Supplementary Material for MarineInst: A Foundation Model for Marine Image Analysis with Instance Visual Description

Ziqiang Zheng[1*], Yiwei Chen[1], Huimin Zeng[2], Tuan-Anh Vu[1], Binh-Son Hua[3], and Sai-Kit Yeung[1]

[1] The Hong Kong University of Science and Technology
[2] Northeastern University
[3] Trinity College Dublin
Project website: https://marineinst.hkustvgd.com

In this supplementary material, we first provide more details about our MarineInst20M dataset (we release our MarineInst20M dataset at https://github.com/zhengziqiang/MarineInst20M) used for optimizing the MarineInst in Section 1. Furthermore, we provide more preliminaries about our proposed MarineInst, as well as the implementation details of our model in Section 2. Then more experimental results on underwater salient object segmentation, underwater object detection, text-to-image synthesis, instruction-following instance understanding, image storytelling, instruction-following segmentation, ablation studies, comparison with Mask R-CNN, and more qualitative results are provided in Section 3. More discussions about the failure cases and generalization ability of MarineInst, the main contribution of MarineInst over existing algorithms, related works, and the potential future directions are provided in Section 4.

1 MarineInst20M

1.1 Dataset Construction

Acquiring annotated marine datasets for training models is challenging since it is difficult to capture high-quality marine/underwater images due to the specific conditions of the environment, and it also requires domain expertise to label the collected imagery. The dynamic nature of underwater environments presents a great challenge for consistent and accurate data collection. Unlike terrestrial datasets, underwater datasets are limited and costly to obtain, which hampers the development of robust and accurate foundation models in the marine field for effective scene understanding from visual images. Factors such as water turbidity, varying light conditions, and the movement of water currents can significantly affect visibility and image quality, making it difficult to obtain clear and reliable visual data.

To address these challenges, our MarineInst20M is constructed from three main data sources: 1) existing public marine/underwater datasets; 2) manually

* Corresponding author: zhengziqiang1@gmail.com

collected and labeled images from private data, existing public datasets, and YouTube videos; and 3) public Internet images. We illustrate the remarkable diversity of our constructed MarineInst20M dataset in Figure 2. The images are with viewpoint variations, and visibility changes, from tiny plankton to large marine mammals and *etc.* We provide a comprehensive overview of our dataset construction procedures in Figure 1. Please follow the sorted procedures to better understand the construction details.

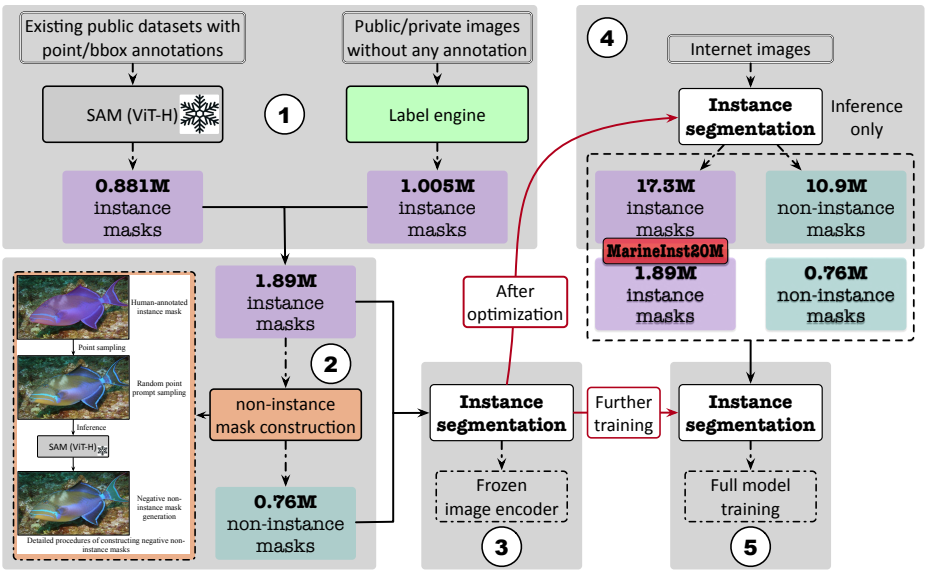


Fig. 1: Construction flow of MarineInst20M. Best viewed in color and follow the sorted procedures.

We provide comprehensive and informative statistics about our constructed MarineInst20M dataset in Table 1. For better illustration, we have also provided information of 1) # of object categories; 2) # of images; 3) annotation type; 4) whether the dataset contains non-organism objects; 5) image diversity; 6) whether the dataset contains complicated objects; 7) the original task/purpose of proposing such datasets or image collections; and 8) detailed statistics of the total images, total instance masks and average instance mask within each image. For better illustration, we have also provided the composition visualization of the instance masks from our MarineInst20M dataset in Figure 3. We refer the readers to check Figure 3 and Table 1 for more details about the composition of our constructed MarineInst20M dataset. In the following subsections, we will discuss the implementation details of building our MarineInst20M dataset in detail.

Table 1: The detailed statistics of the proposed MarineInst20M dataset for optimizing our MarineInst. We provide the information of 1) # of categories; 2) # of images; 3) the annotation type including *category*, *point*, *bounding box*, and *mask*; 4) whether the dataset provides the annotations for non-organism objects (**Non-org.** for short); 5) the diversity richness of various datasets or image collections; 6) whether the dataset contains the complicated objects *e.g.*, camouflaged objects, objects with irregular boundaries and non-rigid objects (**Comp.** for short); 7) the original task and motivation for proposing such datasets or collecting imagery/videos; and 8) the # of images, total instance masks and average instance mask within each image (denoted as Img/Inst./Aver.) after our processing procedures. 🤖 denotes the instance masks are **generated by models** based on various prompts or automatically (no prompts). 🧑 denotes instance masks are annotated by **human annotators** based on our written labeling tool. "–" indicates that the number cannot be reported or it is difficult to provide an accurate statistic. **Foregr.** denotes that the foreground objects are annotated (categories may vary from different images).

Datasets	Categories	Images	Annotation	Non-org.	Diversity	Comp.	Original task and motivation	Img/Inst./Aver.
Mastr1325 [20]	3	1,325	Mask	✓	Medium	✘	Marine obstacle segmentation	178/215/1.21
Marine Fouling [29]	3	267	BBOX	✘	Low	✓	Biological fouling detection	221/508/2.30
LaRS [97]	4	4,006	Mask	✓	Medium	✘	Marine obstacle segmentation	367/562/1.53
Fish4Knowledge [19]	Foregr.	27,370	BBOX	✘	Low	✘	Fish detection and tracking	470/470/1.00
MAS3K [52]	37	3,103	Mask	✘	Medium	✓	Marine animal segmentation	553/651/1.18
SUIM [41]	8	1,500	Mask	✓	Medium	✘	Underwater scene segmentation	589/1,091/1.85
Aquarium [3]	7	638	BBOX	✘	Medium	✘	Underwater object detection	632/4,182/6.62
UTB180 [16]	Foregr.	58,000	BBOX	✘	Low	✘	Underwater visual object tracking	900/900/1.00
TACO [33]	–	1,500	BBOX	✘	Medium	✘	Litter detection	1,109/2,656/2.39
Brackish [62]	Foregr.	15,084	BBOX	✘	Low	✘	Underwater fish detection and tracking	1,423/3,168/2.23
FLOW [27]	Foregr.	2,000	BBOX	✘	Medium	✘	Litter detection	1,825/3,850/2.11
DUO [56]	3	2,227	BBOX	✘	Medium	✘	Underwater object detection	2,170/13,090/6.03
DeepFish [71]	1	39,766	BBOX	✘	Medium	✘	Fish detection	4,396/12,381/2.82
Underwater Garbage [8]	15	416	BBOX	✓	Medium	✘	Underwater garbage detection	4,542/9,386/2.07
CoralNet [18]	191	416,512	Cate./Point	✓	High	✓	Sparse point based coral reef identification	4,615/5,753/1.25
WaterMask [53]	7	4,628	Mask	✓	High	✘	Underwater instance segmentation	4,628/28,410/6.14
IOCFish5k [72]	–	5,637	Point	✘	High	✓	Underwater object counting	5,382/192,900/35.84
OZFish [11]	Foregr.	9,242	BBOX	✘	Medium	✘	Underwater fish detection	6,235/38,875/6.23
URPC [12]	4	6,626	BBOX	✘	Medium	✘	Underwater object detection	6,330/38,307/6.05
TrashCan [39]	–	7,212	BBOX	✓	Medium	✘	Underwater trash detection	6,465/9,855/1.52
Trash-ICRA19 [33]	–	7,668	BBOX	✓	Medium	✘	Underwater trash detection	7,307/18,822/2.58
MarineDet [35]	821	22,679	BBOX	✓	High	✓	Open-marine object detection	22,679/39,243/1.73
FishNet [45]	17,357	94,532	Cate./BBOX	✘	High	✓	Fine-grained fish classification and detection	48,659/49,774/1.02
FathomNet [44]	–	109,871	BBOX	✓	High	✓	Underwater and deep-sea object detection	69,909/121,329/1.74
FishNet Open [13]	34	143,818	BBOX	✓	High	✓	Fish and non-fish detection	82,622/285,170/3.45
Total (1_{st} source)	–	284,206	🤖	✓	High	✓	Image collection of existing public datasets	284,206/881,548/3.10
HK-Reef-Fish [6]	–	730	🧑	✘	Low	✓	Fish identification	729/1,985/2.72
CoralVOS [96]	–	60,456	Mask;🧑	✓	Low	✘	Coral video segmentation	750/2,057/2.74
MVC [88]	–	1,026	🧑	✘	Medium	✘	Underwater object detection and segmentation	1,026/3,516/3.43
Sea Animal [10]	23	13,711	Category;🧑	✘	Medium	✘	Sea animal classification	3,080/7,448/2.42
ImageNet [30]	38	43,907	Category;🧑	✘	Low	✘	Scene classification	3,987/7,175/1.78
MVK [75]	–	4,872	🧑	✓	Medium	✘	Marine video retrieval	4,872/25,077/5.15
Oceanic Life [14]	–	7,990	🧑	✘	High	✘	Collection of Marine Life Imagery	5,029/20,811/4.14
Reef-Life-Survey [1]	–	7,089	🧑	✘	High	✓	Marine creature identification	7,075/12,502/1.77
Corals-of-world [32]	–	8,217	🧑	✘	Medium	✘	Coral reef identification	7,636/17,264/2.26
Wildfish++ [94]	2,348	103,034	Category;🧑	✓	High	✘	Fine-grained fish classification	9,367/17,075/1.82
FishDB [82]	–	10,074	🧑	✘	Medium	✘	Fish species identification	9,905/18,914/1.91
Reeflex [2]	–	15,174	🧑	✘	High	✓	Marine creature identification	15,088/61,656/4.09
Fish-of-Australia [21]	–	20,795	🧑	✘	Medium	✓	Fish species identification	19,269/44,342/2.30
Youtube	–	20,935	🧑	✓	High	✓	Video collection	20,935/201,290/9.61
EOL [9]	–	3,498,763	🧑	✘	High	✓	Species identification	23,141/80,128/3.46
Private data	–	24,420	🧑	✓	High	✓	Surveying; Diving; Snorkeling	24,420/289,898/11.87
Total (2_{nd} source)	–	156,309	🧑	✓	High	✓	Image collection with manual annotations	156,309/811,138/5.19
Internet images [4, 5, 7]	–	35,172	🧑	✓	High	✓	Image collection (human labeled)	35,172/194,010/5.52
Internet images [4, 5, 7]	–	1,945,714	🤖	✓	High	✓	Image collection (automatic mask generation)	1.94M/17.3M/8.89
Total (3_{rd} source)	–	1,980,346	🤖;🧑	✓	High	✓	Image collection of Internet images	1.98M/17.5M/8.84
MarineInst20M	–	2,420,851	🤖;🧑	✓	High	✓	Instance segmentation and captioning	2.42M/19.2M/7.93



Fig. 2: We provide the example images from our MarineInst20M dataset. The images demonstrate a remarkable image diversity. Please zoom in to see more details.

