ORCA: Object Recognition and Comprehension for Archiving Marine Species

Yuk-Kwan Wong[1] Haixin Liang [1] Zeyu Ma[2] Yiwei Chen[1] Ziqiang Zheng[1*]
Rinaldi Gotama[3] Pascal Sebastian[3] Lauren D. Sparks[3] Sai-Kit Yeung[1]

[1]Hong Kong University of Science and Technology
[2]University of Electronic Science and Technology of China [3]Indo Ocean Foundation

Project website: https://orca.hkustvgd.com

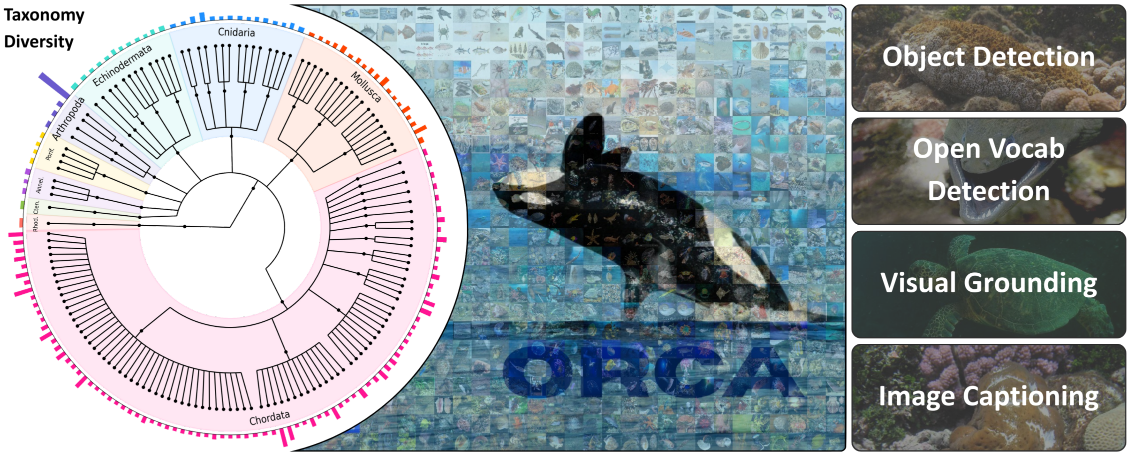


Figure 1. ORCA includes a spectrum of taxonomic marine species regarding both diversity and coverage. The detailed instance-level annotations (BBOX and dense captions enriched with domain-specific knowledge) enable object recognition and further comprehension for archiving the marine species, also supporting various biological applications.

Abstract

Marine visual understanding is essential for monitoring and protecting marine ecosystems, enabling automatic and scalable biological surveys. However, progress is hindered by limited training data and the lack of a systematic task formulation that aligns domain-specific marine challenges with well-defined computer vision tasks, thereby limiting effective model application. To address this gap, we present ORCA, a multi-modal benchmark for marine research comprising 14,647 images from 478 species, with 42,217 bounding box annotations and 22,321 expert-verified instance captions. The dataset provides fine-grained visual and textual annotations that capture morphology-oriented attributes across diverse marine species. To catalyze methodological advances, we evaluate 18 state-of-the-art models on three tasks: object detection (closed-set and open-vocabulary), instance captioning, and visual grounding. Results highlight key challenges, including species diversity, morphological overlap, and specialized domain demands, underscoring the difficulty of marine understanding. ORCA thus establishes a comprehensive benchmark to advance research in marine domain.

[*] Corresponding author: zhengziqiang1@gmail.com

1. Introduction

The ocean, with a vast coverage on the surface of our blue planet, remains a mysterious abyssal region to the best of our knowledge. Advancing knowledge of marine ecosystems is critical for oceanography [12, 80], sustainable resource management [14, 48], and biodiversity conservation [37, 43]. Considerable efforts have been devoted to biological surveys and habitat monitoring [3, 67, 71, 78]. To enhance the scalability and efficiency of in-situ monitoring, researchers are increasingly leveraging computer vision techniques to reduce manual data processing, including image classification [9, 62, 79], object detection [73, 76], and vision–language modeling [27].

Despite the remarkable success enabled by powerful network backbones and domain-specific datasets, significant challenges remain, which can be broadly categorized into issues of **training data** and **task formulation**. Current marine datasets are often restricted to a small set of predefined categories (*e.g.*, seven semantic classes in the UIIS dataset [34]) and are typically collected from limited geographic regions. Such constraints hinder both taxonomic diversity and ecological coverage, limiting the recognition of a broad range of marine species. Even fine-grained datasets with larger category sets, such as [88, 89], remain primarily focused on fish monitoring.

Regarding task formulation, current task definitions remain insufficient for domain requirements. *Image-level classification* [9, 62, 79] may lead to inconsistency between coarse category annotations and image content, which often contains multiple species in a wild environment. *Object detection* [73, 76] is limited by a narrow set of categories. Furthermore, category labels alone cannot capture key biological traits that are essential for ecological monitoring. For *image captioning*, although recent vision–language models (VLMs) [29, 31, 70, 82] are evolving rapidly, their outputs are typically coarse and lack the granularity and domain-specific knowledge needed for description.

