MarineEval: Assessing the Marine Intelligence of Vision-Language Models

Yuk-Kwan Wong Tuan-An To Jipeng Zhang Ziqiang Zheng* Sai-Kit Yeung
Hong Kong University of Science and Technology

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

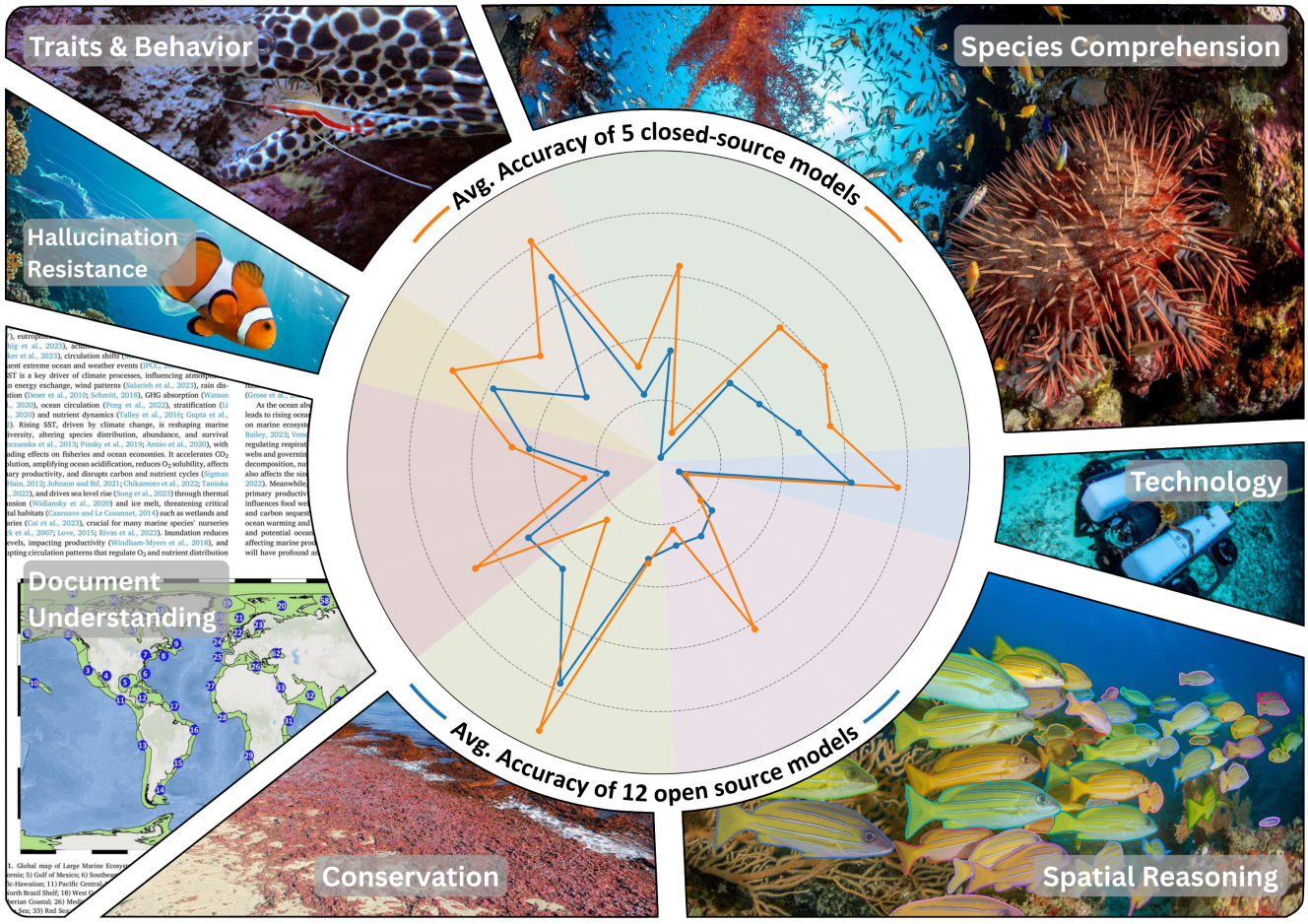

Figure 1. We present MarineEval, the first large-scale dataset and benchmark to comprehensively assess the ability of existing vision language models (VLMs) on marine intelligence.

Abstract

We have witnessed promising progress led by large language models (LLMs) and further vision language models (VLMs) in handling various queries as a general-purpose assistant. VLMs, as a bridge to connect the visual world and language corpus, receive both visual content and various text-only user instructions to generate corresponding responses. Though great success has been achieved by VLMs in various fields, in this work, we ask whether the existing VLMs can act as domain experts, accurately answering marine questions, which require significant domain expertise and address special domain challenges/requirements. To comprehensively evaluate the effectiveness and explore the boundary of existing VLMs, we construct the first large-scale marine VLM dataset and benchmark called MarineEval, with 2,000 image-based question-answering pairs. During our dataset construction, we ensure the diversity and coverage of the constructed data: 7 task dimensions and 20 capacity dimensions. The domain

*Corresponding author: zhengziqiang1@gmail.com

requirements are specially integrated into the data construction and further verified by the corresponding marine domain experts. We comprehensively benchmark 17 existing VLMs on our MarineEval and also investigate the limitations of existing models in answering marine research questions. The experimental results reveal that existing VLMs cannot effectively answer the domain-specific questions, and there is still a large room for further performance improvements. We hope our new benchmark and observations will facilitate future research.

1. Introduction

Vision-Language Models (VLMs) [38, 39, 55, 60, 67, 72, 73, 78] have achieved state-of-the-art results in a wide range of visual understanding tasks, including open-vocabulary object recognition [55], image captioning [38, 39], phrase grounding [53, 56] and interactive visual understanding [51], because of their strong comprehension ability to align visual contents and natural-language description. The growing ability has motivated not only the general public but also domain research from a broad spectrum of scientific and industrial fields to adopt VLMs for domain applications, such as medical analysis [37], mathematical computation [20], and scientific research [44].