1.2 Existing Public Datasets

In our work, we propose to utilize the existing public marine/underwater datasets with various formats of annotations (*e.g.*, point, box, and mask) for optimizing our MarineInst model. As mentioned in our main manuscript, we infer SAM

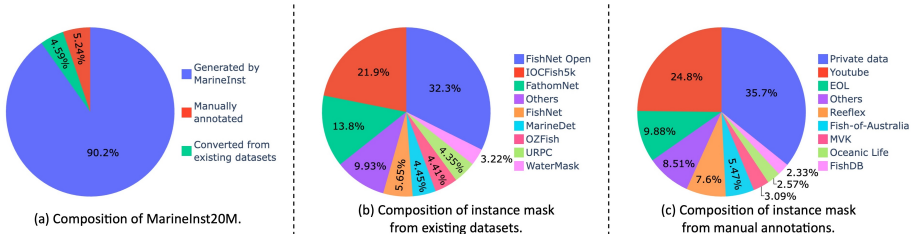


Fig. 3: We provide the visualization of the composition of all the instance masks from our MarineInst20M dataset: a) demonstrates the composition of instance masks from three different sources; b) illustrates the composition of the instance masks from the existing public dataset after our conversion; c) shows the composition of the instance masks labeled by the human annotators in this work. For both b) and c), we only visualize the top 8 components for better readability.

with the ViT-H backbone by point or box prompts to obtain the instance masks. Furthermore, we perform the filtering to remove those generated masks with the value of "predicted IoU" lower than 0.88. For those datasets with mask annotations, we only pick up partial images with satisfactory instance mask annotations, indicating that each mask only exhibits one instance. After converting the annotations of existing public datasets to instance masks, we have obtained around **284K** images with **882K** high-quality instance masks for optimizing our model. To convince the readers, we provide the visualization of the generated instance masks for the images from those datasets in Figure 4. As illustrated, most of the generated instance masks are reasonable and accurate enough to optimize our MarineInst model as the pseudo ground truth. With the training data from various public datasets, our model could be driven to obtain a strong performance across a variety of marine data.

1.3 Data with Manual Annotations

As for our manually labeled data, the marine images mainly come from 1) existing public datasets; 2) marine research websites (including HK-Reef-Fish [6], Reeflex [2], Reef-Life-Survey [1], FishDB [82], Fish-of-Australia [21], Corals-of-world [32] and EOL [9]); 3) YouTube videos; and 4) the private data contributed by local amateurs and marine biologists from different sites. Please check Table 1 for more details.

For the images from the existing datasets [10, 14, 30, 75, 94, 95], which are without any annotations, we manually label the images from these datasets for obtaining the instance masks. Wildfish++ [94], Sea Animal [10], and ImageNet [30] datasets only provide the category annotations. Please note that we only adopt the images from 38 ocean-related categories for labeling on the ImageNet dataset. We randomly select **9,367**, **3,080**, and **3,987** images from these three datasets for instance mask labeling, respectively. The MVC [95], MVK [75],

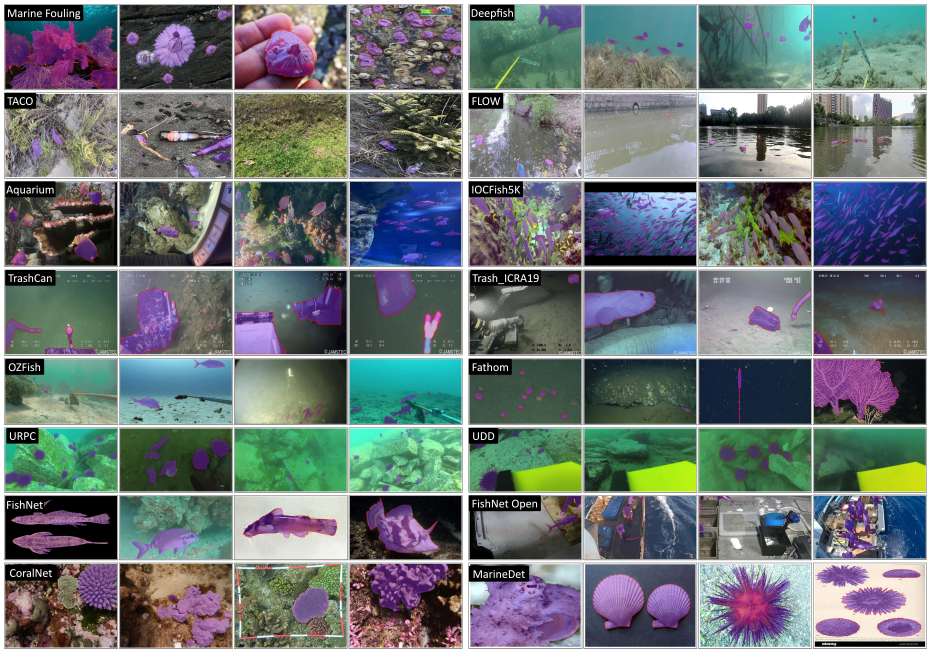


Fig. 4: We provide the example images with the instance mask visualizations from the existing public datasets after our processing procedures, converting the point or bounding box annotations to instance masks. Please zoom in to check more details. `Results shown are not cherry picked.`

and CoralVOS [96] datasets only provide marine video clips. We download the videos from these three datasets and manually from some frames for instance mask labeling. The statistics for the number of cropped images and labeled instance masks are also reported in Table 1.

As for the images from the marine research websites, we utilize our labeling tool for instance mask annotation. Considering the EOL website [9] contains redundant images, we only randomly pick up **23,141** images for labeling. We have labeled nearly all the images from HK-Reef-Fish [6] (**729** images labeled), Reeflex [2] (**15,088** images labeled), Reef-Life-Survey [1] (**7,075** images labeled), FishDB [82] (**9,905** images labeled), Fish-of-Australia [21] (**19,269** images labeled), Corals-of-world [32] (**7,636** images labeled) websites. Among all these images, some images are discarded or ignored due to the following reasons: 1) the images are corrupted; 2) the visibility of the captured images is drastically degraded; and 3) there is no biologically meaningful instance within the image.

For the videos downloaded from YouTube (around 1,000 videos with a time duration of 1,243 hours in total), we manually or automatically crop frames from the videos and then manually label these cropped frames. Finally, we generate **201K** instance masks for **20.9K** images. The private data comes from the con-



Fig. 5: Instance mask visualization of example images from MarineInst20M dataset. The instance masks are all **labeled by human annotators**.

tributions of local amateurs and biologists (domain experts) from various sites around the world. We have finally obtained **24,420** images with **289,808** instance mask annotations from these private data. Such images with annotations are significantly valuable for optimizing a strong model to satisfy the requirements of the domain users. The number of instance masks of each dataset or image source is also provided in Table 1.

Finally, we provide the visualization of the instance masks labeled by our human annotators in Figure 5.

1.4 Public Internet Data

We adopt crowdsourcing techniques to scrape public images, which are mainly from Flickr [4], Gettyimages [5], and Shutterstock [7]. The public images come from many different natural underwater environments, covering marine vision tasks such as ocean exploration, human-computer intelligence cooperation, and underwater autopilot. To ensure the high diversity and comprehensive coverage of the scraped public images. We construct a list of keywords for querying these public image websites. The keywords are:

"underwater"	"marine"	"sea life"	"fish"	"sea creature"
"ocean"	"marine life"	"marine biology"	"diving"	"snorkeling"
"colorful reef creatures"	"marine mammal"	"coral reef"	"marine biodiversity"	"deep sea"
"oceanic abyss"	"crustaceans"	"aquatic"	"microscopic sea life"	"nudibranch"
"sea slug"	"frogfish"	"aquariums"	"shark"	"dolphin"
"underwater flora and fauna"	"mollusca"	"beach"	"turtle"	"sea star"
"starfish"	"scallop"	"sea urchin"	"porifera"	"anemone"
"cnidaria"	"whale"	"orca"	"trua"	"seahorse"
"sea dragon"	"seaweed"	"diver"	"jelly fish"	"seal"
"sea otter"	"marine birds"			

We adopt these keywords to scrape public Internet images. After collecting redundant marine images downloaded from the public Internet, we randomly



Fig. 6: The world cloud visualization of the top 1,000 words in all the extracted phrases from the alt-texts of the public Internet images.

pick up partial images from whole public Internet images and perform manual instance mask labeling: **35,172** images with **194,010** instance masks. It is worth noting that public images from Internet websites contain some text descriptions in the form of alt-text captions. Based on these alt-text captions, we adopt the KeyPhraseTransformer open source[1] to extract the phrases from these alt-texts. We have also provided the statistics for the extracted phrases. There are approximately **470K** different phrases in total, where **24,283, 5,063, 2,490, and 508** phrases appeared over *10, 50, 100, and 500* times, respectively. Furthermore, we provide a world cloud visualization of the top 1,000 phrases in all the extracted phrases from the alt-texts in Figure 6. As demonstrated, the scraped public Internet images have significant diversity and contain comprehensive marine object conceptions.

Finally, we provide the instance masks generated by our MarineInst model in Figure 7. All the images shown in Fig. Figure 7 come from the public Internet. It is worth noting that all the instance masks are automatically generated without any prompts from the users. In other words, no human intervention is involved during the whole instance mask generation procedure. We observe that most of the generated instance masks are reasonable and accurate enough for further optimizing our MarineInst model. We also acknowledge that there are still some **false negatives**, in which our model failed to segment under the automatic setting.

[1] `https://github.com/Shivanandroy/KeyPhraseTransformer`



Fig. 7: Instance mask visualization of example images from MarineInst20M dataset. The instance masks are all **automatically generated by our MarineInst model without any prompts**. `Results shown are not cherry picked.`

1.5 Discussions

Our MarineInst20M dataset contains around 20 million instance masks with detailed and comprehensive captions. Our MarineInst20M dataset could enable semantic segmentation, instance segmentation, and object detection, either individually or in combination. Furthermore, in our MarineInst20M dataset, we extend the marine instance segmentation to the open-vocabulary setting, where the model is optimized to segment the objects within the image based on a language description from humans.

Instance masks with generated instance-level semantic captions from our MarineInst20M dataset are illustrated in Figure 8. As demonstrated, the generated instance masks with captions are reasonable, enabling a comprehensive understanding of marine images of different semantic granularities. Thanks to the combined architecture of automatic instance segmentation and semantic instance captioning, MarineInst produces satisfactory labeling for most samples and can provide more detailed, diverse, and comprehensive annotations. More importantly, the semantic instance understanding has also produced a large number of question-answer pairs: *e.g.,* 1) *single-instruction-input-single-mask-output*; 2) *single-instruction-input-multiple-mask-output*; 3) *single-mask-input-single-caption-output*; 4) *multiple-mask-input-spatial-reasoning-output*; and 5) *multiple-mask-input-relationship-summarizing-output*. We leave more details and discussions of generating such pairs in Section 2.2.