To address these challenges, we introduce ORCA, the first multimodel dataset explicitly designed for marine research. ORCA offers 1) **broad taxonomic coverage** spanning 478 species and 670 common-name categories; 2) **instance-level annotations** enabling both object detection and grounding; and 3) **biology-oriented captions** with diagnostic traits, appearances, behaviors, and habitats, all validated by marine biologists. The dataset comprises 14,647 images with 42,217 bounding boxes, each labeled with both scientific and common names to support diverse usage scenarios. In total, ORCA provides 22,321 expertly verified instance–caption pairs, ensuring terminological accuracy and scientific relevance.

ORCA supports a range of vision-language tasks, including closed-set and open-vocabulary detection, instance-level captioning, and visual grounding. While detection and grounding primarily assess a model's ability to recognize and localize marine species, ORCA further introduces three evaluation settings: *Class-Level*, *Intra-Class*, and *Inter-Class*, to systematically examine how taxonomic hierarchies influence those abilities under the condition of **morphological overlapping**, where closely related species exhibit highly similar traits, thereby complicating species identification. Beyond spatial localization, the captioning and grounding components of ORCA facilitate fine-grained alignment between visual observations and linguistic descriptions. This dual emphasis not only enhances object-level referencing but also supports the structured, biologically meaningful archiving of marine survey data.

We have benchmarked 18 state-of-the-art algorithms across the aforementioned tasks. In summary, our contributions can be outlined as follows:

- We present ORCA, the first large-scale marine dataset with broad taxonomic coverage, bounding box annotations, and rich instance-level captions.
- We conduct an evaluation of 18 models, showing that fine-tuning on ORCA improves performance on localization and captioning tasks.
- We demonstrate that dense, domain-specific captions enable accurate object referencing and resolve challenges posed by morphological overlap, where visual cues are ambiguous and misleading.

- We show that existing captioning models struggle with instance-level descriptions, often producing coarse, image-level captions instead of region-specific outputs.

2. Related Work

Existing marine research. Marine species exhibit high diversity in pose, appearance, and pattern. Robust marine visual understanding can leverage recent algorithms [19, 33, 84, 85] to advance research, conservation, and industry. Several datasets have been introduced, including MAS3K [32, 33], WildFish [88], WildFish++ [89], and SUIM [21], which improve recognition of marine organisms. However, most of them provide only a limited set of predefined categories without detailed captions, restricting their utility for fine-grained marine analysis and large-scale scientific databases. ORCA addresses this gap by introducing a large-scale dataset covering a broad range of marine species with high-quality annotations (bounding boxes and captions).

Object Detection. Object detection is a core computer vision task [36, 55, 56], involving simultaneous object localization and classification. Conventional one-stage [15, 42, 54] and two-stage [17, 55, 56] detectors rely on fixed predefined category sets, which limit their applicability in marine domains, where species diversity varies greatly across regions. Open-vocabulary object detection (OVOD) [24, 66, 72, 75] addresses this challenge by extending detection to unseen categories. OVOD commonly leverages large-scale vision–language pre-training [51] to align visual regions with textual concepts; for instance, RegionCLIP [24] enhances generalization by matching regional features with natural language. These properties make OVOD particularly promising in marine applications, with the ability to recognize novel and diverse species.

Vision–Language Understanding. VLMs [1, 29, 30, 39, 40, 63, 82, 83, 87] have made substantial progress, driven by large-scale datasets such as Visual Genome [26], VizWiz [16], RefCOCO [22], and Objects365 [59]. These models combine visual encoders [13] with large language models [45, 46], trained on massive image–text corpora. CLIP [51] demonstrated strong zero-shot recognition, while BLIP [29, 30] advanced multimodal pre-training through frozen encoder–decoder architectures. Collectively, these works provide the foundation for tasks such as image captioning and grounding, which are critical for automatically documenting and archiving marine observations and discoveries. However, most existing datasets focus on terrestrial objects with very limited marine coverage, restricting VLM effectiveness in this domain. Furthermore, current VLMs struggle with fine-grained, region-level instance understanding essential for marine-specific tasks. To address this gap, ORCA provides high-quality textual annotations to better enable VLM applications in marine research.

3. Orca Construction

We illustrate the construction protocol of ORCA in Figure 2 and subsequently summarize its characteristics and statistics.

3.1. Dataset Construction

Data collection. The process began by compiling a target list of marine taxonomic categories. GPT-4 was employed to generate canonical common names (*e.g.*, seahorse), providing a proxy for vernacular terms most widely used by the public and thereby guiding more effective image searches. Candidate images were then sourced from Google Images, Flickr, and iNaturalist, with URLs retained for copyright attribution. All images underwent manual inspection to remove duplicates and misclassified entries, ensuring both quality and diversity. Each common name was subsequently mapped to its corresponding taxon in the World Register of Marine Species (WoRMS) [2]. Cases where a common name referred to an entire genus or higher taxonomic rank (*e.g.*, "unicorn fish," encompassing the genus *Naso*) were excluded to avoid ambiguity.

Bounding-box annotation. We combined the Segment Anything Model (SAM) [25] with human-supplied point prompts to delineate object masks, which were subsequently converted to axis-aligned bounding boxes. Given the amorphous morphology of marine organisms, we specifically verified that each box fully encompassed the target instance, including translucent fins and slender appendages.