In this work, we focus on the potential ability of powerful VLMs for marine understanding [31, 73], which is overlooked by existing research, but shares invaluable importance for protecting our ecosystem. The oceans, covering around 71% of the area of our blue planet, play vital roles in different fields, making marine research non-negligible. Regarding the importance of marine research, they remain logistically difficult and expensive to observe. Though qualitative results of VLMs in general scenarios so far are encouraging [34–36], quantitative evaluation is of great necessity to systematically evaluate and compare the abilities of various VLMs to conduct marine visual understanding.

Directly applying the existing VLMs for detailed marine visual understanding is non-trivial, and there are still some significant challenges. First of all, the underwater conditions [2, 73] contain the non-trimmed background, lacking prior knowledge for obtaining a reliable and comprehensive marine understanding. Furthermore, strong performance on generic datasets does not guarantee decent accuracy in specialised settings, where data distribution shift, domain gaps, and the lack of domain-specific knowledge can severely degrade model reliability, leading to significant hallucination [74]. We argue that the general-purpose evaluation dataset does not faithfully reveal the VLMs' capability in addressing domain-specific requirements (*e.g.*, biologists favor the population/density estimation [64, 75, 76], object counting [61], and relationship summarization [25]), as it rarely provides tailored tasks or authoritative ground truth for domain research. Consequently, existing research cannot effectively and reliably evaluate the performance of VLMs in handling marine understanding.

To satisfy the need for marine evaluation, a representative and rigorous benchmark tailored to a domain application is indispensable for tracking methodological progress and selecting reliable models. Besides, evaluating the performance of VLMs in marine research will provide valuable insights into the flexibility of existing VLMS as a domain-specific AI assistant. However, there are a few attempts [52, 72] to comprehensively evaluate VLM for more advanced analysis, which requires domain-specific knowledge and expertise. The foregoing analysis implies that a domain-aware evaluation dataset should satisfy two criteria: 1) questions should demand specialised marine knowledge rather than common sense; 2) capability dimensions should be defined at a granularity that reflects specific domain requirements.

In this paper, we take all the above-discussed challenges into consideration, presenting the first large-scale marine VLM dataset and benchmark called MarineEval. Our MarineEval is a multi-task dataset (including diverse question/task formats) with 2,000 manually constructed high-quality image-based question-answering pairs from 7 task dimensions and 20 domain-specific capability dimensions as illustrated in Figure 1. To retain the quality of the constructed benchmark, we have formulated a rigorous pipeline shown in Figure 2 to construct our dataset, which involves *visual necessity testing* and *domain expert verification*. Furthermore, to alleviate subjective grading and promote evaluation efficiency/reliability, we introduce detailed, scalable evaluation procedures to comprehensively assess VLM performance across multiple question formats. We have benchmarked 17 existing SOTA VLMs on our MarineEval on the right of Figure 1, where the best model could only achieve 49.58% accuracy. Substantial progress is still needed to enhance VLMs' performance on marine visual understanding. Our contribution can be summarized:

- We curate the first marine VLM benchmark with 2,000 high-quality image-based question-answering pairs, dedicated to marine analysis, enabling rigorous assessment of models on marine vision language tasks.
- We have included the domain requirements/challenges into our benchmark construction, where 20 manually constructed capacity dimensions could comprehensively measure VLMs' ability for marine understanding.
- Our experimental results and observations reveal limitations of existing VLMs, demonstrating persistent challenges in spatial reasoning, precise localization, species identification, and ecological knowledge integration.

2. Related Work

VLMs. The impressive performance of ChatGPT [49] and GPT-4 [50] has led to increasing attention to produce more

powerful LLMs as an AI assistant. VLMs equip LLMs with the ability to receive visual content, and they have unveiled remarkable zero-shot image-text capabilities in a conversational format. Flamingo [4] pioneered web-scale vision-language pretraining by bridging image and text models. BLIP [38, 39] bootstraps vision-language pre-training from frozen pre-trained image encoders and frozen language decoders. Based on BLIP-2 [39], MiniGPT-4 [78] proposed a projection layer to align pre-trained vision encoders to frozen LLMs (*e.g.* Vicuna [12]), and exhibited respectable zero-shot image comprehension in dialogues. GPT-4V [51] showcased impressive general-purpose visual understanding and reasoning abilities. However, these VLMs may still make mistakes, especially for the domain-specific knowledge, since it is not specifically optimized on the reliable domain-specific corpus/knowledge.

Marine datasets. Several datasets have recently been introduced for the marine domain. They provide insight into VLMs' capacity for marine understanding. MarineInst [73] emphasizes fine-grained perception by generating captions for individual object instances, while SeaFloorAI [47] concentrates on geological and spatial knowledge through question answering. MarineEval aims to evaluate broader model capabilities across the marine domain.

Benchmarking VLMs. Evaluating VLMs remains challenging due to the difficulty of assessing both their visual perception capabilities and their alignment with the inherently subjective and associative nature of human perception [18, 59]. To support systematic evaluation, multiple benchmarks have been introduced. For example, the SEED-Bench series [34–36] assesses VLMs through hierarchical tasks spanning a wide range of capabilities, while MM-Star [10] and related work [24] expose risks of knowledge leakage, whereby models infer answers from contextual cues rather than visual input. Evaluation efforts have also been extended to specialized domains such as recommendation [77], medicine [37, 57], multilingual understanding [29], and mathematics [20, 70], highlighting their potential but also their limitations. Addressing the lack of resources for marine applications, we introduce MarineEval, the first dataset tailored to evaluate VLMs in marine-centric tasks, explicitly incorporating domain-specific requirements.

3. MarineEval

In this section, we detail how we construct our MarineEval, which adheres rigorously to the evaluation criteria, while placing a strong emphasis on addressing specific marine challenges. It provides a tailored framework and well-defined capability dimensions to assess the effectiveness of VLMs in solving marine questions as illustrated in Figure 2.

3.1. General Criteria

MarineEval emphasizes 3 criteria for VLM evaluation.

Visual necessity: VLMs should derive the answers based on visual content, rather than relying solely on the textual inputs. As highlighted in [9, 24], there is a risk of knowledge leakage, where the question itself contains sufficient contextual cues for a well-trained LLM to get the correct answer. Such scenarios compromise the validity of the evaluation, as they fail to faithfully reflect the models' visual comprehension.