Fig. 8: Visualization of the generated instance masks with comprehensive and detailed semantic instance captions. Please zoom in to check more details. `Results shown are not cherry picked.`

2 MarineInst

2.1 Preliminaries

The instance mask generation of MarineInst is built upon SAM [47], the foundation model for generic segmentation driven by the biggest SA-1B dataset for mask prediction to date. SAM consists of three components, a prompt encoder

$\texttt{Prop}(\cdot)$, a heavy image encoder $\texttt{Enc}(\cdot)$, and a lightweight mask decoder $\texttt{Dec}(\cdot)$. As a promotable model, SAM is fed with an image I and a set of user prompts P, including point, box, or coarse mask prompts. SAM utilizes $\texttt{Enc}(\cdot)$ to obtain image embedding, and adopts $\texttt{Prop}(\cdot)$ to encode prompts P of a length k into prompt tokens as:

$$F_I = \texttt{Enc}(I), \quad T_P = \texttt{Prop}(P), \tag{1}$$

where $F_I \in \mathbb{R}^{h \times w \times c}$ and $T_P \in \mathbb{R}^{k \times c}$, with h, w denoting the resolution of the image embedding and $c = 256$ denoting the feature dimension. F_I and T_P are fed into the mask decoder $\texttt{Dec}(\cdot)$ for mask generation. SAM constructs the input tokens of $\texttt{Dec}(\cdot)$ by concatenating the learnable mask tokens T_M generated by $\texttt{Dec}(\cdot)$ and the prompt tokens T_P for generating the mask output, formulated as

$$M = \texttt{Dec}(F_I, \texttt{Concat}(T_M, T_P)), \tag{2}$$

where M denotes segmentation masks yielded by SAM. SAM is primarily optimized by in-air images, making it less effective in segmenting mariner images. Furthermore, SAM has a strong ability to group pixels with similar appearances or textures together for mask generation. But in contrast, it cannot generate instance-level masks due to its semantic-agnostic nature.

To address these issues, we propose to formulate a lightweight binary filtering branch to enable MarineInst to discriminate whether the generated masks are instance masks. We conduct attention-based feature interaction between mask generation and binary instance filtering:

$$\mathcal{L}_{bin.} = -(y \log(p) + (1-y)\log(1-p)),\ p = \texttt{MLP}(F_I, \texttt{Concat}(T_M, T_P)), \tag{3}$$

where y denotes the binary ground truth. T_M and T_P are the learnable mask tokens and prompt tokens, respectively. F_I is the image embedding from the image encoder and $\texttt{MLP}$ is a lightweight MLP layer. Both "positive" instance masks and "negative" non-instance masks are used for optimizing our MarineInst. It is worth noting that our MarineInst also inherits the ability to receive user prompts for generating desired masks.

2.2 Implementation Details

Instance segmentation. We adopt SAM as our backbone and utilize it as an effective network initialization. For our iterative optimization, we first continuously pre-train our MarineInst model on the instance mask data from the public existing datasets and our manual annotations (**1.89M** positive instance masks in total). Both the positive instance masks and negative non-instance masks (**0.76M** non-instance masks) have been utilized for optimizing our MarineInst model. For non-instance mask generation, we randomly pick up one point inside the whole instance mask and infer SAM for mask generation, where we only preserve the masks with predicted IoU over 0.88. We regard the generated mask as non-instance if the overlapping between the generated mask and the original instance mask is below 0.5. We provide the detailed procedures for

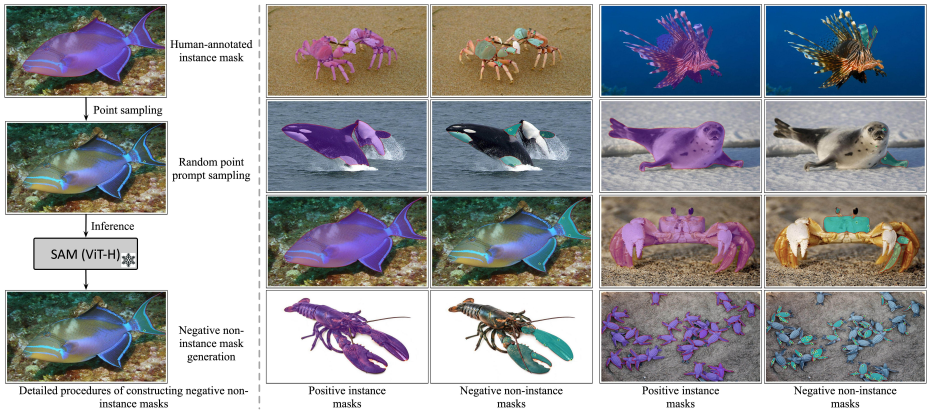
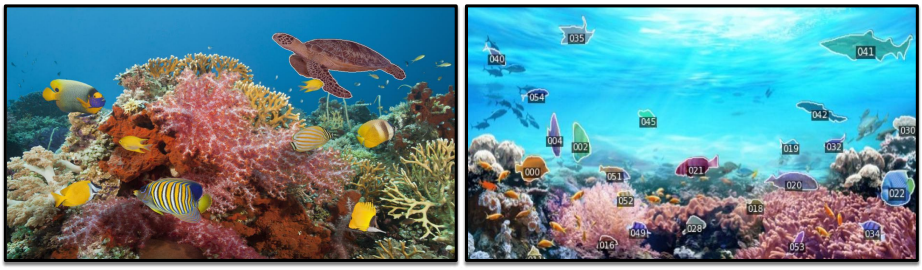


Fig. 9: Details procedures constructing negative masks based on positive instance masks and visualizations of both positive instance masks and negative non-instance masks.

constructing the negative masks in Figure 9. To provide a better illustration, we have also provided the visualization images of both positive instance masks and negative non-instance masks in Figure 9. In total, there are **2.65M** masks (positive and negative) used for optimization. At this stage, we keep the heavy encoder frozen and only optimize the prompt encoder and mask decoder. The training prompt is only the point prompt (three random points inside the whole mask). After optimizing MarineInst (ViT-H backbone) for 3 epochs, we utilize the trained model for generating instance masks for those public Internet images. During the automatic instance mask generation procedure, we follow the automatic mask generation pipeline of SAM and generate the grid points (32×32) as point prompts for automatic instance mask generation. The IoU threshold and stability threshold are set to 0.82 to remove the automatically generated low-quality instance masks. Please note that the automatically non-instance masks (**11.7M** non-instance masks) have also been preserved for further optimizing our MarineInst model.

Then the whole model is continuously pre-trained on our MarineInst20M dataset (with **19.2M** instance masks and **11.7M** non-instance masks) to better extract efficient marine feature representations. Please note that the whole model, including the heavy image encoder, prompt encoder, and mask decoder, has been optimized on our MarineInst20M dataset. The training prompt is a combination of both point and box prompts. We have optimized our MarineInst on MarineInst20M dataset for 3 epochs. To promote the generalization ability and robustness of our trained foundation model, we apply various augmentation techniques to increase our training data. Besides the color jitter adopted in SAM, we also conduct random cropping, rotation, enlarging, and flipping to simulate the high diversity of marine images. We perform experiments on 8 H800



Single mask Multiple masks with mask IDs

Fig. 10: Two settings of instruction-following instance understanding: 1) single mask and 2) multiple masks with mask IDs.

GPUs and set the batch size per GPU to 1. We resize the original images to the required size and set the longest side of the resized image to 1024 while keeping the original image ratio. It takes *1,056*, *1,392*, and *2,064* GPU hours to optimize the MarineInst model with *ViT-B*, *ViT-L*, and *ViT-H* backbones, respectively.

Instance captioning. For semantic caption generation for the instance mask, we first crop image regions from the whole image based on the generated instance masks and then feed the cropped images to the frozen VLM. we adopt the frozen MarineGPT [89] to include a ViT backbone with a pre-trained Q-Former and Vicuna-V0 [28] (tuned from LLaMA-13B [74]) as the decoder to generate responses. It is worth noting that we adopt the pre-trained model at the first stage. The value for beam search is set to 0.1 and the maximum length of the generated tokens for the cropped images is 50. The input size of the fed images is set to 224×224.

We formulate mask-caption pairs (m, θ) to enable instruction-following tasks: (a) *instruction-following instance understanding* and (b) *instruction-following segmentation*. For instance understanding, the mapping $m \rightarrow \theta$ is described by "`Human`: The image is <*`image`*>. Please generate caption for instance <*`mask`*>: m. `Response`: θ", where <*`image`*> and <*`mask`*> are image and mask tokens, respectively. For instruction-following segmentation, the mapping $\theta \rightarrow m$ performs segmentation following user intents (discussed in supplementary). We optimize MarineInst with the constructed instruction-following data and enable MarineInst to handle various tasks aligned with user intents.

Instruction-following instance understanding. We construct the pair of instance masks and the generated semantic captions to optimize our model. We adopt MarineGPT [89] as our baseline and we fine-tune the released pre-trained models to our instruction-following instance understanding tasks. Following the experimental setting of [89], we optimize both the Q-Former and linear layer parts of the whole model. There are two settings for this instruction-following instance understanding task: 1) *single mask* and 2) *multiple masks with randomly assigned mask IDs* as demonstrated in Figure 10.

As for the instruction construction, we only choose one simple instruction for the former *single mask* setting:

- The image is <*image*>. Please describe the object in the mask.

For the latter *multiple masks* setting: we formulate the following 4 different instructions.

- The image is <*image*>. Please generate the caption for the instance mask m with mask ID <*mask ID*>.
- The image is <*image*>. Please describe what the object is doing in the instance mask m with mask ID <*mask ID*>.
- The image is <*image*>. Please explain the relationship between the object in the instance mask m with mask ID <*mask ID*> and the background environments.
- The image is <*image*>. Please explain the relationship between the object in the instance mask m with mask ID <*mask ID*> and the object in the instance mask m with mask ID <*mask ID*>.

Under the two settings, we have formulated **307,272** and **338,650** instruction-following training data for the *single mask* and *multiple masks* settings, respectively. The batch size is set to 4 and the number of total steps in an epoch is 10,000. We perform experiments on 4 Tesla A40 GPUs and optimize our MarineInst by 4 epochs. The image size is set to 384×384 during the tuning procedure.

Instruction-following segmentation. Our MarineInst could also be utilized for instruction-following segmentation, generating required instance masks based on the user instructions. We follow the data preparation of LISA [48] to generate the instruction-following training data. We formulate the instruction as follows: `Human`: The image is <*image*>. Please generate the mask with <*caption*> for me. `Response`: <*SEG*>. <*caption*> and <*SEG*> are the generated semantic captions and corresponding mask annotations generated by our MarineInst, respectively. We construct **307,272** instruction-following training data for tuning our MarineInst to perform segmentation based on user instructions. The batch size is set to 4 and the number of total steps in an epoch is 10,000. We perform experiments on 4 Tesla A40 GPUs and optimize our MarineInst by 4 epochs. Please note that the "instruction-following segmentation" could be regarded as the reverse procedure of "instruction-following instance understanding" (*single mask* setting).

