Supplementary Material for MarineEval

This supplementary contain the details of dataset statistic (Section A), domain gap analysis (Section B) and dataset dimension (Section C).

A. Dataset Statistic

This section contain the details of dataset size (Section A.1), qustion format distribution (Section A.2), and image resolution distribution (Section A.3).

A.1. Dataset Size

MarineEval achieves a balance between *quantity* and *expert-verified quality*. As shown in Table 1, MarineEval provides a comparable quantity to other domain-specific benchmarks while encompassing a broader variety of question formats, enhancing breadth and utility.

A.2. Question Format Distribution

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”, where the qusetion format distribution is shown in Figure 4.

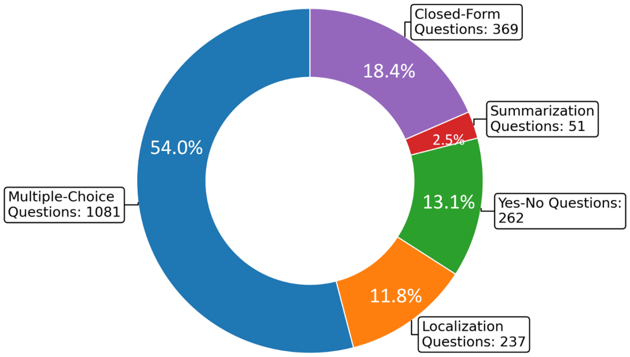


Figure 1. Distribution of various question formats in MarineEval

A.3. Image Resolution Distribution

The image resoluation distribution is illustrated in Figure 2. The maximum resolution is around **2,000 pixels**, with average resolution [629_{height}, 790_{width}].

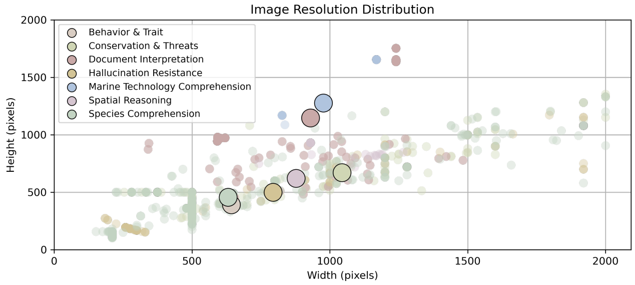


Figure 2. Distribution of image resolutions of MarineEval across capabilities. The large marker denotes **average resolution**.

B. Domain Gap Analysis

From the experiments reported in the main paper, we observed that the model exhibits notable weaknesses in two key aspects: **spatial reasoning** and **species comprehension**. In this section, we further investigate whether these shortcomings primarily arise from insufficient marine science domain knowledge or from fundamental limitations of current VLMs.

To this end, we constructed a human-verified evaluation set encompassing both general and marine-specific contexts, each consisting of 200 questions. The dataset was carefully designed to ensure both quality and comprehensive coverage. We then assessed the performance of the best-performing open-source and closed-source models, as summarized in Table 2. Our analysis reveals two major findings:

- **Spatial reasoning** performance remains consistently low across both domains, suggesting that this limitation stems from intrinsic weaknesses in current VLM architectures or optimization strategies rather than from a lack of domain-specific data.
- **Species comprehension** declines sharply in the marine context, highlighting these models’ heavy dependence on specialized knowledge and insufficient adaptation to niche domains.

To further contextualize these findings, we conducted a human evaluation study using MarineEval. In this experiment, individuals from both general and marine science backgrounds were invited to answer the same set of questions used for model evaluation. The results in Figure 3 reveal that participants with a marine science background substantially outperform both VLMs and general-background participants on average, particularly in Species Comprehension (*SC*).

Domain-Specific Benchmarks			
Benchmark	# Questions	Question Format	Domain
Heron-Bench	102	Free-Form	Japanese
KOFFVQA	275	Free-Form	Korean
CulturalVQA	2,378	Free-Form	Culture
VQA-RAD	3,515	Yes/No, Free-Form	Medical
MarineEval (Ours)	**2,000**	**Yes/No, Multiple Choice, Localization, Free-Form**	**Marine**
General-Domain Benchmarks			
MM-VET	200	Free-Form	General
MME	2,194	Yes/No	General
SEED-Bench	19,242	Multiple Choice	General

Table 1. Comparison of VQA benchmarks statistics.

General Domain	Spatial Reasoning	Species Comprehension
Best open-source model	18.00	24.50
Best close-source model	33.00	62.00
Marine Domain	Spatial Reasoning	Species Comprehension
Best open-source model	$17.00_{-01.00}$	$23.00_{-01.50}$
Best close-source model	$32.00_{-01.00}$	$35.71_{-26.29}$

Table 2. Performance comparison under general and marine settings. Each setting contains 200 questions.