Objectivity: The evaluation rubrics should be clearly defined to avoid any subjective judgment. Existing works [29] adopt Likert-scale scoring, where the response is rated on a range (*e.g.,* 1 to 5). Even though such scoring allows for more precise performance analysis, it may result in fluctuation across adjacent scores if the criteria are not clearly defined or the evaluators do not reach a common sense. This lack of clarity results in subjective and inconsistent evaluations, reducing the reliability of the results.

Stability: Evaluation results should demonstrate stability and consistency across repeated trials. However, existing approaches frequently depend on human annotators, whose judgments are inherently subjective and susceptible to cognitive bias. This reliance introduces variability in outcomes when different groups of evaluators conduct the same experiments, thereby undermining the reproducibility and comparability of the results.

With the consideration of the above criteria, we formulate our dataset construction (Section 3.2) and evaluation process (Section 4.1) accordingly.

3.2. Dataset Construction

MarineEval employs a systematic multi-step process for dataset construction, as shown in Figure 2:

Data collection: We first harvest a large range of diverse public datasets by aggregating and post-processing the corresponding visual annotations from these candidate sources, including public classification datasets [26, 46, 63, 67, 71], object detection datasets [33, 48], counting dataset [61], marine-related books [5, 23, 28], scientific papers [8, 13–16, 19, 21, 32, 40, 45, 54, 58, 62, 66, 68], authoritative webpages [1, 17, 30], search engine, and private data. Detailed data source distribution is included in the Appendix.

Visual necessity testing: To eliminate questions that can be answered without visual content, a visual necessity test was conducted. Specifically, each question-answer pair in the candidate dataset was tested by removing the associated image and inputting it into five VLMs (Claude-3.7-Sonnet-Vision [6], Gemini-2.0-Flash-Vision [22], Grok-2-Vision [65], GPT-4o-Vision [51], and Qwen-VL-Plus [7]). If any one of the models can infer the answer without relying on the visual content, the corresponding question will be deemed to exhibit data leakage and will be excluded from further consideration. We emphasize that such visual necessity testing could better fairly evaluate the ability of existing VLMs to truly understand the given visual contents.

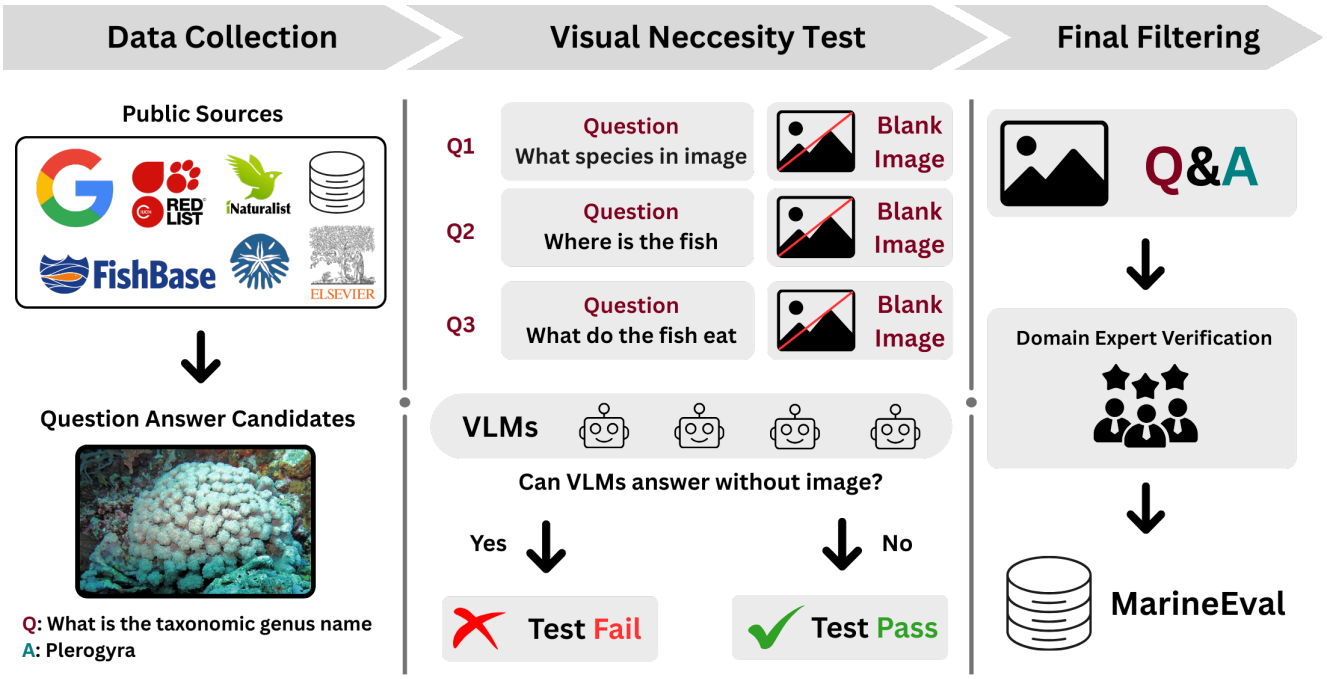

Figure 2. Workflow of MarineEval construction. 1) We first harvest diverse sources of candidate question-answer pairs, where the ground truth answer is post-processed by automatic programs or GPTs. 2) We adopt a visual necessity test to filter out pairs that are answerable without visual inputs. 3) Finally, domain experts construct and verify 2,000 high-quality pairs to constitute the final dataset.

Final filtering. Besides the automatic and program-based construction, we also design a human-in-the-loop procedure to manually formulate the final dataset, where the ground truth answers are verified by domain experts.

3.3. Evaluation Dimentions

Our MarineEval could be systematically categorized into 7 overarching task dimensions and 20 capacity dimensions. The 7 tasks are summarized below, and some concrete examples are shown in Figure 3. Details of 20 subfields are provided in the Appendix.