User studies. We perform user studies to evaluate the accuracy of the generated semantics of the four different algorithms: SSA [23], OVSAM [84], MarineInst (BLIP2 [50]) and MarineInst (MarineGPT [89]). We randomly pick up 1,000 mask-caption pairs generated by these four algorithms each. Please note that we first randomly choose 1,000 generated instance masks from the whole pool. Then BLIP2 and MarineGPT have been utilized for generating the semantic captions. For BLIP2, we adopt the pre-trained model "blip2-opt-6.7b" with the LLM "OPT-6.7B". MarineGPT is based on the Vicuna-V0 [28] (tuned from LLaMA-13B [74]). It is worth noting that the input for BLIP2 and MarineGPT are the

Table 2: The underwater salient object segmentation results on the USOD10K dataset [40]. "Depth" ("Edge") indicates that the depth map (edge map) of the original RGB images has been utilized for optimization as the additional clues.

Method	Additional clues	$S_m \uparrow$	$E_\epsilon^{max} \uparrow$	max$F \uparrow$	MAE $\downarrow$
SVAM-Net [42]	None	0.7465	.7649	0.6451	0.0915
CTDNet [87]	Edge	0.9085	0.9531	0.9073	0.0285
CDINet [85]	Depth	0.7049	0.8644	0.7362	0.0904
SGL-KRN [80]	Depth & Edge	0.9214	0.9633	**0.9245**	0.0237
TC-USOD [40]	Depth & Edge	**0.9215**	**0.9683**	0.9236	**0.0201**
MarineInst	None	0.9103	0.9411	0.8876	0.0256

same. As for OVSAM, we regard the BBOX of the generated instance masks as the box prompt to generate mask prediction with semantic category annotation. Similarly, we infer SSA with the same box prompts to generate masks with semantics. For subject fidelity, we asked 3 students from the marine biology field to answer 1,000 scoring questions, totaling 12,000 ($4 \times 3 \times 1000$) answers. The students are asked to answer the question: "Please give your satisfactory score (from 1 to 5) based on the correctness, helpfulness, and information richness of generated captions for the instance mask". Then we compute the average satisfactory score for each algorithm. The higher satisfactory score indicates a higher accuracy of the generated semantic captions for the instance masks.

Comparisons. In this work, we mainly include SAM [47], Semantic-SAM [49], SSA [23], OVSAM [84] and Grounded SAM [69] for comparison. Both SAM and MarineInst could be inferred under the automatic, point prompt based, and box prompt based settings. Under the automatic setting, SAM and MarineInst automatically generate masks based on 32×32 grid points. Semantic-SAM produces instance masks by setting the semantic granularity to 3. SSA is based on SAM so we do **not** compute the quantitative results of the instance segmentation for SSA. We adopt the predicted IoU score as the confidence score when evaluating the instance segmentation performance. For evaluating OVSAM, we adopt its official configuration with "sam_r50x16_fpn" as the network backbone, where the IoU branch has been discarded due to knowledge distillation. Thus, we choose the confidence score for category prediction as the confidence score for evaluating the instance segmentation performance. For Grounded SAM, we first utilize Grounding DINO [58] to yield the bounding box predictions based on the text query. We preserve the predictions with a similarity score over 0.25 as suggested in Grounding DINO. Then the box predictions are regarded as box prompts to generate dense instance masks.

Table 3: We report the quantitative object detection results of MarineDet and MarineInst-Det on URPC [12] dataset. "MarineInst20M (**human-annotated**)" indicates that we only utilize the bounding box annotations from the human-annotated instance masks (converted from existing public datasets and manual annotations) for pre-training. "MarineInst20M (**human-annotated+model-generated**)" indicates that bounding box annotations from all the instance masks in our MarineInst20M dataset have been used for pre-training.

Method	Pre-training data	Sea urchin	Scollop	Starfish	Sea cucumber	mAP_{50}
MarineDet [35]	MarineDet dataset	86.4	83.8	45.8	66.6	70.6
MarineInst-Det	MarineInst20M (human-annotated)	89.5$_{+3.1}$	86.6$_{+2.8}$	60.2$_{+14.4}$	69.4$_{+2.8}$	76.4$_{+5.8}$
MarineInst-Det	MarineInst20M (human-annotated+model-generated)	**90.2**$_{+3.8}$	**87.7**$_{+3.9}$	**63.2**$_{+17.4}$	**70.6**$_{+4.0}$	**77.9**$_{+7.3}$

3 More Experiments

3.1 Underwater Salient Object Segmentation

With pre-trained on huge instance masks from the redundant marine images, our MarineInst model could then be fine-tuned to perform underwater salient object segmentation. We present the qualitative results of our MarineInst model on the USOD10K dataset and more quantitative result comparisons with existing state-of-the-art algorithms in Table 2. Please note that we only utilize the height and width of the input images as the box prompt to yield the salient predictions. Our model is optimized without the supervision of the additional depth maps and the edge maps, which require additional pre-processing procedures to obtain side clues and are not easy to obtain for wide underwater scenarios. Our method could achieve comparable performance with CTDNet [87], which utilizes edge information as the additional supervision. When compared with SVAM-Net [42] under the same experimental setting without any additional clues, our MarineInst could achieve much better results.

3.2 Underwater Object Detection

Since our MarineInst20M dataset has large-scale instance masks, we could easily obtain the bounding box annotations for various marine objects. With such redundant BBOX annotations for a wide spectrum of marine objects, we aim to optimize a powerful region proposal network (RPN) model. We follow the experimental setting of MarineDet [35] and utilize the bounding box annotations of our MarineInst20M dataset to pre-train a powerful detection model (denoted as **MarineInst-Det**). We adopt the RegionCLIP [90] as our baseline and continuously pre-train the model with ResNet-50 backbone on our MarineInst20M dataset. Then we fine-tune our pre-trained model to the downstream URPC dataset [12]. The quantitative object detection results of our MarineInst-Det and existing MarineDet are in Table 3. We conduct our MarineInst-Det pre-training

under two settings: 1) **human-annotated**: only the bounding box annotations from the converted instance masks and our manually labeled instance masks have been utilized for pre-training; 2) **human-annotated+model-generated**: the bounding box annotations from all the instance masks have been used for pre-training, including the converted instance masks, manually labeled instance masks, and model-generated instance masks. As reported, our model outperforms the existing MarineDet by a large margin. We attribute such observable performance gains to our huge pre-training data. Meanwhile, we also observe that the model-generated instance masks could also promote the downstream fine-tuning performance (77.9 vs. 76.4) due to more training data involved. This observation also demonstrates that the model-generated instance masks can further promote marine object detection, indicating reasonable instance mask generation to some extent. MarineInst-Det has a stronger ability to extract efficient and effective features from visual images, even under some challenging conditions. Our MarineInst20M dataset and pre-trained object detection model will be a significant contribution to effective and efficient object detection in the marine field.

3.3 Text-to-Image Synthesis

Marine text-to-image synthesis is a cutting-edge technology with various applications in marine science, research, and education. It combines natural language descriptions with advanced image generation techniques [63, 70] to create vivid and realistic visual representations of underwater environments and marine life. To demonstrate that our MarineInst20M dataset could promote the marine text-to-image synthesis. We fine-tune the "stable-diffusion-v1-5" (**SD1.5** for short) model based on the training data, where we construct **2M** image-text pairs: **1M** pairs are from the public Internet images with alt-text captions and another **1M** pairs are from the cropped images with the model generated instance captions. For the latter 1M pairs, we crop the images based on the instance masks, and the area of the cropped images is required to be larger than 256×256. The cropped close-up images provide valuable guidance for synthesizing reasonable marine creatures. We fine-tune the pre-trained stable diffusion model based on our training data for 10,000 steps and the batch size is set to 192. We provide some example results in Figure 11 under two settings: 1) without fine-tuning (original pre-trained frozen model) and 2) with fine-tuning. To help the readers better compare the image synthesis results, we have also provided the reference images (real images) for each required marine species. Please note that we fed the pre-trained model and our fine-tuned model with the same text prompts generated by ChatGPT. As illustrated in Figure 11, further fine-tuning on our MarineInst20M dataset could lead to much better image synthesis, which aligns the text prompts better. Our trained model has a stronger ability to synthesize more reasonable images, which comply with physical and biological laws. We attribute this improvement to the knowledge injection driven by our high-quality text-instance pairs. Finally, to quantitatively measure the text-to-image synthesis performance, we compute the FID [38] scores (lower is better) between

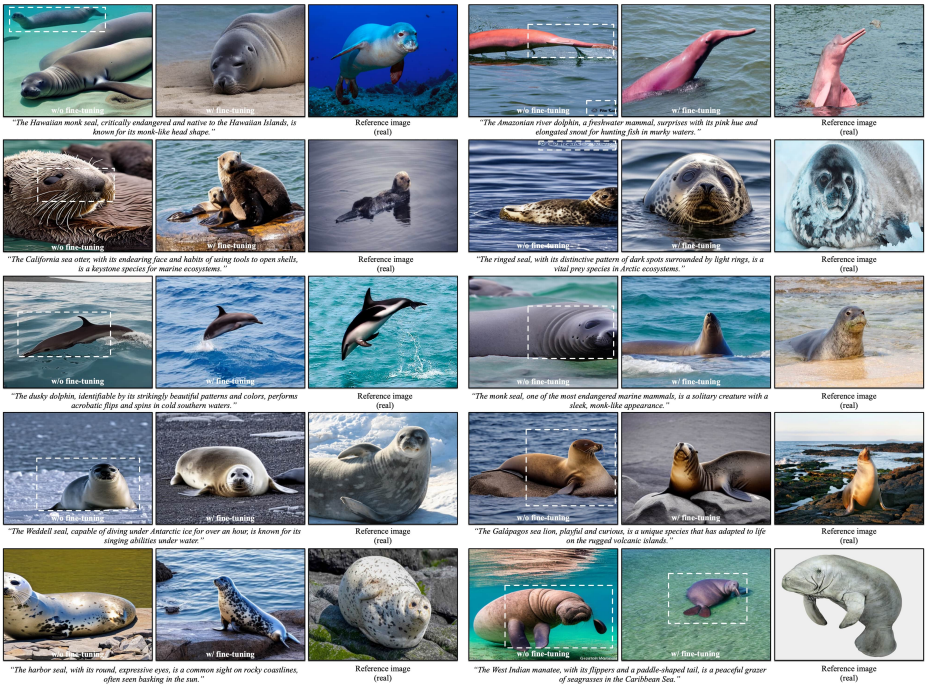


Fig. 11: We report the marine text-to-image synthesis results under two settings: a) without fine-tuning and b) with fine-tuning on our MarineInst20M dataset. The reference images from the required marine species have also been provided for the readers to better compare the synthesis performance. Best viewed in color. `Results shown are not cherry picked.`

10,000 synthesized marine images (from *vanilla* and *fine-tuned* SD1.5 model, respectively) and aligned real images (we adopt the text captions of the real images as the text prompts for generating the marine images). The vanilla SD1.5 achieves 25.91 while our fine-tuned counterpart could obtain 19.22 in terms of FID metric.

3.4 Instruction-following Instance Understanding

We have also quantitatively evaluated the ability of MarineInst to perform the instruction-following instance understanding. We first construct 1,000 testing images and corresponding human-constructed reference captions (describing appearance, pose, activity, event, and other attributes) for instance understanding. The average word length of reference captions is *44.21* for instruction-following instance understanding. The instruction is "*Describe the object in mask*". We use widely used captioning metrics[2] to compute scores between model-generated responses/captions and reference captions, comparing with general-purpose MiniGPT-

Table 4: We report the quantitative results of instruction-following instance understanding. The generic MiniGPT-4 and the domain-specific MarineGPT are included for comparison.