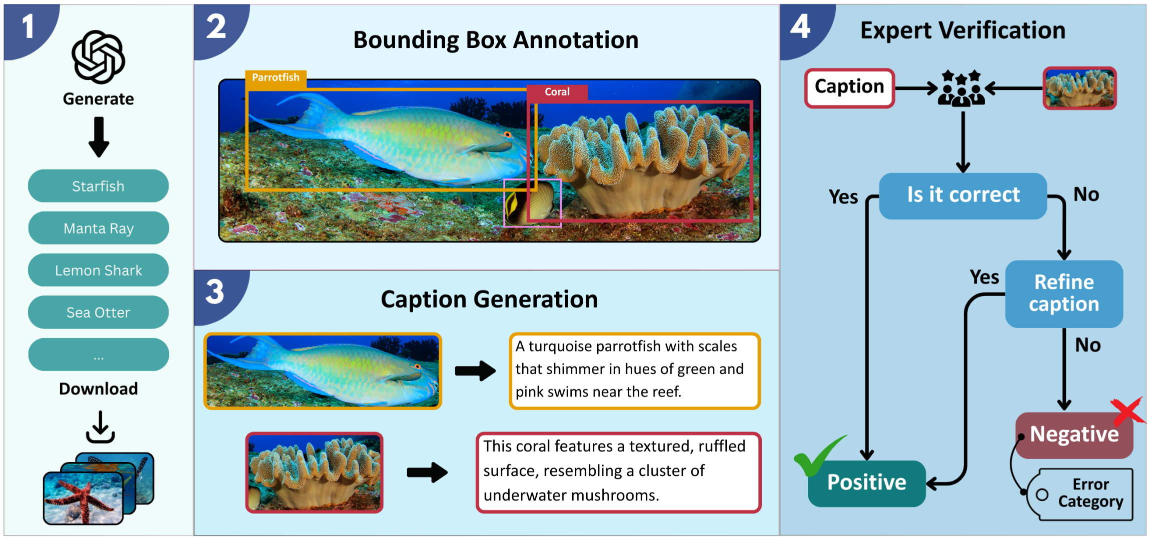


Figure 2. Overview of ORCA construction process. It begins with image collection, followed by bounding box annotation and caption generation for each box. Domain experts then verify all of them and refine at least one caption per image.

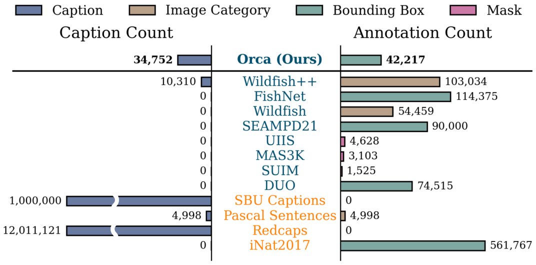


Figure 3. ORCA offers a balanced and sufficient amount of visual and textual annotations, compared to general and domain-specific datasets.

Dataset	Image Count	Visual Annotation	Lingistic Annotation	Category Count	Taxonomy Supported
DUO [38]	7,782	BBOX	-	4	✗
SUIM [21]	1,525	Mask	-	8	✗
MAS3K [32]	3,103	Mask	-	37	✗
UIIS [34]	4,628	Mask	-	7	✗
SEAMPD21 [5]	28,328	BBOX	-	130	✗
Wildfish [88]	54,459	Category	-	1,000	✗
FishNet [23]	94,532	BBOX	-	17,357	✓
Wildfish++ [89]	2,348	Category	Image-Level	2,348	✓
Redcaps [11]	12,011,121	-	Image-Level	-	✗
Pascal Sentences [53]	1,000	Category	Image-Level	20	✗
SBU Captions [47]	1,000,000	-	Image-Level	-	✗
iNat2017 [20]	859,000	BBOX	-	5,089	✓
Orca (Ours)	14,645	BBOX	**Instance-Level**	670	✓

Table 1. Statistic comparison with other general and domain-specific datasets.

Caption generation. Existing datasets [57] mainly utilized alt-texts to formulate the image-text pairs. However, the texts suffer from limited information (short captions), misalignment with visual contents, and deviation from domain-specific requirements. Instead, we generate rich instance-level descriptions. For every large bounding box ($> 1{,}024$ pixels), the cropped region is passed to MarineGPT [82] to produce captions tailored to the marine research.

3.2. Dataset Statistic and Comparison

Caption refinement. The generated caption is then passed to domain experts for verification and refinement along four dimensions: 1) *Unique morphological traits*, such as color, shape, injuries, *etc*; 2) *Spatial context* (absolute and relative positions); 3) *Environmental background*; and 4) *Behavioral cues* (individual or inter-species interactions). To enhance labeling efficiency, experts are required to refine at least one caption per image. The remaining are labeled *positive* if error-free or *negative* otherwise. We intentionally retained the *negative* captions, proving harder negatives than prior work that substitutes random nouns [74, 81]. Finally, ORCA contains 34,752 captions (with 12,873 *refined*, 9,448 *positive* and 12,431 *negative* captions). We further codify 11 error categories responsible for negative labels, where details are provided in the supplementary material.

