 2

[29] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. Underwater image enhancement benchmark dataset and beyond. In *IEEE Transactions on Image Processing (TIP)*, 2020. 1, 2

[30] Chen Li, Yixiao Ge, Dian Li, and Ying Shan. Vision-language instruction tuning: A review and analysis. *arXiv preprint arXiv:2311.08172*, 2023. 2

[31] 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. 2

[32] 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. 2, 5

[33] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *arXiv preprint arXiv:2401.12345*, 2024. 3

[34] 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. 1, 2

[35] Huimin Ma, Zhen Wang, Mingqiang Xu, Qi Wu, and Jun Liu. Underwater image segmentation using deep learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. 1, 2

[36] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. *arXiv preprint arXiv:2402.45678*, 2024. 3

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 6

[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. U^2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 2020. 1, 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 3, 5, 4

[40] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 4

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 2

[42] Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. *arXiv preprint arXiv:2411.09219*, 2024. 3, 4

[43] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. *arXiv preprint arXiv:2404.34567*, 2024. 3

[44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[45] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2402.04567*, 2024. 3, 4

[46] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2403.67890*, 2024. 3

[47] Zhihui Wang, Chao Wang, Zheng Li, and Zhigeng Pan. Underwater image segmentation with adversarial networks. In *IEEE International Conference on Image Processing (ICIP)*, 2019. 1, 2

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[49] 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. 3

[50] 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. 3

[51] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[52] 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. 2, 6

[53] Xiaobo Yang and Xiaojin Gong. Tuning-free universally-supervised semantic segmentation. *arXiv preprint arXiv:2402.98765*, 2024. 3

[54] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing correlations in clip with off-the-shelf foundation models for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.10086*, 2024. 3, 4, 5

[55] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 6

[56] 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. 3

[57] 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. 3

[58] 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.

[59] 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. 3

[60] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*, 2023. 1

[61] Ziqiang Zheng, Yiwei Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. Marineinst: A foundation model for marine image analysis with instance visual description. In *European Conference on Computer Vision*, pages 239–257. Springer, 2024. 1

[62] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[63] Yifan Zhou, Ming Zhao, Ke Sun, Wei Li, and Shiqi Wang. A survey on underwater image enhancement and segmentation: From traditional methods to deep learning. *IEEE Access*, 2023. 1, 2