Methods	CIDEr [76]↑	METEOR [17]↑	BLEU-4 [61]↑	CLIP-S [37]↑	RefCLIP-S [37]↑
MiniGPT-4 [93]	13.18	12.19	6.10	73.82	71.27
MarineGPT [89]	16.78	13.00	6.78	74.60	71.89
MarineInst	**25.06**	**15.77**	**9.51**	**75.71**	**76.01**

Table 5: We report the quantitative results of marine image storytelling. The generic MiniGPT-4 and the domain-specific MarineGPT are included for comparison.

Methods	CIDEr [76]↑	METEOR [17]↑	BLEU-4 [61]↑	CLIP-S [37]↑	RefCLIP-S [37]↑
MiniGPT-4 [93]	18.43	13.39	7.05	74.42	71.53
MarineGPT [89]	27.19	16.32	9.74	75.08	75.68
MarineInst	**30.43**	**17.06**	**10.76**	**76.47**	**76.41**

4 and domain-specific MarineGPT. The quantitative results are reported in Table 4. As observed, MiniGPT-4 and MarineGPT suffer from a weaker ability to understand specified instances compared with our MarineInst due to the two models were not optimized by instance-level supervision.

3.5 Image Storytelling

A picture is said to be worth a thousand words, conveying complicated conceptions and relationships between objects. Marine image storytelling could bring a deeper understanding of the marine realm to both scientists and the general public. It enables the automatic generation of descriptive and informative captions for marine images, allowing researchers to annotate vast datasets efficiently. Furthermore, marine enthusiasts and amateurs can gain valuable insights into the intricate marine world beneath the waves by receiving detailed explanations and context for the visuals they encounter. Considering our MarineInst could generate comprehensive and meaningful semantic instance captions for each instance mask. We ask ChatGPT-3.5 to generate a summary of all the generated semantic instance captions for the instance masks within the images for performing **image storytelling**. The merged caption is regarded as the image-level caption for the whole image. In this way, we could perform more fine-grained image-level captioning and understanding. We provide some examples in Figure 12. GPT-4V is included for comparison. The effective and detailed marine image captioning based on our MarineInst could enrich the connection with the ocean and its myriad inhabitants, making it a labeling tool in marine exploration and communication.

Similarly, we perform the evaluation of the marine image storytelling. Following the same evaluation pipeline as the instruction-following instance un-

[2] https://github.com/jmhessel/clipscore



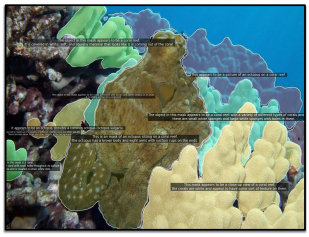
This image reveals the rich tapestry of marine biodiversity, where **a black sea urchin** rests unassumingly on the coral reef floor. Encircled by an array of small rocks and teeming with various sea life, the urchin's spiny appearance adds texture and depth to the underwater landscape. Sharing this aquatic realm, **an octopus with a captivating pattern of brown and white stripes stretches out across the seabed**. Its two large, watchful eyes and all eight legs splayed outwards, the octopus is an embodiment of the adaptability and complexity of oceanic creatures, each contributing to the dynamic ecosystem of the reef.

The image captures **an octopus** in its natural habitat on a vibrant coral reef. The octopus has a mottled brown and tan body, blending seamlessly with its surroundings, which is indicative of its camouflage capabilities. Its eight arms, covered in suckers, are sprawled across the reef, allowing it to anchor and maneuver with ease. The creature's large, prominent eyes are fixated in the direction of the camera, giving it a seemingly aware and intelligent expression. The ocean floor is a collage of marine life, including various corals, sponges, and small fish, creating a dynamic and biodiverse ecosystem. The image, **credited to David Comas 2019**, offers a glimpse into the rich and complex world that thrives beneath the ocean's surface.



This image juxtaposes **two formidable vessels**, each serving a unique purpose on the vast expanse of the sea. The first is **a large, red and white deep-sea scientific research vessel**, its open hull indicative of its exploratory mission. Equipped with two large cranes on the back deck and a substantial scientific instrument atop, it is the epitome of human ingenuity in the quest for oceanographic knowledge. Nearby, **a large tanker ship asserts its presence with a robust hull designed for endurance and capacity,** its deck crowned by several towering masts that speak to its commercial navigational prowess. Together, these ships represent the diverse nature of maritime endeavors, from the pursuit of scientific discovery to the necessities of global commerce.

The image depicts a maritime scene with **two ships** on a calm sea under an overcast sky. In the foreground, a large red and white vessel, likely a **deep-sea scientific research ship**, is equipped with cranes and various instruments, indicating its purpose for oceanographic exploration. Its open deck and the configuration suggest active work or the capability to perform complex tasks at sea. In the background, another ship, **possibly a commercial cargo vessel or tanker due to its size and structure**, is visible through a misty haze, emphasizing the depth of the scene. The tranquil water and subdued lighting convey a serene oceanic ambiance.



This captivating image showcases a vibrant underwater scene, featuring an octopus with a brown body, casually perched on the diverse ecosystem of a coral reef. The octopus, possibly a **common octopus (Octopus vulgaris)**, displays its eight arms adorned with suction cups, essential for navigation and hunting on the reef. Amidst this marine spectacle is a **striking coral**, characterized by its pink and white hues, round shape, and surface dotted with small holes, contributing to the reef's complexity. The coral reef itself is a tapestry of life, composed of variously colored corals with distinctive brown and white stripes, and an assortment of sponges, some small and compact, others large with conspicuous openings. The **white, soft, and squishy textures emerging from the corals** add a three-dimensional aspect to the scene, while the background is peppered with rocks.

The photograph captures **an octopus nestled amongst coral,** displaying its remarkable camouflage skills. The octopus has a textured skin that mirrors the multifaceted surface of the surrounding coral, with shades of brown and tan that blend into the underwater landscape. Its body is contoured to the shape of the coral it rests upon, illustrating its ability to adapt its posture to diverse environments. Suckers are visible along its tentacles, hinting at its capacity for tactile sensation and grip. In the background, we see a contrast between the rough textures of the nearby coral and the smooth, distant forms of **coral structures, creating a layered perspective of the reef habitat.** This scene is a testament to the octopus's mastery of disguise and the dynamic biodiversity of coral reef ecosystems.



In this enchanting underwater snapshot, **two distinct fish are depicted in an aquarium setting**, each illuminated by a tranquil blue light. The first is a striking **black and white striped fish**, notable for its elongated, flowing tail, elegantly swimming through the clear blue waters. Complementing this is another sizable fish, captured in the same serene environment, its form highlighted by the blue lighting that reveals a **pointed nose and a lengthy tail**, suggesting a sense of graceful motion within the aquatic space. Together, these fish, each with their own unique features and illuminated by a calming blue glow, embody the dynamic and diverse life that thrives beneath the surface.

The image captures **a single fish with striking black and white stripes**, navigating the tranquil blue waters of its aquatic environment. The fish's fins are well-spread, indicating either a state of gliding or active swimming. Its body shape is angular and laterally compressed, typical of many tropical fish species found in coral reefs, which aids in quick maneuvering through complex underwater structures. The **eyes of the fish are large and prominent**, suggesting a high visual acuity for detecting predators or prey. The overall composition, with the contrasting colors of the fish against the deep blue background, creates a vivid and dynamic marine scene.

Semantic instance captions | MarineInst+GPT-3.5 | GPT-4V

Fig. 12: We optimize our MarineInst to generate comprehensive and detailed semantic instance captions for each generated instance mask. Then we utilize ChatGPT-3.5 to generate the merged caption as the image-level caption based on the generated instance captions. GPT-4V is included for comparison, where texts in green are correct responses and red are wrong responses.

derstanding, we first construct 1,000 testing images and corresponding human-constructed reference captions (describing appearance, pose, activity, event, and

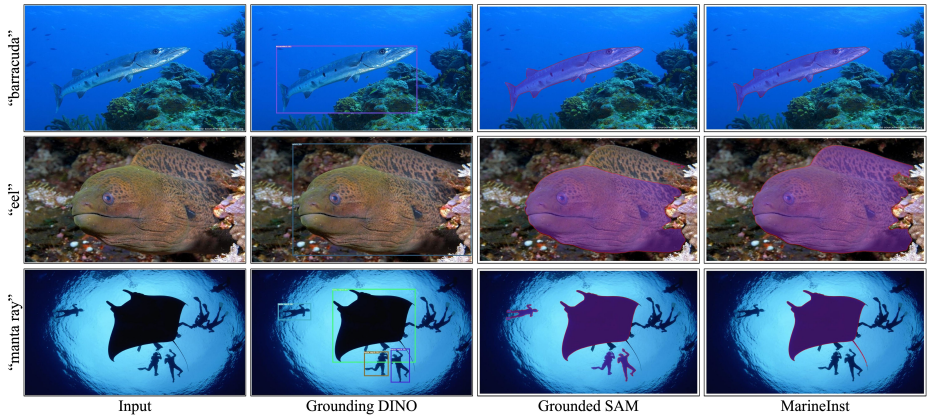


Fig. 13: We compare our MarineInst with Grounded SAM for instruction-following segmentation. The generated bounding boxes from Grounding DINO and corresponding confidence scores have also been provided for better illustration.

other attributes) for the whole marine image. The average word length of reference captions is *53.78*. The general-purpose MiniGPT-4 and domain-specific MarineGPT were included for comparison. The quantitative results are reported in Table 5. Through combining the semantic instance captions yielded by MarineInst, we could generate comprehensive image captioning and achieved better image storytelling performance than MiniGPT-4 and MarineGPT.

3.6 Instruction-following Segmentation

Instruction-following segmentation is a powerful technique that combines the precision of segmentation with the guidance of instructions from humans. Interpreting textual commands allows computer vision systems to identify and segment the specific objects or regions of interest within an image accurately. It also enables the extraction of precise anatomical structures indicated by marine professionals, aiding in species identification and environmental monitoring for marine research. In this work, MarineInst facilitates more sophisticated capabilities by allowing the model to understand and delineate the specific elements within the visual field. The concurrent Grounded SAM [69] (Grounding DINO [58] + SAM [47]) could also generate the desired mask based on the text queries. Grounded SAM is first asked to generate the bounding box based on the text prompt and then the generated bounding boxes from Grounding DINO are regarded as box prompts for SAM to generate instance masks with semantics. The semantics of Grounded SAM inherit the text prompts from users. We provide the qualitative comparison between Grounded SAM and our MarineInst in Figure 13. The object detection results of Grounding DINO have also been provided. Grounding DINO could generate reasonable object detection results for the "barracuda" and "eel", but produces false positives for the "manta ray".

Table 6: Instance segmentation results of various algorithms under settings: "Ⓐ" - Automatic; "★" - Point; "▣" - BBOX. **SAM-F**: fine-tuned SAM on our MarineInst20M dataset. – indicates that the results cannot be computed.