Our dataset introduces domain-specific features that distinguish it from both general-purpose and

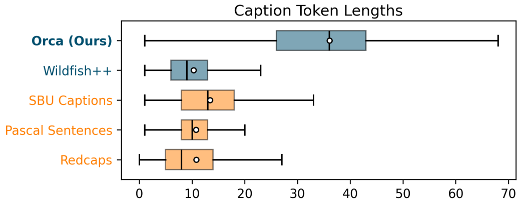
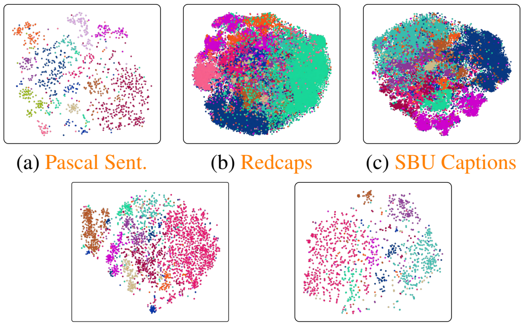


Figure 4. Caption tokens length for general datasets and domain-specific datasets. The white circle represents the mean caption length. Outliers have been filtered out.



(a) Pascal Sent. (b) Redcaps (c) SBU Captions

(d) ORCA (Ours) (e) Wildfish++

Figure 5. t-SNE of the vocabulary used in general datasets and domain-specific datasets). Pascal Sent. stand for Pascal Sentence for better visualization.

existing marine datasets: 1) it provides comprehensive instance-level annotations, with each bounding box larger than 1,024 pixels paired with a caption and mapped to marine taxonomic categories, as shown in Table 1; 2) it ensures balanced visual–textual supervision, offering comparable scales across both modalities to support a wide range of vision–language tasks, unlike other datasets that emphasis on one modality, as illustrated in Figure 3; and 3) it includes dense and diverse captions for each organism, yielding high caption density and substantial vocabulary diversity, as shown in Figure 4 and Figure 5 respectively.

4. Experiments

We benchmark 18 existing SOTA models on ORCA from three representative tasks, including object detection (closed-set and open-vocabulary settings), instance captioning, and visual grounding.

4.1. Object Detection

Experimental settings. We evaluate the capacity of closed-set and open-vocabulary object detection models to both localize and identify marine creatures. It is notoriously challenging, even for experienced biologists, because of the morphological overlap among species, where those belonging to the same higher-level taxon often exhibit similar physical characteristics. For OVOD, we devise three settings: *Class-Level*, *Intra-Class*, and *Inter-Class*.

Class-Level. We group species at "Class" level in the taxonomic hierarchy. Specifically, 670 vernacular categories are consolidated into 33 Class-level taxonomic categories, with 24 *seen* and 9 *unseen* categories. Certain vernacular categories (*e.g.*, *Bryozoa*) correspond to higher taxonomic ranks (*e.g.*, *phylum*) and are therefore excluded.

Intra-Class setting refines the task further by requiring models to identify vernacular categories within the aforementioned 33 Class-level taxonomic groups. From these, we sample 555 vernacular categories as *seen* and 109 as *unseen*.

Inter-Class. We adopt a more granular approach by sampling one vernacular category as *unseen* for every four categories within each "Class", while designating the remaining three vernacular categories as *seen*. "Classes" with fewer than four categories are excluded. As a result, this setup includes 482 *seen* and 161 *unseen* vernacular categories.

Close-set object detection. We mainly include 3 representative close-set object detection algorithms (Faster-RCNN [55], YOLOX [15], and GridR-CNN [44]) and report the mAP$_{50}$ of 24 *seen* categories under three settings. Our implementation of these models is based on MMDetection [6] using the official experimental setting. Please note that we do not evaluate these closed-set object detection algorithms on the *unseen* categories.

OVOD. We evaluate the performance of 3 open-vocabulary object detection algorithms (UniDetector [66], RegionCLIP [86], and DECOLA [7]). We follow the official experimental setting and fine-tune the model on our ORCA dataset. Particularly, we adopt the single-dataset training strategy for UniDetector [66] to continuously optimize it in an end-to-end fashion. For DECOLA [7], we utilize

Method	Seen			Unseen		
	Class-level	Intra-Class	Inter-Class	Class-level	Intra-Class	Inter-Class
FasterRCNN [55]	28.7	17.6	16.7	-	-	-
YOLOX [15]	27.5	21.7	21.0	-	-	-
GridRCNN [44]	32.7	28.1	28.6	-	-	-
UniDetector [66]	31.5	23.3	24.1	8.2	0.4	0.7
RegionCLIP [86]	39.8	34.1	29.8	12.2	6.2	0.4
DECOLA [7]	**66.7**	**88.8**	**86.9**	**37.7**	**51.6**	**52.3**

Table 2. Quantitative object detection results (mAP$_{50}$) under close-set and open-vocabulary settings.

their best-performing model with Swin-B backbone (phase 1) as the pre-trained model. We inherit the language-conditioned detection training procedure of DECOLA while keeping other configurations the same. We report the quantitative result in Table 2 where mAP$_{50}$ is computed.