1. **Species Comprehension** examines the capability of VLMs to identify and interpret species-level visual information, thereby contributing to biodiversity monitoring and ecological research.
2. **Behavior & Trait Extraction** focuses on the ability to derive meaningful insights into the behavior and physical traits of marine organisms, facilitating advancements in automated observational and documentary records.
3. **Document Interpretation** evaluates the capacity of VLMs to analyze and derive insights from scientific literature and documentary sources. This functionality is especially critical for enhancing scientific understanding and generating insightful ecological reports.
4. **Conservation & Threat Analysis** emphasizes the ability of VLMs to accurately interpret domain-specific content, particularly in the context of endangered species and disaster classification.
5. **Spatial Reasoning** measures spatial comprehension ability. While it is commonly evaluated in general scenarios, MarineEval specifically investigates whether VLMs sustain high performance in marine environments.
6. **Marine Technology Understanding** evaluates understanding of marine technologies, which constitute a critical component of marine research.
7. **Hallucination Resistance** tests the robustness of VLMs in avoiding erroneous or hallucinatory outputs. Specifically, it involves pairing generally true statements with images that depict corner cases or counterexamples to assess whether the VLM is susceptible to being misled by the accompanying statements.

3.4. Dataset Statistics and Specific Features

MarineEval consists of 2,000 image-based question-answer pairs that span across 7 tasks and 20 capacity dimensions. To comprehensively evaluate the abilities of VLMs, we designed five distinct question formats: "Yes-No questions", "multiple-choice questions", "localization questions", "closed-form questions", and "summarization questions", as illustrated in Table 1. This diversity in question types enables MarineEval to assess a wide spectrum of capabilities for marine visual understanding, from basic factual judgement to complex reasoning and summarization tasks.

Our dataset contains three key features compared with existing general-purpose benchmarks:

Question Format	Description
Yes-No Question	Models make binary classification to determine whether a statement is true or false.
Multiple-Choice Question	Models select one or more than one correct option from at least four choices.
Localization Question	Models are asked to provide bounding box of target objects in COCO format.
Closed-Form Question	Models respond in a restricted format (*e.g.,* give a number or short phrases).
Summarization (Open-ended)	Models are asked to summarize the insight of the given image in free format.

Table 1. Explanations of different question formats in MarineEval.

1. Domain–specific marine knowledge requirements. The majority of questions in MarineEval demand specialised expertise in marine science, such as *taxonomic classification*, *IUCN conservation status*, and *biogeographic distribution* of specific organisms. This emphasis probes a knowledge space largely absent from mainstream training corpus, and thereby challenges the existing VLMs not only to retrieve and synthesize information but also to operate effectively within specialized knowledge domains.

2. Pronounced visual domain shift. The collected images in MarineEval diverge markedly from the general-purpose dataset that focuses on common scenarios or human-centered events. Differently, MarineEval features a large proportion of underwater photographs, exhibiting low contrast, motion blur, colour attenuation, and a large range of perspectives. The images often capture complex habitats such as reef communities and pelagic schools, while some of the images are satellite imagery. These modalities create a substantial distribution shift and introduce visual complexities, thereby providing a robust test bed to stress the zero-shot visual generalisation of VLMs.

3. Practical evaluation setting with specific domain requirements. MarineEval intentionally upheld both closed-form questions and open-ended questions to better represent real-world scenarios. While existing VLM datasets prioritize ease of evaluation by only providing "Yes-No" or "multiple-choice questions", MarineEval comprises 420 closed-form and open-ended questions (nearly one-fourth of the datasets) to measure models' ability to perform nuanced interpretation and free-form reasoning. Our design enables a more faithful measure of practical utility, where a fixed answer set is not available for a question.

4. Experiments

We first detail our experimental setting and then evaluate 17 SOTA VLMs on MarineEval by conducting a quantitative analysis and summarizing key findings regarding limitations of VLMs in marine visual understanding.

4.1. Experiment Settings

We start by explaining how we evaluate the existing VLMs. To ensure *objectivity*, *stability*, and *scalability*, as outlined in Section 3.1, we adopt a **binary judgement** evaluation strategy and report the final accuracy. To clearly verify the model responses, MarineEval classifies model outputs to either **correct** or **wrong**, regardless of their format or associated capability dimensions. Such an evaluation design introduces two benefits:

1. **Clear marking rubrics**. Unlike Likert-scale scoring, where marking criteria can often be ambiguous, binary judgment lowers the evaluation difficulty by reducing the scoring task to a binary classification. It minimizes subjective interpretations and thus promotes greater objectivity and reproducibility in evaluation.
2. **Easy comparison**. Binary judgement standardizes comparison by using *accuracy* as a universal metric across different models and dimensions. By maintaining a consistent evaluation standard, comparisons across models and dimensions become more straightforward.

We then provide the detailed evaluation metrics for computing the final accuracy regarding different question formats. For the "Yes-No" and "Multiple-Choice" questions, we first utilize the template matching to compute the accuracy of generated responses of various VLMs. During the evaluation procedure, we found some VLMs frequently violated the required response format, as these VLMs cannot strictly follow the user instructions to generate the required responses. To address this issue, we evaluate the models by appending each option to the question and computing the log-probability of the entire answer sequence. Instead of using raw logits, we compare the summed $\log p$ values over all tokens in the option. This avoids numerical issues and allows for fair comparison across options of varying lengths. The option with the highest total (or average) log-probability is selected as the final answer, and then we can eliminate the format issues for both "Yes-No" and "Multiple-Choice" questions. We adopt models' native decoding strategies for computing the probabilities.

Localization Questions. We evaluate the ability of VLMs to localize the specified object by the user instructions in the given image. We ask VLMs to follow the COCO format: (x, y, w, h) to yield the bounding box prediction. Then we compute the Intersection-of-Union (IoU) between the prediction and ground truth. We regard the output with an IoU score over 0.3 as an accurate prediction.