Method	binary instance filtering	human-annotated instance masks	model-generated instance masks	AP ↑		AP_s ↑		AP_m ↑		AP_l ↑	
				bbox	segm	bbox	segm	bbox	segm	bbox	segm
SAM Ⓐ [47]	✘	✘	✘	5.9	5.8	0.3	0.4	3.2	3.5	15.2	14.7
SAM-F Ⓐ	✘	✓	✘	$23.0_{+17.1}$	$24.8_{+19.0}$	$5.9_{+5.6}$	$6.8_{+6.4}$	$22.2_{+19.0}$	$25.5_{+22.0}$	$33.3_{+18.1}$	$33.7_{+19.0}$
MarineInst Ⓐ	✓	✓	✘	$28.2_{+22.3}$	$30.1_{+24.3}$	$7.2_{+6.9}$	$8.3_{+7.9}$	$29.9_{+26.7}$	$33.4_{+29.9}$	$37.0_{+21.8}$	$37.7_{+23.0}$
SAM-F Ⓐ	✘	✓	✓	$24.0_{+18.1}$	$25.8_{+20.0}$	$6.0_{+5.7}$	$7.0_{+6.6}$	$22.7_{+19.5}$	$26.0_{+22.5}$	$34.8_{+19.6}$	$35.4_{+20.7}$
MarineInst Ⓐ	✓	✓	✓	$30.8_{+24.9}$	$32.7_{+26.9}$	$7.6_{+7.3}$	$8.8_{+8.4}$	$32.1_{+28.9}$	$35.5_{+32.0}$	$40.2_{+25.0}$	$40.8_{+26.1}$
SAM ★ [47]	✘	✘	✘	59.0	63.0	64.3	77.8	70.2	77.3	48.5	47.4
SAM-F ★	✘	✓	✘	$67.9_{+8.9}$	$70.7_{+7.7}$	$71.2_{+6.9}$	$81.2_{+3.4}$	$78.7_{+8.5}$	$83.8_{+6.5}$	$57.8_{+9.3}$	$56.8_{+9.4}$
MarineInst ★	✓	✓	✘	$69.9_{+10.9}$	$72.5_{+9.5}$	$74.2_{+9.9}$	$84.1_{+6.3}$	$79.7_{+9.5}$	$84.7_{+7.4}$	$60.5_{+12.0}$	$59.2_{+11.8}$
SAM-F ★	✘	✓	✓	$71.6_{+12.6}$	$74.2_{+11.2}$	$73.8_{+9.5}$	$83.9_{+6.1}$	$81.3_{+11.1}$	$85.9_{+8.6}$	$62.8_{+14.3}$	$61.9_{+14.5}$
MarineInst ★	✓	✓	✓	$73.1_{+14.1}$	$75.4_{+12.4}$	$77.5_{+13.2}$	$86.6_{+8.8}$	$82.5_{+12.3}$	$86.7_{+9.4}$	$64.1_{+15.6}$	$62.8_{+15.4}$
SAM ▣ [47]	✘	✘	✘	–	93.5	–	95.3	–	95.7	–	92.2
SAM-F ▣	✘	✓	✘	–	$94.6_{+1.1}$	–	$95.8_{+0.5}$	–	$96.3_{+0.6}$	–	$93.4_{+1.2}$
MarineInst ▣	✓	✓	✘	–	$94.8_{+1.3}$	–	$96.1_{+0.8}$	–	$96.3_{+0.6}$	–	$93.8_{+1.6}$
SAM-F ▣	✘	✓	✓	–	$95.1_{+1.6}$	–	$96.0_{+0.7}$	–	$96.8_{+1.1}$	–	$93.9_{+1.7}$
MarineInst ▣	✓	✓	✓	–	$95.4_{+1.9}$	–	$96.4_{+1.1}$	–	$97.3_{+1.6}$	–	$93.8_{+1.6}$

Meanwhile, even with reasonable object detection results, SAM cannot always obtain precise instance mask predictions (refer to the "eel" case). Grounded SAM still struggles with error accumulation, where the false positive issues cannot be addressed in Grounded SAM. In contrast, MarineInst could generate reasonable mask outputs, which align the user instructions well. Furthermore, we provide a quantitative comparison with Grounded SAM based on our constructed 1,000 instruction-mask testing pairs. We adopt the IoU (binary segmentation: foreground object instances and backgrounds) as the evaluation metric. Grounded SAM achieved **23.32** while MarineInst got **39.65** in terms of IoU score. MarineInst demonstrates a stronger ability to understand the text prompts and recognize the described marine creatures. Finally, MarineInst is performing the instruction-following segmentation in an end-to-end manner. Our MarineInst, with its fusion of language and vision, paves the way for more efficient and versatile computer vision systems across various applications.

3.7 Ablation Studies

In this section, we aim to provide more analysis of the ablation studies, evaluating the effectiveness of the binary instance filtering and the model-generated annotations. The quantitative results are reported in Table 6. We directly fine-tune SAM (denoted as **SAM-F**) on our MarineInst20M as a comparison. Two experimental settings are designed: 1) **human-annotated**: only the human-annotated instance masks (**1.89M** instance masks) are used for training and 2) **human-annotated+model-generated**: both human-annotated and model-generated instance masks are used for fine-tuning. Please note that we also utilize the non-instance masks (**0.76M** non-instance masks under the "**human-annotated**"

Table 7: Quantitative comparisons between Mask R-CNN [36] and MarineInst on the UIIS dataset.

Methods	mAP↑	AP_{50}↑
Mask R-CNN [36]	23.3	40.8
MarineInst	**26.6**	**43.4**

setting and **11.7M** non-instance masks under the "**human-annotated+model-generated**" setting) for optimizing MarineInst with the binary instance filtering. As illustrated in Table 6, fine-tuning SAM (SAM-F) on our MarineInst20M dataset with redundant instance masks could lead to observable performance gains under all the settings. Furthermore, with the proposed binary instance filtering, our MarineInst could achieve better performance gains than SAM-F by effectively alleviating the over-segmentation and partial-segmentation. Especially, MarineInst has achieved larger performance improvements over SAM-F under the automatic setting. Meanwhile, with the non-instance masks together, the ability of MarineInst to generate precise masks with point or box prompts could also be slightly promoted. By comparing the performance under the "**human-annotated**" and "**human-annotated+model-generated**" settings, we conclude that the model-generated instance masks are also valuable in promoting the zero-shot instance segmentation ability, leading to a stronger marine foundation model for instance segmentation.

3.8 Comparison with Mask R-CNN

The traditional Mask R-CNN [36] could also perform effective instance segmentation. We provide additional analysis and comparison between Mask R-CNN and MarineInst to explore the effectiveness driven by large-scale datasets and powerful foundation models. We compare Mask R-CNN on the UIIS dataset by customizing MarineInst to *fixed-category* instance segmentation by extending binary instance filtering to multiple category classification. We continuously fine-tune MarineInst on the UIIS dataset with the instance masks with semantic category annotations. Following the experimental setting of [53], we report the experimental results in Table 7. The results confirm that our method outperforms Mask R-CNN. We attribute the performance improvement to the redundant mask annotations during the pre-training procedure and a more powerful network backbone. MarineInst demonstrates a stronger ability to perform precise instance segmentation.

3.9 More Results

In this section, we present more qualitative results as follows:
Comparison with SOTAs. Figure 14 illustrates more result comparisons between our MarineInst and existing SOTA algorithms. As illustrated, our Marine-

Inst could generate reliable and accurate instance masks with comprehensive and detailed semantic descriptions.

Head-to-head comparison with SAM. Figure 15 shows more automatic instance mask generation performance of our MarineInst on marine images. SAM is included for head-to-head comparison. As demonstrated in Figure 15, MarineInst could effectively alleviate the over-segmentation and partial-segmentation issues, leading to more effective instance segmentation.

Instruction-following instance understanding. Figure 16 presents more results of MarineInst on the instruction-following instance understanding task. We present the results under both settings: 1) *single mask* and 2) *multiple masks with assigned mask IDs*.

Hallucination. Figure 17 reports the failure cases (hallucination) of our MarineInst on generating semantic captions for the instance mask. The generated instance captions may not reflect the content of the image due to the hallucinations. Our studies indicate that the instance mask based image croppings cannot guarantee a satisfactory performance when there are multiple instances crowded within the same image cropping. Furthermore, the VLM also tends to describe the foreground objects rather than the background environments. We believe that further instruction-following instance understanding could help alleviate such hallucinations.

4 Discussions

4.1 Failure Cases and Generalization Ability

Failure cases. There are still some failure cases in MarineInst. MarineInst struggles with crowded scenes (*e.g.*, a school of tiny fish, making it difficult to define the separated instances); and the objects in the shadow and with low visibility and self-occlusions. We illustrate some failure cases in Figure 18(a).

Generalization ability to terrestrial images in Figure 18(b). We evaluate whether MarineInst could generate accurate instance masks for the terrestrial images (no overlapping with our MarineInst20M). As illustrated, MarineInst has demonstrated a satisfactory generalization ability to the animals on land, generating precise instance masks by learning some shared common sense knowledge.

4.2 Contribution Claim

To the best of the knowledge of the authors, our MarineInst is the first attempt to automatically generate instance masks with detailed and comprehensive semantic instance captions, describing the appearance, textures, pose, activity, and other attributes of the instance. The formulated instance visual description is a pioneering attempt toward dense instance understanding within the images. We utilize the powerful VLMs for more detailed and comprehensive captions and harness the power of the LLM. The generated captions are not limited to providing the category-level semantics, but also the semantic understanding from

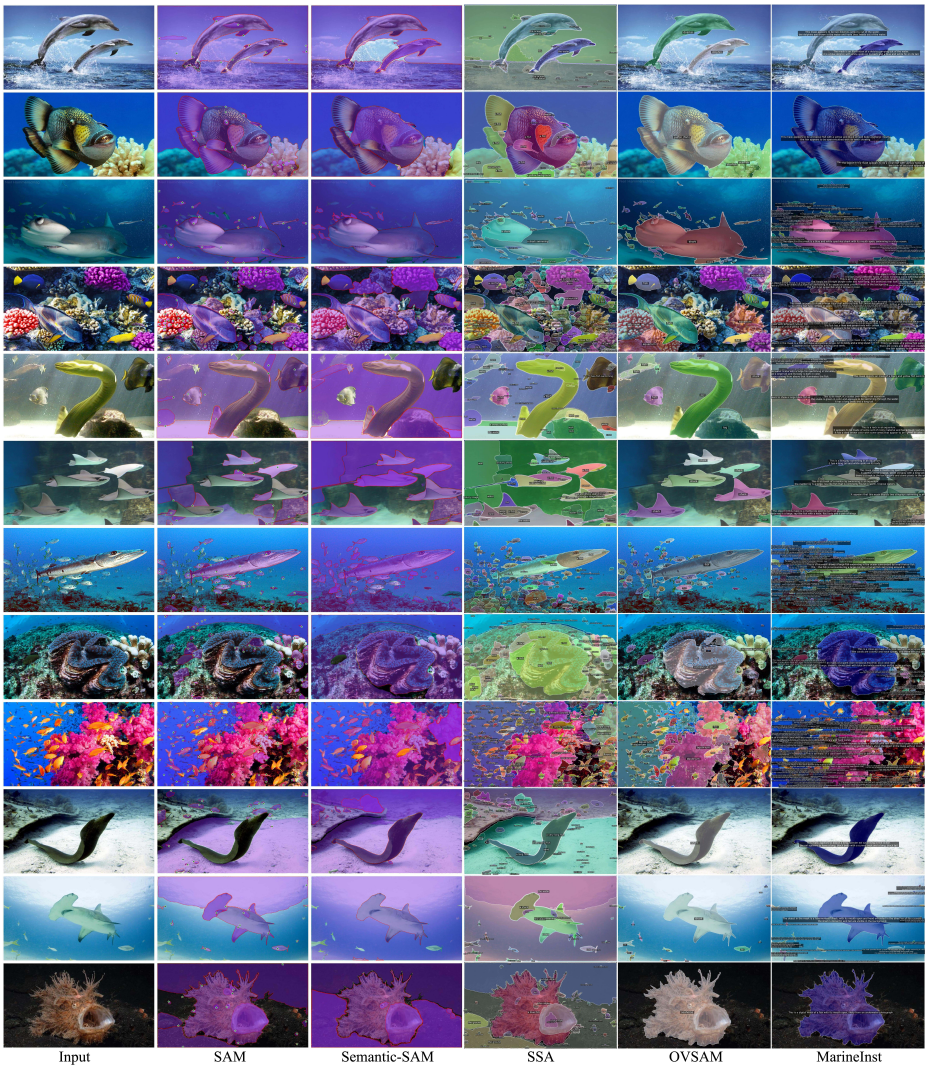


Fig. 14: Comparison between MarineInst and the existing SOTA algorithms. Both SAM and MarineInst generate masks based on automatically generated grid points. SSA yields semantic predictions based on the automatically generated masks by SAM. We set the semantic granularities of Semantic-SAM to 3. OVSAM is inferred by the box prompts. Please zoom in to see more details. `Results shown are not cherry picked.`

various semantic granularities. The generated instance masks could be utilized for marine species identification, object counting, coverage estimation, species interaction, and robotics applications to name a few. Our MarineInst emphasizes both the *data* and *modeling* approaches: we assemble the largest pre-training