This convergence between human and model performance gaps highlights that *SC* is the dimension most reliant on domain-specific expertise. Together with the quantitative results in Table 2, these findings underscore the essential role of domain adaptation in advancing multimodal understanding within specialized scientific fields such as marine ecology.

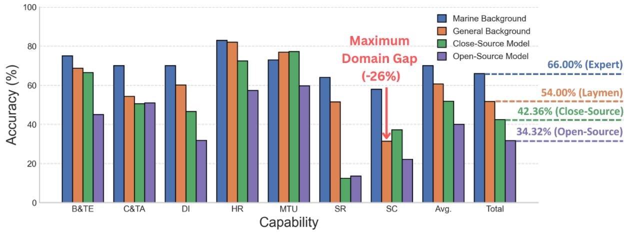


Figure 3. Performance comparison of humans and VLMs.

C. Dimensions

This section show the details of each dimension in 7 capabilities, including Species Comprehension (Section C.1), Behaviour & Trait Extraction (Section C.2), Document Interpretation (Section C.3), Conservation & Threat Analysis (Section C.4), Marine Technology Understanding (Section C.5), Spatial Reasoning (Section C.6), Hallucination Resistance (Section C.7). The overview of all 7 task dimensions and 20 capacity dimensions is shown in Table 3

C.1. Species Comprehension

This dimension evaluates the ability of VLMs to identify and interpret species-specific visual information, thereby supporting biodiversity monitoring and ecological research. We detail various subtasks in the following subsections.

C.1.1. Species Identification

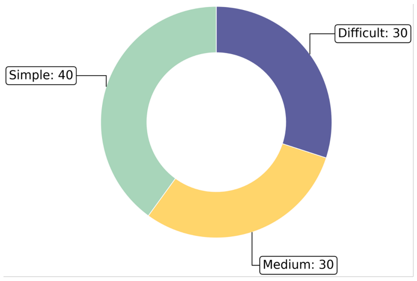


Figure 4. Distribution of difficulties for Species Identification task.

This dimension assesses the capacity of VLMs to perform taxonomic classification of organisms depicted in images. To provide a more granular evaluation of VLMs' performance, the questions are categorized into three levels of difficulty based on the taxonomic rank. The definitions of these difficulty levels and their distribution are presented in Table 4 and Figure 4, respectively. A question sample is shown in Figure 6 for better illustration.

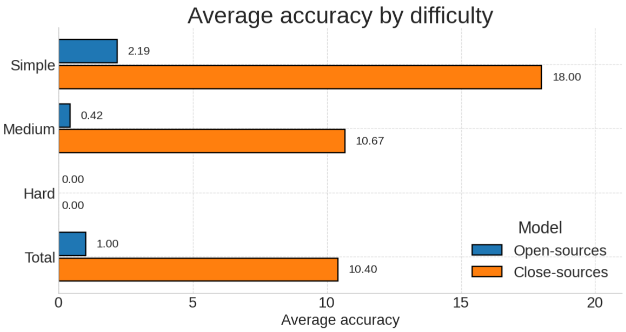


Figure 5. Model performance for the Species Identification task.

Capability	Dimension	Description
Species Comprehension	Species Identification	Determine the scientific (binomial) name of a species from a single specimen image.
	Cross-Image Matching	Assess whether two independently presented images depict the same species.
	Biodiversity Recognition	Identify and list all distinct species visible within a complex ecological scene.
	Ecological Attribute Inference	Infer ecological attributes of a given species—such as habitat, geographic range, or trophic role—from its image.
	Inter-Species Relationship Reasoning	Analyze pairs of species images to deduce their ecological relationship (e.g., predation, mutualism, symbiosis).
	Camouflage Localization	Detect and spatially delineate organisms exhibiting camouflage through disruptive coloration or texture blending.
Behaviour & Trait Extraction	Trait Extraction	Extract key morphological or visual traits of a species from a single image.
	Behavioural Classification	Identify and classify behavioral patterns displayed in a short sequence of frames.
Document Interpretation	Figure Understanding	Interpret and derive insights from scientific figures or graphical data representations.
	Book Understanding	Comprehend and extract information from book pages.
	Paper Understanding	Analyze, and interpret research findings presented in scientific journal articles with extensive textual content.
Conservation & Threat Analysis	Disaster Diagnosis	Identify the type of environmental disaster represented in an image.
	Pollutant Localization	Detect and localize anthropogenic pollutants or contaminants within the visual scene.
	Threat-Status Determination	Determine the International Union for Conservation of Nature (IUCN) conservation status of the focal organism.
Marine Technology Understanding	Instrument Function Identification	Identify the function or operational role of marine equipment shown in an image.
Spatial Reasoning	Visual Grounding	Locate within an image the species or object referenced in a textual query.
	Numerosity Estimation	Count the number of instances of relevant entities (e.g., fish, vessels) present in an image.
	Depth Ordering	Determine which of several designated points in an image is closest to the observer or camera viewpoint.
	Spatial Relation Assessment	Infer the relative spatial arrangement between two organisms (e.g., left of, above).
Hallucination Resistance	Hallucination Resistance	Assess whether the model produces fabricated or unsupported outputs when prompted with ambiguous or controlled inputs.