LLM for Judgement. While existing datasets often use multiple-choice questions to reduce ambiguity, MarineEval includes complex closed-form and open-ended questions to better reflect real-world marine scenarios and enable more

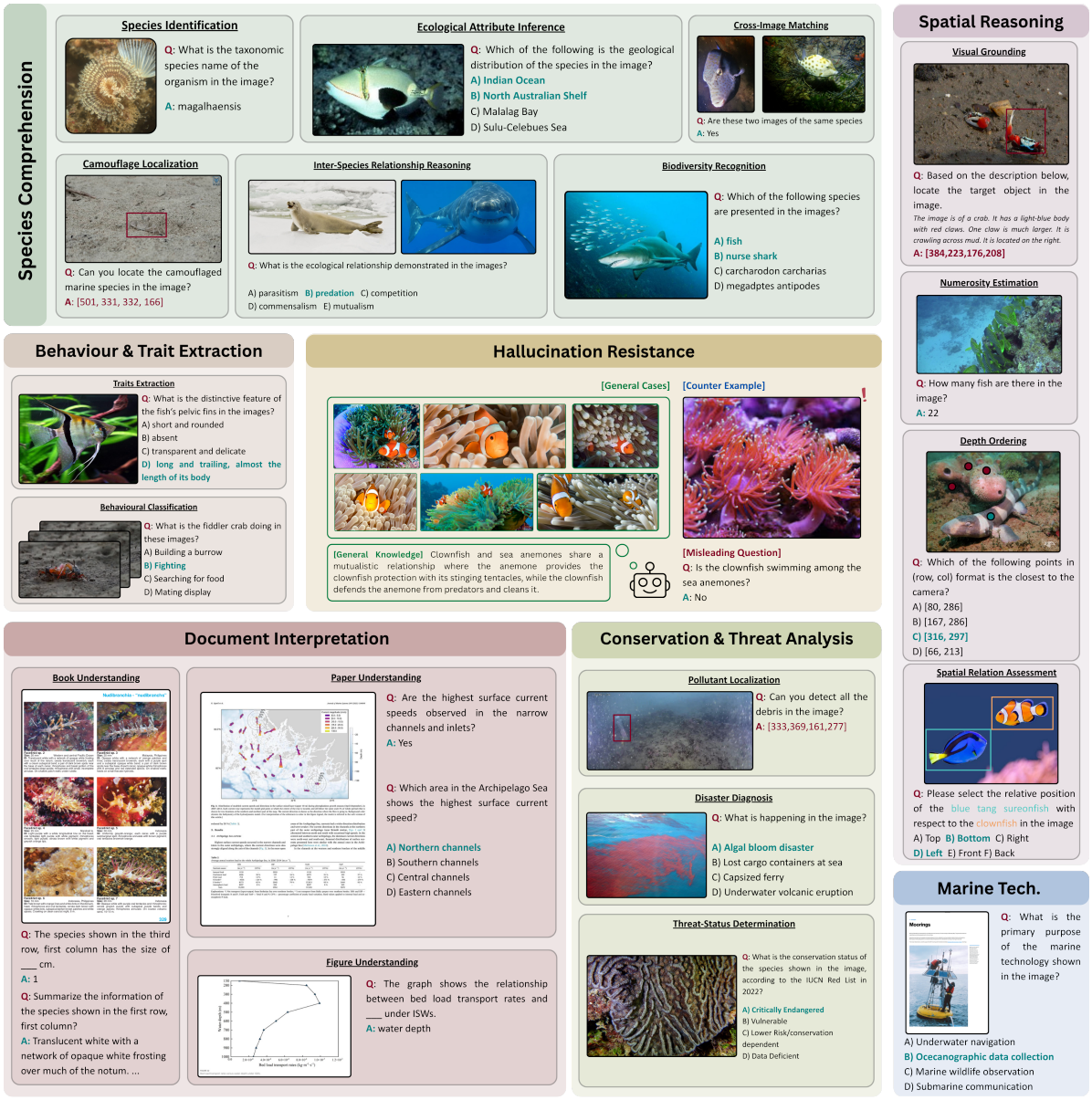

Figure 3. Overview of 7 task dimensions and 20 capacity dimensions of MarineEval. Best viewed in color.

comprehensive VLM evaluation. Evaluating open-ended responses is challenging due to nuanced interpretation, as generated answers may differ syntactically yet match semantically with ground truth. We employ LLMs for automated accuracy assessment, enabling scalable semantic comparison between ground truth and candidate responses. In detail, for both closed-form and open-ended questions, we first construct ground truth answers by domain experts as illustrated in Figure 4, where keypoints are summarized. Then we feed both the generated responses from the VLMs and the summarized keypoints by the humans to the LLMs to perform the matching from two aspects: whether there are 1) missing contents or 2) extraneous contents in the generated responses compared with the ground truth answers. In our experiments, considering there are some potential biases within the LLMs, we have chosen 3 powerful LLMs: GPT-4o-mini, Grok-3-mini, and DeepDeek-chat in our experiments.

4.2. Baselines

In our experiments, we have included 12 open-source VLMs: DeepSeek-VL-chat [43], OpenFlamingo [3], Mini-Monkey [27], InternVL-2.5 [11], LLaVA-1.6 Vicuna [41], InternLM-XComposer2.5 [69], LLaVA-Next [42] and InternVL-2 [11]; and 5 close-source VLMs: Claude-3.7-Sonnet-Vision [6], Gemini-2.0-Flash-Vision [22], Grok-2-Vision [65], GPT-4o-Vision [51], and Qwen-VL-Plus [7] for evaluation. In detail, we adopt the official models released on Huggingface for the open-source model, where the model

Q: Summarize the traits.

[Ground Truth]:
- Black and white banded eyestalks
- Bluish gray eyes
- Black claws

[Missing Content]:
This creature has eyestalks with a black and white striped pattern and bluish-gray eyes that stand out.

[Semantically Equivalent]:
The species is recognized by its banded eyestalks in black and white, grayish-blue eyes, and distinctively dark claws.

[Extraneous Content]:
Known for its black and white banded eyestalks, bluish-gray eyes, and black claws, this species is commonly found in coastal areas, often inhabiting discarded shells for protection.

Figure 4. Open-ended responses frequently include omissions or irrelevant content, which hinders reliable evaluation. We employ the LLM for judgments to compare the ground truth answers and generated responses.

size ranges from 1.8B to 38B. All experiments for the open-source model are conducted using 6 NVIDIA GeForce RTX 4090 D. All the used prompts and hyperparameter configurations will also be released. For the closed-source models, we conduct evaluations by calling their official APIs, with a total of 10,000 inference calls (5 closed-source models for 2,000 questions) made to ensure robust and fair comparison.