Fig. 15: Head-to-head comparison between SAM and MarineInst on instance mask generation. The ⋆ indicates the automatically generated grid point prompts. SAM suffers from over-segmentation and partial-segmentation issues, generating redundant meaningless masks. MarineInst demonstrates a stronger ability than SAM on instance mask generation. `Results shown are not cherry picked.`

dataset for marine visual analysis, as well as propose a powerful and flexible marine foundation model. Our dataset and foundation model vastly improve the monitoring and study of marine ecosystems. Our data collection strategies and

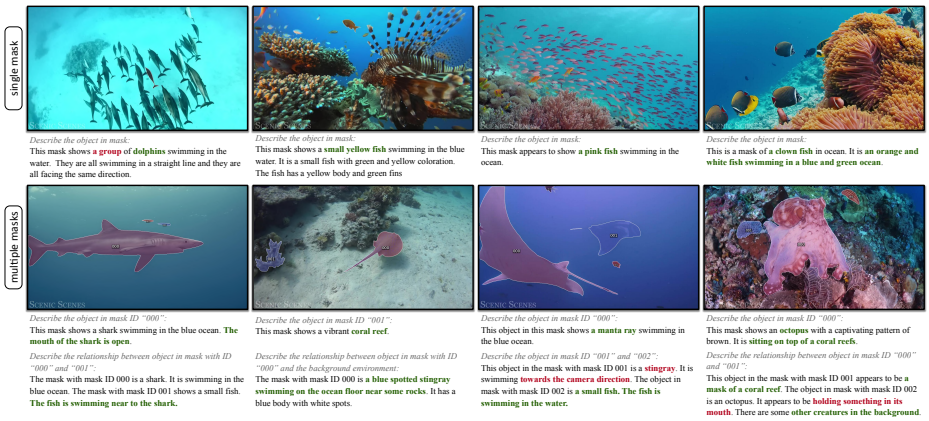


Fig. 16: The instruction-following instance understanding results of MarineInst under two settings: 1) *single mask* and 2) *multiple masks with assigned mask IDs*. The texts in green are correct responses and the texts in red are wrong responses.

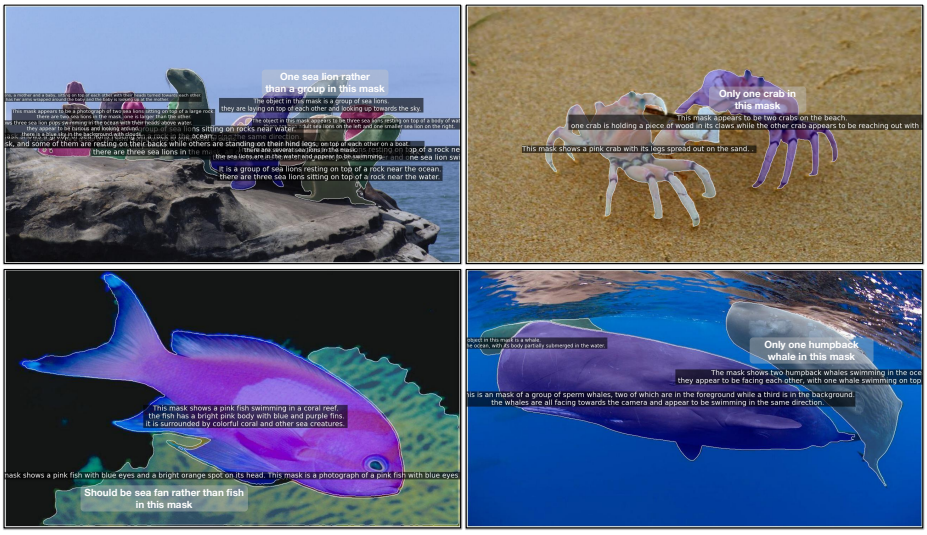


Fig. 17: There are still some hallucinations in the generated semantic captions for the instance mask. Best viewed in color.

model design could also be extended to other domains as well, providing valuable insights for the computer vision community.

In this work, we do not propose specially designed model modifications to explicitly address the intrinsic challenges of marine images (*e.g.*, low visibility [60] and color distortion [15]). In contrast, we aim to alleviate the data distribution

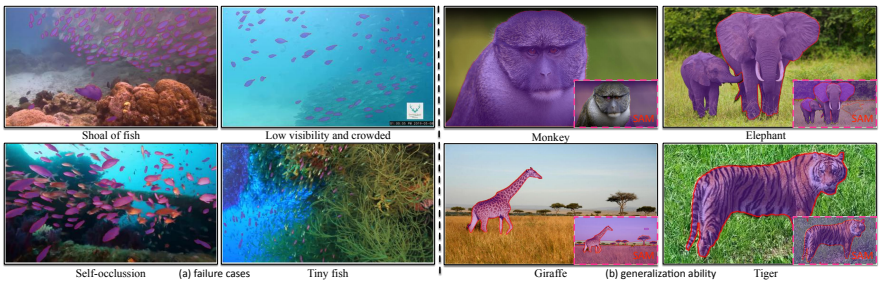


Fig. 18: (a) Our failure cases. (b) Generalization ability of MarineInst to terrestrial images. The bottom-right inline image shows results from SAM.

shift challenge in a data-driven manner, assembling the largest marine image dataset for training our foundation model to promote its generalization ability and robustness. Furthermore, we do not choose image enhancement as our key task since it cannot provide semantic understanding for marine images or instances within the images.

Potential broader impact. MarineInst not only transforms how the existing algorithms analyze and interpret images but also presents new opportunities and challenges for both marine and computer vision communities. MarineInst could also serve as a teaching aid, providing visual aids and analyses across subjects like biology, enhancing monitoring capabilities, and enabling more sophisticated and efficient image captioning, manipulation, and synthesis.

Limitation and future work. **Hallucinations**. There are still some hallucinations in the generated semantic captions from MarineInst. Multiple instances within one image cropping lead to inaccurate captions for the required instance mask. We believe further instruction-following instance understanding could help alleviate the hallucinations even based on noisy data for training. We leave more accurate instance captioning as our future work.

4.3 Related Works

Utilizing CLIP for open-vocabulary tasks. As CLIP models [66] are optimized contrastively at the global image-text level, they cannot directly output the dense predictions (*e.g.*, bounding box, and segmentation predictions) at the region or pixel level. Recent works [35,54,59,65,81,91] demonstrate the feasibility of adopting and locking a pre-trained CLIP model for open-vocabulary object detection and image segmentation tasks. The open-vocabulary setting suits the general open-world visual perception, where the target of interest is recognized based on a natural language description. RegionCLIP [46] proposed to perform the regional visual feature and the textual conception alignment to promote the generalization ability to unseen categories. SegCLIP [59] proposed to perform open-vocabulary segmentation in an annotation-free manner by gathering patches with learnable centers to semantic regions. MaskCLIP [31] alters the

last pooling layer of CLIP to produce dense predictions and utilize the generated pseudo-labels to train a segmentation model. MasQCLIP [81] then further proposed to perform knowledge distillation through the self-training with pseudo labels. However, these algorithms are mainly optimized and evaluated on the in-air datasets, such as COCO [55] and LVIS [34]. Our MarineInst presents the first attempt to perform open-vocabulary instance segmentation in the marine field. Another key difference between our MarineInst over the existing open-vocabulary dense prediction algorithms is that MarineInst performs open-ended semantic instance caption generation while existing algorithms mainly focus on category-level semantics.

SAM and SAM Variants. SAM has been widely used for medical image segmentation [79], satellite images [24,68], remote sensing [77], camouflaged object segmentation [43,73], challenging scenarios [25] and other applications [67,83,92]. MSA [79] designed an adapter design for transferring SAM to a counterpart in segmenting medical images and SAM-adapter [25] proposed to perform camouflaged object segmentation and shadow detection. ClassWise-SAM-Adapter [64] proposed the class-wise adapter to perform the semantic segmentation. However, the number of adapters is subject to the number of semantic classes. RSPrompter [24] proposed to fine-tune SAM to the satellite images and perform instance segmentation through the designed prompt, while the object instances are quite limited. However, these algorithms failed to generate instance masks and still require prompts from the users for required mask generation.

VLM. BLIP series [50, 51] bootstrap vision-language pre-training from frozen pre-trained image encoders and frozen language decoders. Based on BLIP-2 [50], MiniGPT-4 [93] proposed a projection layer to align pre-trained vision encoder to frozen LLMs, and exhibited respectable zero-shot image comprehension in dialogues. LLaVA [57] aimed to optimize the linear layer based on the constructed instruction-follow data. MarineGPT [89] is the first vision-language model in the marine field. It is further optimized by the domain-specific data and demonstrates a strong zero-shot recognition ability for marine data. Moreover, MarineGPT could usually generate more detailed and comprehensive captions than BLIP series [50,51], by utilizing a frozen LLM. Our MarineInst proposes to harness the power of VLM for generating detailed and comprehensive instance captions.

4.4 Future Directions

Video understanding. Currently, MarineInst mainly focuses on the image-level instance visual description. How to extend our MarineInst to video field [86] for temporal understanding will be our future work.

3D reconstruction and scene understanding. Our method could also be utilized for promoting the 3D reconstruction and scene understanding [26]. We could employ our MarineInst to perform instance visual descriptions and then utilize the instance masks for highlighting the foreground objects to obtain high-quality point clouds. In this way, we can effectively bridge the 3D semantic gap at the instance level.

Underwater enhancement. The generated instance masks could also be utilized for promoting the underwater image enhancement performance by highlighting the foreground objects with precise boundaries. The automatic instance mask generation and underwater image enhancement could formulate a mutually beneficial system.
Instance-level VLM. In this work, we demonstrate that our MarineInst could be utilized for instruction-following tasks, including both instruction-following instance understanding and segmentation. With such instance-level instruction-following data, we can further optimize the instance-level VLM. A dataset with instance-level captions could be generated using our pre-trained models. A new set of evaluation benchmarks and metrics for measuring the performance of instance understanding are required.
Controllable image synthesis. Similarly, MarineInst produces valuable and scalable training data for controllable image synthesis [78]. The precise localization and comprehensive semantic captions will guide the model for better controllable image synthesis performance.
Spatial reasoning. Spatial reasoning capability [22] not only empowers the model with common sense knowledge about object sizes but also makes it useful for interaction tasks. To achieve this, the spatial localization and the instance captions of the instance masks are valuable for constructing the instruction-following data to enable spatial chain-of-thought for solving complex spatial reasoning tasks.