Table 3. Overview of the seven task dimensions and twenty capability dimensions in our MarineEval.

Difficulty	Description
Simple	Models are asked to determine the taxonomic `Family` name.
Medium	Models are asked to determine the taxonomic `Genus` name
Hard	Models are asked to determine the taxonomic `Species` name

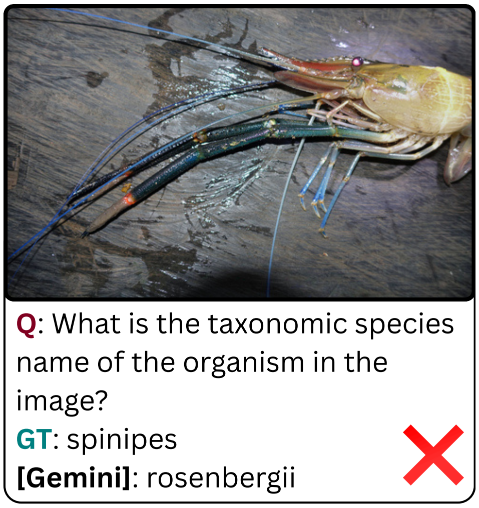
Table 4. Definition of difficulty for the Species Identification.

Data collection: The data for this dimension is sourced from BioTrove [35], which provides taxonomic classifications at various levels along with corresponding images. All classification labels have been verified by more than three domain experts and researchers on iNaturalist [12].

Performance: As reported in Figure 5, both open-source and close-source VLMs face challenges in accurately identifying organisms at the `Species` level. However, close-source models generally outperform open-source models across different levels of classification. We attribute this phenomenon to the larger training corpus of these close-source VLMs.

C.1.2. Cross-Image Matching

This dimension evaluates whether VLMs can identify the same species across different images or differentiate between distinct species with similar appearances. VLMs are presented with two independent images and tasked with determining if both depict the same species. The evaluation is



Q: What is the taxonomic species name of the organism in the image?

GT: spinipes

[Gemini]: rosenbergii

Figure 6. Question sample of the Species Identification task.

categorized into three difficulty levels based on taxonomic classification, as species within the same higher-level taxonomic group often exhibit similar physical traits, making differentiation more challenging. A visual sample is shown in Figure 9 for better explanation. Table 5 and Figure 7 illustrate the definitions and distributions of these difficulty levels, respectively.

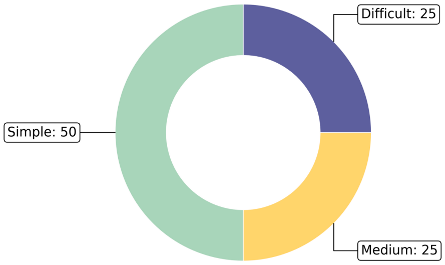


Figure 7. Difficulty distribution in the Cross-Image Matching task.

Data collection: The data used for this evaluation is also derived from BioTrove [35], which provides taxonomic classifications and corresponding images. All classification labels were verified by at least three domain experts and researchers on iNaturalist [12]. For each difficulty level, we sampled two image pairs to construct a cross-image matching question.

Performance. As shown in Figure 8, model performance declines as task difficulty increases, indicating greater challenges in distinguishing closely related species. However, closed-source models show no significant performance drop from the difficulty level of *Medium* to *Hard*, potentially due to their stronger visual encoders.

C.1.3. Biodiversity Recognition

This dimension evaluates a model's ability to identify every taxon present in a visually complex scene. Given an input image, the VLM must enumerate all species it can observe. A representative sample is provided in Figure 10.