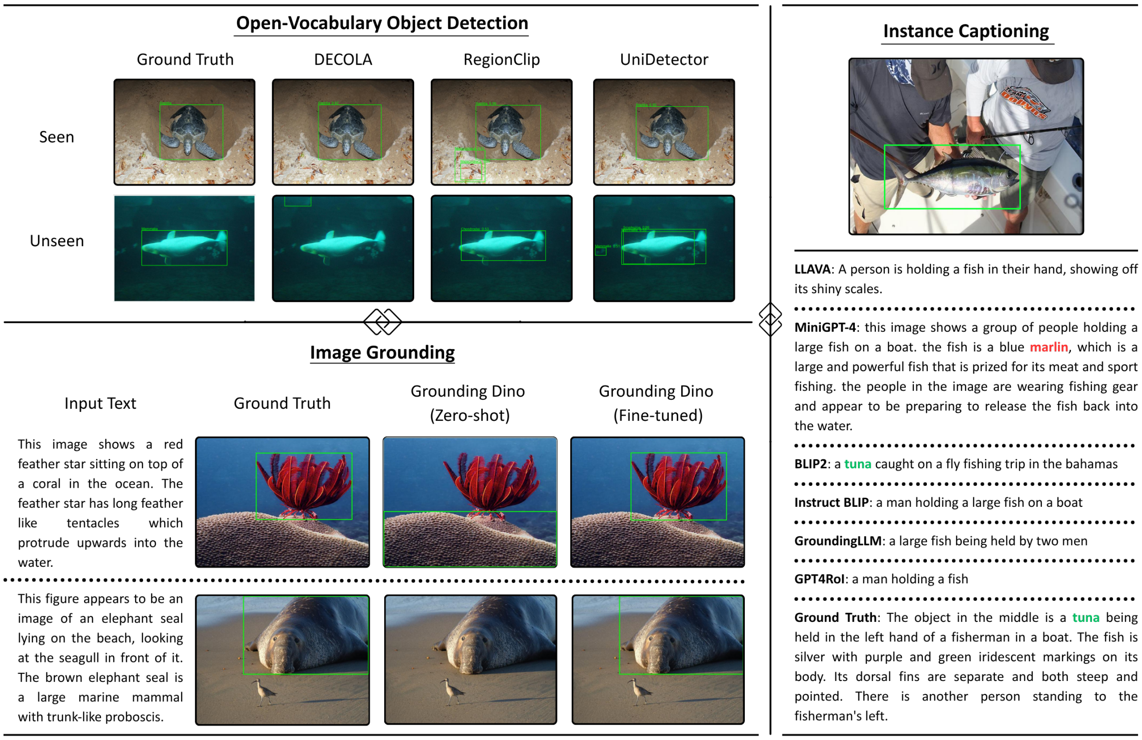
Comparison and analysis. Detecting marine organisms poses significant challenges for general object detection models. As summarized in Figure 6 and Table 2, while these models effectively locate objects, they struggle with accurate classification. We summarize two observations: 1) Morphological overlap across species confuses models for species identification, resulting in lower performance for Intra- and Inter-Class compared to Class-Level. 2) Relying solely on visual cues is insufficient for species identification. Performance in closed-set object detection (which relies exclusively on visual features) is generally lower than in open-vocabulary object detection (which incorporates both visual features and category labels). DECOLA demonstrates a clear advantage in recognizing fine-grained marine species. We attribute this to its language-conditioned query selection strategy.

4.2. Instance Captioning

Experimental settings. We benchmark off-the-shelf VLMs from two aspects: image-level and region-level. The former image-level VLMs (LLAVA [40], MiniGPT-4 [87], BLIP2 [31], and InstructBLIP [8]) were optimized by image-level captions and lacked the ability to understand specific object instances. We evaluate these image-level VLMs based on the following user instruction: "*describe the object in this figure*". The latter region-level VLMs (GroundingLMM [52], GPT4RoI [77]) were optimized by paired image region prompts and the corresponding instance captions. We provide the BBOX annotation in the given text prompt following the experimental setting of [52, 77]. We perform the evaluations based on expert-verified instance captions to analyze their capability in describing marine instance objects. To quantitatively measure the performance of various algorithms, we adopt the widely used captioning metrics (including CLIPScore, RefCLIPScore [18], CIDEr [65], BLUE-4 [49], METEOR [4], and Rouge [35]) to compute quantitative results in Table 3. Besides the human-constructed instance captions proposed in ORCA, we also construct a starting sentence to include the category information for the selected object instance: "This is a <*Category Name*>.", where the <*Category Name*> is the placeholder to compensate for the scientific category-level information of each instance. In this way, by penalizing generated plausible but not domain-specific responses (*e.g.,* "fish", "animal", and "mammal"), we encourage the model to generate the scientific captions to satisfy the domain requirements.

Implementation details. We perform the evaluation only based on the released official models provided by various algorithms on ORCA and our experiments were conducted using an NVIDIA L20 GPU. For LLAVA [40], we choose its V1.5-7b version for evaluation. The language model of MiniGPT-4 [87] is set to LLaMA-2 [64]. As for the GroundingLMM [52], we report the results of the models fine-tuned on RefCOCOg dataset [22] and Visual Genome (VG) dataset [26], respectively. For MiniGPT-4 fine-tuning, we train it on 4 NVIDIA A100-40GB for 5 epochs while other training parameters remain the same.

Comparison and analysis. Based on the results in Table 3 and Figure 6, we summarize the following



This image shows a red feather star sitting on top of a coral in the ocean. The feather star has long feather like tentacles which protrude upwards into the water.