4.3. Results and Observations

The quantitative result comparisons between all the benchmarked VLMs are reported in Table 2. We report the detailed accuracy of 7 task dimensions in our MarineEval, the average accuracy across these 7 tasks, and the total accuracy (the primary performance metric) on the total dataset. We also recruited participants from both general and marine backgrounds to answer the questions as a reference to upper-bound human performance. We have the following observations regarding the results:

Inefficacious spatial and species understanding. Spatial Reasoning (*SR*) and Species Comprehension (*SC*) remain among the most challenging capabilities for all evaluated models. Spatial reasoning tasks, such as image grounding and depth ordering, require fine-grained geometric representations that are insufficiently captured by current VLMs. Species comprehension, on the other hand, involves taxonomic identification and the inference of ecological attributes, which are beyond the scope of general-purpose VLMs. Further analysis on the impact of the domain gap (please refer to supplmentary) indicates that the limited performance in species comprehension is largely attributed to the models' lack of domain-specific knowledge. In contrast, the poor performance in spatial reasoning primarily stems from an inherent deficiency in spatial understanding in the general setting.

Ecological insight scarcity. The performance of Conservation & Threat Analysis (*C&TA*) is still low for all models, *C&TA* questions involve disaster diagnostics and IUCN conservation status prediction, which represent corner cases and rare knowledge that is sparsely represented on the open web. Our results suggest that simply enlarging general-purpose corpora fails to cover long-tail ecological phenomena and the specialised reasoning they require.

Model choice. Model scale is not a reliable performance predictor of performing marine visual understanding. InternVL-2.5 (4B size) outperforms several larger models (even double-sized) and surpasses the closed-source models on multiple axes. This outcome underscores that architectural choices, vision encoders, and training strategy can outweigh parameter count. It also suggests diminishing returns for brute-force scaling when domain-specific supervision is scarce.

4.4. Further Analysis

In this section, we provide more experimental analysis.

Visual Necessity. We investigate the necessity of visual input in answering the questions, ensuring that the model cannot infer correct answers solely from the textual content. To this end, we conduct experiments under two settings: 1) *with visuals*, where the actual visual input is provided, and 2) *without visuals*, where a meaningless blank image is used as input. We evaluate all five closed-source VLMs and report their average accuracy across these settings. As shown in Table 3, model performance declines substantially when visual information is removed. We also include the accuracy of random guessing as reference. Notably, a small subset of closed-form questions are answered correctly even without visual input, primarily in counting tasks where models' random guesses coincide with the ground truth. Overall, these results confirm that the construction process of MarineEval does not unintentionally leak information to the VLMs, thereby ensuring the validity and integrity of the evaluation.

LLM reliability. We then investigate the reliability of using LLMs as judges, focusing on two key aspects: *stability* and *human alignment*. To assess stability, we repeat the evaluation procedure three times and report the mean and standard deviation in Table 4. The experimental results show that incorporating LLMs does not introduce instability when evaluating the responses generated by VLMs under our experimental setup. For human alignment, we randomly sample 500 question–response pairs, which are independently evaluated by both human annotators and LLM judges. The results indicate that the final judgments produced by the LLMs achieve a *95.40% agreement rate* with human evaluators, demonstrating the reliability of using LLMs.

Open-source VLMs										
Model	#. Params.	B&TE	C&TA	DI	HR	MTU	SR	SC	Avg.	Total
DeepSeek-VL-chat [43]	1.3B	27.86	39.33	11.00	59.00	34.31	22.25	18.33	30.30	24.96
OpenFlamingo [3]	2B	20.90	40.33	5.33	60.00	21.57	8.25	9.83	23.74	17.62
Mini-Monkey [27]	2B	44.28	50.33	33.00	58.00	74.51	12.75	27.67	42.93	34.45
InternVL-2.5 [11]	4B	65.17	56.67	54.00	64.00	**80.39**	16.75	29.33	52.33	42.54
LLaVA-1.6 Vicuna [41]	7B	68.66	52.00	38.67	53.00	71.57	**34.00**	37.33	50.75	44.73
InternLM-XComposer2.5 [69]	7B	64.18	60.33	49.33	52.00	75.49	14.00	30.17	49.36	41.14
LLaVA-Next [42]	8B	44.78	**69.67**	25.67	32.00	54.90	32.00	26.67	40.81	37.54
InternVL-2 [11]	8B	55.22	55.00	46.00	65.00	78.43	16.50	34.17	50.05	41.44
InternVL-2.5 (26B) [11]	26B	35.32	41.67	47.00	66.00	74.51	25.00	32.33	45.98	38.59
LLaVA-Next-Qwen [42]	32B	67.16	60.00	38.33	65.00	72.55	16.50	**43.67**	51.89	44.78
LLaVA-1.6 Hermes-Yi [41]	34B	68.66	52.00	38.67	53.00	71.57	**34.00**	37.33	50.75	44.73
InternVL-3 [79]	38B	**74.13**	48.33	**60.33**	**68.00**	78.43	22.50	39.83	**55.94**	**47.53**
Avg. across models	–	53.81	52.77	37.19	59.42	71.19	21.23	30.27	46.14	39.17
Close-source VLMs										
Model	#. Params.	B&TE	C&TA	DI	HR	MTU	SR	SC	Avg.	Total
Claude-3.7-Sonnet-Vision [6]	–	68.16	53.67	52.33	71.00	83.33	24.50	45.17	56.88	48.93
Gemini-2.0-Flash-Vision [22]	–	65.17	**60.67**	**59.67**	**74.00**	**87.25**	29.00	**55.33**	**61.59**	**55.07**
Grok-2-Vision [65]	–	**77.61**	54.67	27.33	**74.00**	70.59	**34.50**	54.00	56.10	50.42
GPT-4o-Vision [51]	–	69.15	44.67	51.67	72.00	62.75	26.50	40.50	52.46	45.58
Qwen-VL-Plus [7]	–	52.24	41.00	42.00	71.00	85.29	25.00	39.50	50.86	42.39
Avg. across models	–	66.07	50.34	46.20	72.40	77.64	27.90	46.10	55.18	48.08
Human Performance										
General Background	–	68.65	54.33	60.17	82.00	76.96	51.50	31.42	60.72	51.75
Marine Background	–	75.00	70.33	69.67	83.00	72.00	64.00	57.50	70.31	66.35

Table 2. The average accuracy across 7 task dimensions. Abbreviation: Behavior & Trait Extraction (B&TE), Conservation & Threat Analysis (C&TA), Document Interpretation (DI), Hallucination Resistance (HR), Marine Technology Understanding (MTU), Spatial Reasoning (SR), Species Comprehension (SC). We calculate the average accuracy across 7 tasks and the total accuracy of 2,000 questions. – indicates the number cannot be computed.