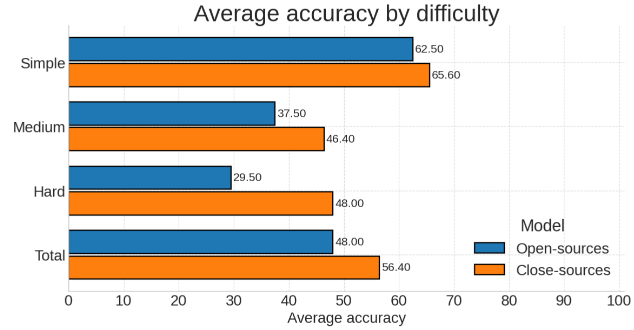
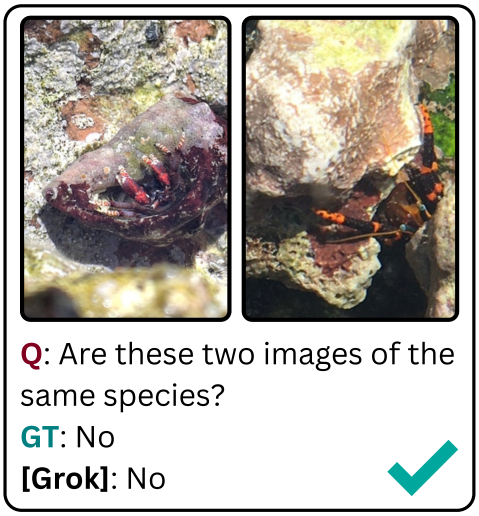


Figure 8. Model performance for the Cross-Image Matching task.



Q: Are these two images of the same species?

GT: No

[Grok]: No

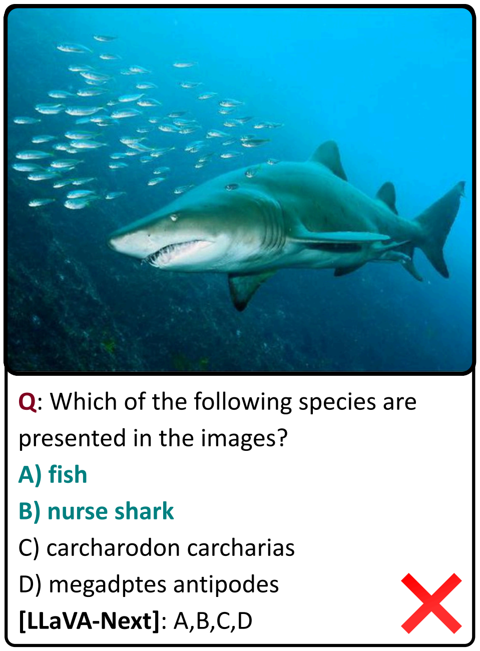
Figure 9. Question sample of the Cross-Image Matching task.

Difficulty	Description
Simple	Two images depict either the same species or different species from distinct taxonomic `Families`.
Medium	Two images depict different species within the same taxonomic `Family` but belonging to separate `Genera`.
Hard	Two images depict different species within the same taxonomic `Genus` but belonging to separate `Species`.

Table 5. Definition of difficulty for Cross-Image Matching task.

Data collection. The images are drawn from a self-annotated object-detection corpus. Images with fewer than two species are discarded to ensure sufficient task difficulty. For the multiple-choice question, we keep only the class labels and augment each question with three distractors drawn uniformly from the remaining label pool.

Performance. As illustrated in Figure 11, the closed-source models outperform the open-source models. However, despite the smaller pre-training budgets, InterVL-2 [3] and InternLM-XComposer-2.5 [37] achieve similar performance to the leading proprietary models. It suggests that excellence on generic multimodal benchmarks does not automatically indicate a strong fine-grained ecological understanding



Q: Which of the following species are presented in the images?

A) fish

B) nurse shark

C) carcharodon carcharias

D) megadptes antipodes

[LLaVA-Next]: A,B,C,D ✗

Figure 10. Question sample of the Biodiversity Recognition task.