This figure appears to be an image of an elephant seal lying on the beach, looking at the seagull in front of it. The brown elephant seal is a large marine mammal with trunk-like proboscis.

LLAVA: A person is holding a fish in their hand, showing off its shiny scales.

MiniGPT-4: this image shows a group of people holding a large fish on a boat. the fish is a blue marlin, which is a large and powerful fish that is prized for its meat and sport fishing. the people in the image are wearing fishing gear and appear to be preparing to release the fish back into the water.

BLIP2: a tuna caught on a fly fishing trip in the bahamas

Instruct BLIP: a man holding a large fish on a boat

GroundingLLM: a large fish being held by two men

GPT4RoI: a man holding a fish

Ground Truth: The object in the middle is a tuna being held in the left hand of a fisherman in a boat. The fish is silver with purple and green iridescent markings on its body. Its dorsal fins are separate and both steep and pointed. There is another person standing to the fisherman's left.

Figure 6. Quantitative results of open-vocabulary object detection, visual grounding, and image captioning.

Method	CLIPScore↑	RefCLIPScore↑	CIDEr↑	BLUE-4↑	METEOR↑	ROUGE↑
LLAVA [40]	73.78	72.27	4.93	8.77	7.70	20.76
MiniGPT-4 [87]	74.48	73.43	5.72	7.18	**16.90**	**28.03**
BLIP2 [31]	76.22	73.73	9.96	8.16	5.95	18.96
InstructBLIP [8]	**76.60**	**75.25**	**12.09**	**13.94**	7.40	21.31
GroundingLMM (RefCOCOg) [52]	73.04	70.97	4.37	4.39	4.60	16.37
GroundingLMM (VG) [52]	71.15	69.04	4.06	2.47	4.11	15.22
GPT4RoI [77]	71.28	68.71	3.53	2.81	4.07	15.08

Table 3. Results of various algorithms (image-level and region-level) on instance captioning.

observations: 1) The generic captioning model predominantly generates coarse phrases that are short and lack domain-specific knowledge. This aligns with the findings in Figure 4, where the training caption provided by the general dataset is also brief, in terms of length. The models frequently use everyday vocabulary, such as describing an object as "a large fish" instead of the more specific term, "marlin". Additionally, the models are prone to misclassifying objects. 2) The models primarily produce image-level captions and struggle to capture fine-grained features, such as morphology, color patterns, and textures. This limitation highlights a significant gap in the ability of general captioning models to support marine-specific tasks effectively. Fine-tuning MiniGPT-4 on ORCA further demonstrates that domain-specific training enhances image captioning performance. Additional details are provided in the supplementary material.

4.3. Visual Grounding

Experimental settings. We evaluate visual grounding models under both zero-shot and fine-tuned settings. Specifically, the expert-verified captions

Zero-shot						
Method	Unseen					
	Class-Level	Intra-Class	Inter-Class	Class-Level	Intra-Class	Inter-Class
GroundingVLP [60]	0.5183	0.5148	0.518	0.5816	0.5543	0.5837
TransVG [10]	0.5191	0.5025	0.5048	0.5849	0.5492	0.5773
GroundingDino [41]	0.5674	0.5606	0.5853	0.6324	0.5868	0.6278
HiVG [69]	0.4751	0.4386	0.4471	0.5459	0.4743	0.5399
Dynamic-MDETR [61]	0.5261	0.5004	0.5092	0.5856	0.5484	0.5792
CLIP-VG [68]	0.5499	0.5357	0.5346	0.6281	0.5789	0.6233
Fine-tuned						
Method	Seen			Unseen		
	Class-Level	Intra-Class	Inter-Class	Class-Level	Intra-Class	Inter-Class
TransVG [10]	0.6294	0.7213	0.6401	0.6984	0.7854	0.7216
GroundingDino [41]	**0.8114**	**0.8011**	**0.8077**	**0.8832**	**0.8554**	**0.8744**
HiVG [69]	0.6602	0.731	0.7235	0.7373	0.7892	0.8176
Dynamic-MDETR [61]	0.7494	0.7166	0.7511	0.8223	0.7762	0.8176
CLIP-VG [68]	0.7724	0.6191	0.6603	0.8569	0.6711	0.7433

Table 4. Visual-grounding performance reported as top-1 bounding-box accuracy at an *IoU* threshold of 0.5.

from our dataset are used as queries. The algorithms then predict a grounding box, and top-1 accuracy is reported at an Intersection over Union (IoU) threshold of 0.5. We select five models (TransVG [10], GroundingDINO [41], HiVG [69], Dynamic-MDETR [61], and CLIP-VG [68]) to assess performance in both zero-shot and fine-tuned scenarios. GroundVLP, [60] (with ALBEF [28] employed), a pipeline leveraging pretrained models, is evaluated exclusively in the zero-shot setting.

Implementation details. To ensure fair comparisons, we adhere strictly to official configuration files and evaluation scripts. For consistency with the GroundingDINO evaluation, which permits only one caption per image, we use the first annotation of each image to construct the testing dataset. In the zero-shot setting, we employ the publicly released models pre-trained on Flickr30K Entities [50], Objects365 [58], ReferItGame [22], based on the setting in the original paper accordingly.

For the fine-tuned setting, we retrain each model on ORCA using the same architecture as zero-shot setting and keep other hyperparameters the same.