References

1. https://reeflifesurvey.com/
2. https://www.reeflex.net/
3. Aquarium dataset. https://public.roboflow.com/object-detection/aquarium
4. Flickr. https://www.flickr.com/
5. Getty images. https://www.gettyimages.com/
6. Hk reef fish. https://www.114ehkreeffish.org/
7. Shutterstock. https://www.shutterstock.com/
8. Underwater trash detection dataset. https://universe.roboflow.com/071bct525prasanga-pcampus-edu-np/underwater-trash-detection-4i8eg
9. Encyclopedia of life. http://eol.org (2018)
10. Sea animal image dataset. https://www.kaggle.com/datasets/vencerlanz09/sea-animals-image-dataste (2018)
11. Ozfish dataset - machine learning dataset for baited remote underwater video stations (2019)
12. Urpc dataset. https://openi.pcl.ac.cn/OpenOrcinus_orca/URPC2020_dataset/datasets (2020)
13. Fishnet open images database. https://www.fishnet.ai/ (2022)
14. Oceanic life dataset. https://www.kaggle.com/datasets/cyanex1702/oceanic-life-dataset (2023)
15. Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 1682–1691 (2019)

16. Alawode, B., Guo, Y., Ummar, M., Werghi, N., Dias, J., Mian, A., Javed, S.: Utb180: A high-quality benchmark for underwater tracking. In: Asian Conference on Computer Vision (ACCV). pp. 3326–3342 (2022)
17. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
18. Beijbom, O., Edmunds, P.J., Roelfsema, C., Smith, J., Kline, D.I., Neal, B.P., Dunlap, M.J., Moriarty, V., Fan, T.Y., Tan, C.J., et al.: Towards automated annotation of benthic survey images: Variability of human experts and operational modes of automation. PloS one **10**(7), e0130312 (2015)
19. Boom, B.J., Huang, P.X., He, J., Fisher, R.B.: Supporting ground-truth annotation of image datasets using clustering. In: International Conference on Pattern Recognition (ICPR). pp. 1542–1545. IEEE (2012)
20. Bovcon, B., Muhovič, J., Perš, J., Kristan, M.: The mastr1325 dataset for training deep usv obstacle detection models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3431–3438. IEEE (2019)
21. Bray, D.J. & Gomon, M.e.: Fishes of australia. `http://fishesofaustralia.net.au/` (2018)
22. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. arXiv preprint arXiv:2401.12168 (2024)
23. Chen, J., Yang, Z., Zhang, L.: Semantic segment anything. `https://github.com/fudan-zvg/Semantic-Segment-Anything` (2023)
24. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Transactions on Geoscience and Remote Sensing (TGRS) (2024)
25. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023)
26. Chen, Z., Li, B.: Bridging the domain gap: Self-supervised 3d scene understanding with foundation models. arXiv preprint arXiv:2305.08776 (2023)
27. Cheng, Y., Zhu, J., Jiang, M., Fu, J., Pang, C., Wang, P., Sankaran, K., Onabola, O., Liu, Y., Liu, D., Bengio, Y.: Flow: A dataset and benchmark for floating waste detection in inland waters. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10953–10962 (October 2021)
28. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., et al.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna. lmsys. org (2023)
29. Chin, C.: Marine fouling images (2019). `https://doi.org/10.21227/k07g-3t57`, `https://dx.doi.org/10.21227/k07g-3t57`
30. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. Ieee (2009)
31. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10995–11005 (2023)
32. E., T., L.M, D.: Corals of the world. `http://coralsoftheworld.org/v0.01(Beta)` (2016)

33. Fulton, M., Hong, J., Islam, M.J., Sattar, J.: Robotic detection of marine litter using deep visual detection models. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 5752–5758. IEEE (2019)
34. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5356–5364 (2019)
35. Haixin, L., Ziqiang, Z., Zeyu, M., Yeung, S.K.: Marinedet: Towards open-marine object detection. arXiv preprint arXiv:2310.01931 (2023)
36. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
37. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)
38. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural information processing systems (Neurips) **30** (2017)
39. Hong, J., Fulton, M., Sattar, J.: Trashcan: A semantically-segmented dataset towards visual detection of marine debris. arXiv preprint arXiv:2007.08097 (2020)
40. Hong, L., Wang, X., Zhang, G., Zhao, M.: Usod10k: a new benchmark dataset for underwater salient object detection. IEEE Transactions on Image Processing (TIP) (2023)
41. Islam, M.J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S.S., Sattar, J.: Semantic segmentation of underwater imagery: Dataset and benchmark. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1769–1776. IEEE (2020)
42. Islam, M.J., Wang, R., Sattar, J.: Svam: saliency-guided visual attention modeling by autonomous underwater robots. Robotics: Science and Systems (2022)
43. Ji, G.P., Fan, D.P., Xu, P., Cheng, M.M., Zhou, B., Van Gool, L.: Sam struggles in concealed scenes–empirical study on segment anything. arXiv preprint arXiv:2304.06022 (2023)
44. Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B., et al.: Fathomnet: A global image database for enabling artificial intelligence in the ocean. Scientific reports **12**(1), 15914 (2022)
45. Khan, F.F., Li, X., Temple, A.J., Elhoseiny, M.: Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20496–20506 (2023)
46. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11144–11154 (2023)
47. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
48. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023)
49. Li, F., Zhang, H., Sun, P., Zou, X., Liu, S., Yang, J., Li, C., Zhang, L., Gao, J.: Semantic-sam: Segment and recognize anything at any granularity. arXiv preprint arXiv:2307.04767 (2023)
50. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. International Conference on Machine Learning (ICML) (2023)

51. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML). pp. 12888–12900. PMLR (2022)
52. Li, L., Dong, B., Rigall, E., Zhou, T., Dong, J., Chen, G.: Marine animal segmentation. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **32**(4), 2303–2314 (2021)
53. Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S.: Watermask: Instance segmentation for underwater imagery. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1305–1315 (2023)
54. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7061–7070 (2023)
55. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
56. Liu, C., Wang, Z., Wang, S., Tang, T., Tao, Y., Yang, C., Li, H., Liu, X., Fan, X.: A new dataset, poisson gan and aquanet for underwater object grabbing. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **32**(5), 2831–2844 (2021)
57. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Neural Information Processing Systems (Neurips) (2023)
58. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
59. Luo, H., Bao, J., Wu, Y., He, X., Li, T.: Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: International Conference on Machine Learning (ICML). pp. 23033–23044. PMLR (2023)
60. Marques, T.P., Albu, A.B.: L2uwe: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In: IEEE/CVF conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 538–539 (2020)
61. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
62. Pedersen, M., Haurum, J.B., Gade, R., Moeslund, T.B., Madsen, N.: Detection of marine animals in a new underwater dataset with varying visibility. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (June 2019)
63. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
64. Pu, X., Jia, H., Zheng, L., Wang, F., Xu, F.: Classwise-sam-adapter: Parameter efficient fine-tuning adapts segment anything to sar domain for semantic segmentation. arXiv preprint arXiv:2401.02326 (2024)
65. Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., et al.: Freeseg: Unified, universal and open-vocabulary image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19446–19455 (2023)

66. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
67. Rajič, F., Ke, L., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Segment anything meets point tracking. arXiv preprint arXiv:2307.01197 (2023)
68. Ren, S., Luzi, F., Lahrichi, S., Kassaw, K., Collins, L.M., Bradbury, K., Malof, J.M.: Segment anything, from space? In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8355–8365 (2024)
69. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
70. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (2023)
71. Saleh, A., Laradji, I.H., Konovalov, D.A., Bradley, M., Vazquez, D., Sheaves, M.: A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. Scientific Reports **10**(1), 14671 (2020). https://doi.org/https://doi.org/10.1038/s41598-020-71639-x
72. Sun, G., An, Z., Liu, Y., Liu, C., Sakaridis, C., Fan, D.P., Van Gool, L.: Indiscernible object counting in underwater scenes. In: IEEE/CVF International Conference on Computer Vision and Patern Recognition (CVPR) (2023)
73. Tang, L., Xiao, H., Li, B.: Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709 (2023)
74. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
75. Truong, Q.T., Vu, T.A., Ha, T.S., Lokoč, J., Wong, Y.H., Joneja, A., Yeung, S.K.: Marine video kit: a new marine video dataset for content-based analysis and retrieval. In: International Conference on Multimedia Modeling (MMM). pp. 539–550. Springer (2023)
76. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
77. Wang, D., Zhang, J., Du, B., Tao, D., Zhang, L.: Scaling-up remote sensing segmentation dataset with segment anything model. arXiv preprint arXiv:2305.02034 (2023)
78. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. arXiv preprint arXiv:2402.03290 (2024)
79. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
80. Xu, B., Liang, H., Liang, R., Chen, P.: Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In: Association for the Advancement of Artificial Intelligence (AAAI). vol. 35, pp. 3004–3012 (2021)
81. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 887–898 (2023)

82. Yang, L., Xu, Z., Zeng, H., Sun, N., Wu, B., Wang, C., Bo, J., Li, L., Dong, Y., He, S.: Fishdb: an integrated functional genomics database for fishes. BMC genomics **21**(1), 1–5 (2020)
83. Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., Chen, Z.: Inpaint anything: Segment anything meets image inpainting. arXiv preprint arXiv:2304.06790 (2023)
84. Yuan, H., Li, X., Zhou, C., Li, Y., Chen, K., Loy, C.C.: Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. arXiv preprint (2024)
85. Zhang, C., Cong, R., Lin, Q., Ma, L., Li, F., Zhao, Y., Kwong, S.: Cross-modality discrepant interaction network for rgb-d salient object detection. In: ACM international conference on multimedia (ACM MM). pp. 2094–2102 (2021)
86. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024)
87. Zhao, Z., Xia, C., Xie, C., Li, J.: Complementary trilateral decoder for fast and accurate salient object detection. In: ACM international conference on multimedia (ACM MM). pp. 4967–4975 (2021)
88. Zheng, Z., Ha, T.S., Chen, Y., Liang, H., Chui, A.P.Y., Wong, Y.H., Yeung, S.K.: Marine video cloud: A cloud-based video analytics platform for collaborative marine research. In: OCEANS. pp. 1–6. IEEE (2023)
89. Zheng, Z., Zhang, J., Vu, T.A., Diao, S., Tim, Y.H.W., Yeung, S.K.: Marinegpt: Unlocking secrets of ocean to the public. arXiv preprint arXiv:2310.13596 (2023)
90. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16793–16803 (2022)
91. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision (ECCV). pp. 696–712. Springer (2022)
92. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can sam segment polyps? arXiv preprint arXiv:2304.07583 (2023)
93. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
94. Zhuang, P., Wang, Y., Qiao, Y.: Wildfish++: A comprehensive fish benchmark for multimedia research. IEEE Transactions on Multimedia (TMM) **23**, 3603–3617 (2020)
95. Ziqiang, Z., Tan-Sang, H., Yingshu, C., Haixin, L., Apple Pui-Yi, C., Yue-Him, W., Sai-Kit, Y.: Marine video cloud: A cloud-based video analytics platform for collaborative marine research (2023)
96. Ziqiang, Z., Yaofeng, X., Haixin, L., Zhibin, Y., Yeung, S.K.: Coralvos: Dataset and benchmark for coral video segmentation. arXiv preprint arXiv:2310.01946 (2023)
97. Žust, L., Perš, J., Kristan, M.: Lars: A diverse panoptic maritime obstacle detection dataset and benchmark. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)