Question Format	Acc. (w/ visuals)	Acc. (w/o visuals)	Random guessing
Yes-No	62.24	42.66	50.00
MCQ	43.28	19.00	23.77
Localization	01.27	00.00	–
Closed-form	20.85	04.34	–
Summarization	12.67	00.00	–
Total	35.84	13.83	–

Table 3. Effectiveness of visual inputs on MarineEval, where average accuracy of all 5 closed-source VLMs in each question format is reported. "w/o visuals" indicates the VLMs are not given with any visual input or given with a meaningless blank image. We also report the accuracy of random guessing if available.

Potential data contamination. We acknowledge the potential data contamination issue in MarineEval, where some evaluation data might overlap with the training data of existing VLMs since this overlap could result in a biased or unfair comparison. Considering that all the benchmarked VLMs were optimized on diverse and extensive training corpora, it is inherently challenging to guarantee that all the testing data in MarineEval is entirely unseen by these models, as part of the data sources of MarineEval also come from publicly available datasets or public websites. Finally, it is important to emphasize that the primary objective of MarineEval is to explore the strengths and limitations of current VLMs in addressing marine challenges, not to fully address the data contamination issue.

Question Format	Acc. (mean_{std})
Closed-form	$20.83_{00.04}$
Summarization	$12.67_{00.00}$

Table 4. Mean and standard deviation of accuracy among 3 trials of using LLMs for measuring the generated responses from VLMs.

5. Conclusion and Discussion

In this work, we investigated whether existing VLMs can serve as domain experts in marine understanding, a field demanding specialized knowledge and nuanced understanding. Through the construction of MarineEval, the first large-scale marine VLM benchmark encompassing 2,000 image-based QA pairs across 7 task dimensions and 20 capacities, we rigorously evaluated 12 open- and 5 closed-source VLMs. Our experiments revealed a critical gap: while general-purpose VLMs excel in broad tasks, they struggle with marine understanding and exhibit notable hallucinations when addressing spatial localization and species identification tasks.

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] S. el. al. Ahyong. World register of marine species (worms). =https://www.marinespecies.org, 2025. Accessed: 2025-07-18. 3

[2] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6723–6732, 2018. 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 6, 8

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[5] Gerald R. Allen. *Reef Fish identification: Tropical pacific*. New World Publications, 2015. 3

[6] Anthropic. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family, 2024. Accessed: 2024-07-18. 3, 6, 8

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 3, 6, 8

[8] Gloria Castellano-González, Bruno Macena, Tiago Bartolomeu, A Passos, Pedro Afonso, and Jorge Fontes. Ecological aspects and hydrodynamics of hitchhiking remoras (remora sp.) associated with sicklefin devil rays (mobula tarapacana). *Marine Ecology Progress Series*, 752, 2024. 3

[9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. 3

[10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 3

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6, 8

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 3

[13] Marina Chiappi, Yolanda Stranga, Chrysanthi Kalloniati, Konstantinos Tsirintanis, George Tsirtsis, Ernesto Azzurro, and Stelios Katsanevakis. Cimpal expanded: unraveling the cumulative impacts of invasive alien species, jellyfish blooms, and harmful algal blooms. *Frontiers in Marine Science*, Volume 12 - 2025, 2025. 3

[14] Allison Dawn, Lisa Hildebrand, Florence Sullivan, Dawn Barlow, and Leigh Torres. Intermittent upwelling impacts zooplankton and their gray whale predators at multiple scales. *Marine Ecology Progress Series*, 752, 2024.

[15] Heather Doig, Oscar Pizarro, and Stefan Williams. Training marine species object detectors with synthetic images and unsupervised domain adaptation. *Frontiers in Marine Science*, Volume 12 - 2025, 2025.

[16] Justin Forget, Zou Zou Kuzyk, C.J. Mundy, and Céline Guéguen. Dissolved organic matter (dom) and barium in james bay: Distribution, sources, and climate change implications. *Journal of Marine Systems*, 250:104084, 2025. 3

[17] R. Froese and D. (Editors) Pauly. Fishbase. https://www.fishbase.se, 2025. Accessed: 2025-07-19. 3

[18] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3

[19] Yuri Fukai, A Fujiwara, S Nishino, S Kimura, M Itoh, and K Suzuki. Characteristics of autumn phytoplankton communities in the chukchi sea: Resuspension of settled diatoms to the surface during strong wind events. *Marine Ecology Progress Series*, 752, 2024. 3

[20] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 2, 3

[21] Claudio Garbelli, Miles Lamare, and Elizabeth Harper. Brachiopod shells as archives of seasonality: insights from

growth lines, microstructure, c and o values in calloria inconspicua. *Marine Biology*, 172, 2025. 3

[22] Google. Gemini 2.0 flash model card. https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf, 2024. 3, 6, 8

[23] Terrence Goslinear, Angel Valdes, and David W. Behrens. *Nudibranch & Sea Slug Identification: Indo-Pacific*. New World Publications, 2019. 3

[24] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. 3

[25] Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li, and Yi Wang. Coralvqa: A large-scale visual question answering dataset for coral reef image understanding. *arXiv preprint arXiv:2507.10449*, 2025. 2

[26] 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. 3

[27] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Multi-scale adaptive cropping for multimodal large language models. *arXiv preprint arXiv:2408.02034*, 2024. 6, 8

[28] Paul Humann and Ned DeLoach. *Reef creature identification: Tropical pacific*. New World Publications, 2010. 3