Comparison and analysis. From Table 4, we summarize two observations: 1) Detailed captions facilitate species identification in visual grounding tasks. Unlike object detection, which suffers from significant performance drops in both Intra-Class and Inter-Class settings in Section 4.1, visual grounding tasks demonstrate no notable performance decline, even in the zero-shot setting. This underscores the importance of detailed captions in improving model robustness. 2) Fine-tuning on the ORCA yields significant performance improvements regarding visual grounding, with top-1 accuracy increasing by at least 10 percentage points for both *seen* and *unseen* categories across all three settings. These results indicate that while detailed captions enable general models to perform reasonably well in unseen marine scenarios, domain-specific supervision provides substantial additional gains.

5. Discussion and Conclusion

New benchmark. ORCA introduces a comprehensive and diverse benchmark specifically curated for marine research. Designed to advance the evaluation of algorithms for marine visual understanding, it encompasses a broad spectrum of marine species across varied environments, offering a valuable platform for testing and developing new models.

Limitation. Despite our efforts to include the most representative marine species, the diversity of marine life far exceeds the current set of categories. We plan to continually expand the dataset to incorporate additional marine objects over time.

Conclusion. This work presents the first large-scale marine dataset that supports both object recognition and detailed visual understanding. It enables multiple tasks, including *object detection*, *instance captioning*, and *visual grounding*. Our comprehensive evaluation highlights the strengths and limitations of both general-purpose and domain-specific algorithms, providing valuable insights for future research in marine applications.

6. Acknowledgement

This project was partially supported by Bridging Horizons: An AI-Powered STEM Learning Initiative in Space and Marine Education under the EdUHK–HKUST Joint Centre for Artificial Intelligence, the HKUST Marine Robotics and Blue Economy Technology Grant, and the Marine Conservation Enhancement Fund (MCEF20107 and MCEF22112).

The authors would also like to express their sincere gratitude to the "Sustainable Smart Campus as a Living Lab" (SSC) program at HKUST for its vital support. The program and its dedicated staff not only contributed essential funding and coordination but also fostered the integration of sustainability into campus operations, providing a real-world demonstration of the principles that underpin this research.

References

[1] Gpt-4v(ision) system card. 2023. 3

[2] et al. Ahyong, S. World register of marine species (worms). `https://www.marinespecies.org`, 2025. Accessed: 2025-05-24. 3

[3] Inyeong Bae and Jungpyo Hong. Survey on the developments of unmanned marine vehicles: Intelligence and cooperation. *Sensors*, 23(10):4643, 2023. 2

[4] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005. 6

[5] Océane Boulais, Simegnew Yihunie Alaba, John E Ball, Matthew Campbell, Ahmed Tashfin Iftekhar, Robert Moorehead, James Primrose, Jack Prior, Farron Wallace, Henry Yu, et al. Seamapd21: A large-scale reef fish dataset for fine-grained categorization. In *Proceedings of the FGVC8: The Eight Workshop on Fine-Grained Visual Categorization, Online*, page 2, 2021. 4

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[7] Jang Hyun Cho and Philipp Krähenbühl. Language-conditioned detection transformer. *arXiv preprint arXiv:2311.17902*, 2023. 5, 6

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

[9] Kubra Demir and Orhan Yaman. Projector deep feature extraction-based garbage image classification model using underwater images. *Multimedia Tools and Applications*, 83(33):79437–79451, 2024. 2

[10] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. *arXiv preprint arXiv:2104.08541*, 2021. 8

[11] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 4

[12] Changming Dong, Guangjun Xu, Guoqing Han, Brandon J. Bethel, Wenhong Xie, and Shuyi Zhou. Recent developments in artificial intelligence in oceanography. *Ocean-Land-Atmosphere Research*, 2022, 2022. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[14] Elizabeth J Drenkard, Charles Stock, Andrew C Ross, Keith W Dixon, Alistair Adcroft, Michael Alexander, Venkatramani Balaji, Steven J Bograd, Momme Butenschön, Wei Cheng, Enrique Curchitser, Emanuele Di Lorenzo, Raphael Dussin, Alan C Haynie, Matthew Harrison, Albert Hermann, Anne Hollowed, Kirstin Holsman, Jason Holt, Michael G Jacox, Chan Joo Jang, Kelly A Kearney, Barbara A Muhling, Mercedes Pozo Buil, Vincent Saba, Anne Britt Sandø, Désirée Tommasi, and Muyin Wang. Next-generation regional ocean projections for living marine resource management in a changing climate. *ICES Journal of Marine Science*, 78(6):1969–1987, 2021. 2

[15] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 5, 6

[16] Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[19] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 3

[20] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018. 4

[21] Md Jahidul Islam, Chelsey Edge, Yuyang Xiao, Peigen Luo, Muntaqim Mehtaz, Christopher Morse, Sadman Sakib Enan, and Junaed Sattar. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1769–1776. IEEE, 2020. 3, 4

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3, 6, 8

[23] Faizan Farooq Khan, Xiang Li, Andrew J. Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20496–20506, 2023. 4

[24] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11144–11154, 2023. 3

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3, 6

[27] Huanyu Li, Hao Wang, Ying Zhang, Li Li, and Peng Ren. Underwater image captioning: Challenges, models, and datasets. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:440–453, 2025. 2