[29] Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. Heron-bench: A benchmark for evaluating vision language models in japanese. *arXiv preprint arXiv:2404.07824*, 2024. 3

[30] IUCN. The IUCN red list of threatened species. https://www.iucnredlist.org, 2022. Accessed on [day month year]. 3

[31] Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, et al. 1st workshop on maritime computer vision (macvi) 2023: Challenge results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 265–302, 2023. 2

[32] Gen Kume, Hiroki Takahira, Masafumi Kodama, Kazuhiko Anraku, Tomonari Kotani, Junya Hirai, and Toru Kobari. Analyses of gut content and isotopic composition of japanese eel anguilla japonica glass eels and elvers from an estuary in southern japan. *Marine Biology*, 172, 2025. 3

[33] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. *Asian Conference on Computer Vision*, 2020. 3

[34] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 2, 3

[35] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[36] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024. 2, 3

[37] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 2, 3

[38] 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

[39] 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. 2, 3

[40] Risto Lignell, Elina Miettunen, Harri Kuosa, Janne Ropponen, Laura Tuomi, Irma Puttonen, Kaarina Lukkari, Marie Korppoo, Markus Huttunen, Karel Kaurila, Jarno Vanhatalo, and Frede Thingstad. Modeling how eutrophication in northern baltic coastal zone is driven by new nutrient inputs, internal loading, and 3d hydrodynamics. *Journal of Marine Systems*, 249:104049, 2025. 3

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 6, 8

[42] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 8

[43] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 6, 8

[44] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[45] Rosemary Morrow and Elodie Kestenare. 30 years of sea surface temperature and salinity observations crossing the southern ocean near 140°e: Trends and rollercoaster variability. *Journal of Marine Systems*, 249:104048, 2025. 3

[46] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19023–19034, 2022. 3

[47] Kien X. Nguyen, Fengchun Qiao, Arthur Trembanis, and Xi Peng. Seafloorai: A large-scale vision-language dataset for seafloor geological survey. In *Advances in Neural Information Processing Systems*, pages 22107–22123. Curran Associates, Inc., 2024. 3

[48] Thanh-Danh Nguyen, Anh-Khoa Nguyen Vu, Nhat-Duy Nguyen, Vinh-Tiep Nguyen, Thanh Duc Ngo, Thanh-Toan

Do, Minh-Triet Tran, and Tam V. Nguyen. The art of camouflage: Few-shot learning for animal detection and segmentation. *IEEE Access*, 12:103488–103503, 2024. 3

[49] OpenAI. Introducing chatgpt. 2022. 2

[50] OpenAI. Gpt-4 technical report, 2023. 2

[51] OpenAI. Gpt-4o system card, 2024. 2, 3, 6, 8

[52] Aadi Palnitkar, Rashmi Kapu, Xiaomin Lin, Cheng Liu, Nare Karapetyan, and Yiannis Aloimonos. Chatsim: Underwater simulation with natural language prompting. *arXiv preprint arXiv:2308.04029*, 2023. 2

[53] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[54] Miguel Perea-Brugal, Amelly Hyldaí Ramos-Díaz, Perla Fernández-García, Carlos Cruz-Cruz, and Fernando Perea. Linking lunar phases to reproductive behaviors of the green sea turtle in the gulf of mexico. *Marine Biology*, 172, 2025. 3

[55] 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. 2

[56] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2

[57] Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*, 2024. 3

[58] Shunsuke Sen-ju, Sota Nakajo, and Akio Tamaki. Assessment of sediment-ejection activity of callianassid ghost shrimp on an intertidal sandflat in relation to seasonal hydrodynamic conditions, temperatures, and shrimp reproductive states. *Marine Biology*, 172, 2025. 3

[59] Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, and Gust Verbruggen. Assessing gpt4-v on structured reasoning tasks. *arXiv preprint arXiv:2312.11524*, 2023. 3

[60] 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

[61] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *IEEE/CVF International Conference on Computer Vision and Patern Recognition (CVPR)*, 2023. 2, 3

[62] Roberto Mario Venegas, Malika Kheireddine, Juan Pablo Rivera Caicedo, and Eric A. Treml. Climate-driven warming, deoxygenation, and desertification in large marine ecosystems. *Journal of Marine Systems*, 249:104053, 2025. 3

[63] smithin Reddy vighnesh anand, Yara Mamdouh. Oil spill dataset- binary image classification. Kaggle, https://www.kaggle.com/datasets/vighneshanand/oil-spill-dataset-binary-image-classification, 2024. 3

[64] 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

[65] xAI. Grok-2 beta release. https://x.ai/news/grok-2, 2024. 3, 6, 8

[66] Mei Xue and Guoping Zhu. Autumn trophic niche partitioning reduces interspecific competition between co-existing antarctic krill (euphausia superba) and hyperiid amphipod (themisto gaudichaudii). *Marine Biology*, 172, 2025. 3

[67] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024. 2, 3

[68] Zinaida Zabudkina, Alexander Osadchiev, Vladimir Ivanov, Mikhail Makhotin, and Viktor Merkulov. Interannual variability of the barents sea branch water in the northeastern part of the barents sea and the st. anna trough. *Frontiers in Marine Science*, Volume 12 - 2025, 2025. 3

[69] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 6, 8

[70] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multimodal llm truly see the diagrams in visual math problems?, 2024. 3

[71] Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Ran He. Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8458–8472, 2021. 3

[72] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*, 2023. 2

[73] 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. 2, 3

[74] Ziqiang Zheng, Yiwei Chen, Jipeng Zhang, Tuan-Anh Vu, Huimin Zeng, Yue Him Wong Tim, and Sai-Kit Yeung. Exploring boundary of gpt-4v on marine analysis: A preliminary case study. *arXiv preprint arXiv:2401.02147*, 2024. 2

[75] 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. 2

[76] 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. 2

[77] Peilin Zhou, Meng Cao, You-Liang Huang, Qichen Ye, Peiyan Zhang, Junling Liu, Yueqi Xie, Yining Hua, and Jaeboum Kim. Exploring recommendation capabilities of gpt-4v (ision): A preliminary case study. *arXiv preprint arXiv:2311.04199*, 2023. 3

[78] 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. 2, 3

[79] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 8