[28] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 8

[29] 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 (ICML)*, pages 12888–12900. PMLR, 2022. 2, 3

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning (ICML)*, 2023. 3

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 6, 7

[32] Lin Li, Eric Rigall, Junyu Dong, and Geng Chen. Mas3k: An open dataset for marine animal segmentation. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 194–212. Springer, 2020. 3, 4

[33] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 32(4):2303–2314, 2021. 3

[34] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. Watermask: Instance segmentation for underwater imagery. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 1305–1315, 2023. 2, 4

[35] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[37] Maria Giulia Lionetto, Roberto Caricato, and Maria Elena Giordano. Pollution biomarkers in the framework of marine biodiversity conservation: State of art and perspectives. *Water*, 13(13):1847, 2021. 2

[38] Chongwei Liu, Haojie Li, Shuchang Wang, Ming Zhu, Dong Wang, Xin Fan, and Zhihui Wang. A dataset and benchmark of underwater object detection for robot picking. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 4

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Neural Information Processing Systems (Neurips)*, 2023. 3

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3, 6, 7

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 8

[42] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference Computer Vision*, pages 21–37. Springer, 2016. 3

[43] Heike K. Lotze. Marine biodiversity conservation. *Current Biology*, 31(19):R1190–R1195, 2021. 2

[44] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7363–7372, 2019. 5, 6

[45] OpenAI. Introducing chatgpt. 2022. 3

[46] OpenAI. Gpt-4 technical report, 2023. 3

[47] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011. 4

[48] Ryan J. O'Connor, Ana K. Spalding, Alison W. Bowers, and Nicole M. Ardoin. Power and participation: A systematic review of marine protected area engagement through participatory science methods. *Marine Policy*, 163:106133, 2024. 2

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[50] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017. 8

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[52] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 6, 7

[53] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles, 2010. Association for Computational Linguistics. 4

[54] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 5, 6

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 3

[57] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4

[58] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. 8

[59] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3

[60] Haozhan Shen, Tiancheng Zhao, Mingwei Zhu, and Jianwei Yin. Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection, 2023. 8

[61] Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2024. 8

[62] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19412–19424, 2024. 2

[63] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[64] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[65] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[66] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11433–11443, 2023. 3, 5, 6

[67] Yuk Kwan Wong, Ziqiang Zheng, Mingzhe Zhang, David J Suggett, and Sai-Kit Yeung. Coralscop-lat: Labeling and analyzing tool for coral reef images with dense semantic mask. *Ecological Informatics*, page 103402, 2025. 2

[68] Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 2023. 8

[69] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *ACM Multimedia 2024*, 2024. 8

[70] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data, 2024. 2

[71] Zongyao Yang, Xueying Yu, Simon Dedman, Massimiliano Rosso, Jingmin Zhu, Jiaqi Yang, Yuxiang Xia, Yichao Tian, Guangping Zhang, and Jingzhen Wang. Uav remote sensing applications in marine monitoring: Knowledge visualization and review. *Science of The Total Environment*, 838:155939, 2022. 2

[72] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023. 3

[73] Haifeng Yu, Xinbin Li, Yankai Feng, and Song Han. Multiple attentional path aggregation network for marine object detection. *Applied intelligence*, 53(2): 2434–2451, 2023. 2

[74] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 4

[75] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, 2021. 3

[76] Ruolan Zhang, Shaoxi Li, Guanfeng Ji, Xiuping Zhao, Jing Li, and Mingyang Pan. Survey on deep learning-based marine object detection. *Journal of Advanced Transportation*, 2021(1):5808206, 2021. 2

[77] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 6, 7

[78] Shu Zhang et al. Applications of marine geographic information systems (gis) in ocean surveying. *Journal of Environmental and Building Engineering*, 1 (1), 2024. 2

[79] Weidong Zhang, Gongchao Chen, Peixian Zhuang, Wenyi Zhao, and Ling Zhou. Catnet: Cascaded attention transformer network for marine species image classification. *Expert Systems with Applications*, 256:124932, 2024. 2

[80] Qianlong Zhao, Shiqiu Peng, Jingzhen Wang, Shaotian Li, Zhengyu Hou, and Guoqiang Zhong. Applications of deep learning in physical oceanography: a comprehensive review. *Frontiers in Marine Science*, 11, 2024. 2

[81] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 4

[82] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Tim Yue Him Wong, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public, 2023. 2, 3, 4

[83] 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. *ECCV*, 2024. 3

[84] Ziqiang Zheng, Haixin Liang, Binh-Son Hua, Yue Him Wong, Put ANG Jr, Apple Pui Yi CHUI, and Sai-Kit Yeung. CoralSCOP: Segment any COral image on this planet. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[85] Ziqiang Zheng, Yuk-Kwan Wong, Binh-Son Hua, Jianbo Shi, and Sai-Kit Yeung. Coralsrt: Revisiting coral reef semantic segmentation by feature rectification via self-supervised guidance. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[86] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, 2022. 5, 6

[87] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 6, 7

[88] Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish: A large benchmark for fish recognition in the wild. In *ACM international conference on Multimedia (ACM MM)*, pages 1301–1309, 2018. 2, 3, 4

[89] Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Transactions on Multimedia (TMM)*, 23:3603–3617, 2020. 2, 3, 4