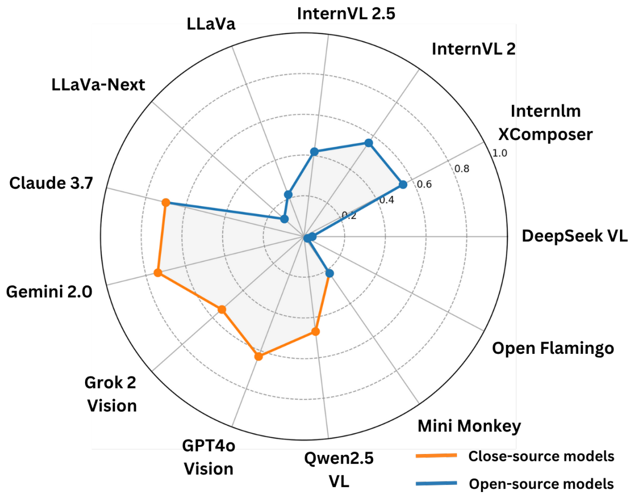


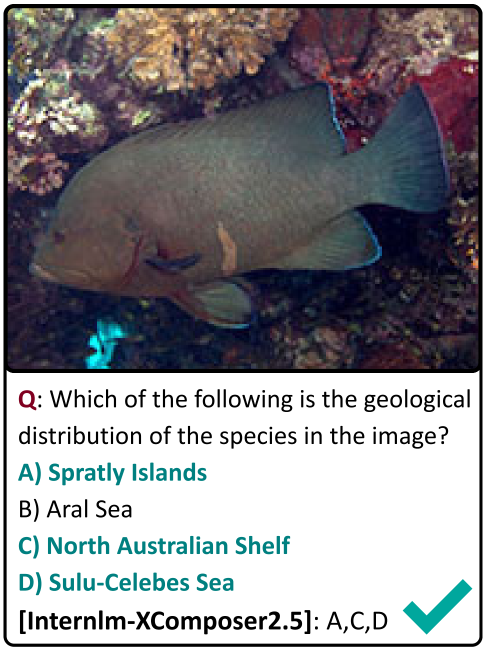
Figure 11. Model accuracy for the Biodiversity Recognition task.

ability. Instead, domain-aware pre-training and instruction tuning are essential.

C.1.4. Ecological Attribute Inference

This task assesses whether a VLM can deduce species–specific ecological traits from a single species image. Concretely, the model must answer multiple–choice questions concerning a taxon's geographic range, preferred habitat, reproductive strategy, or trophic niche. An illustrative example is provided in Figure 12.

Data collection Images and ground–truth attributes are harvested from the BioTrove [35] and FishBase [8] repositories. FishBase, in particular, provides expert–validated ecological



Q: Which of the following is the geological distribution of the species in the image?

A) Spratly Islands

B) Aral Sea

C) North Australian Shelf

D) Sulu-Celebes Sea

[Internlm-XComposer2.5]: A,C,D ✓

Figure 12. Question sample of the Species Ecology task.

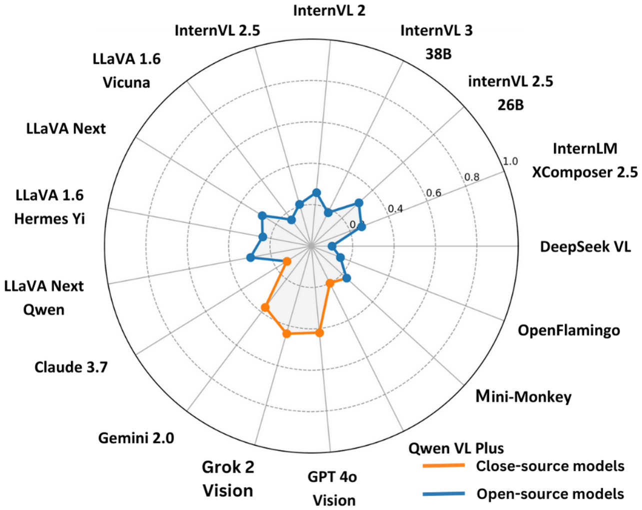


Figure 13. Model accuracy for the Species Ecology task.

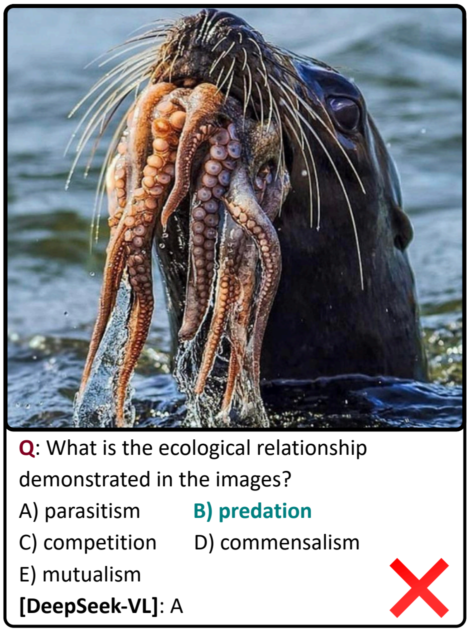
annotations with primary literature references, ensuring high label fidelity.

Performance. Figure 13 reveals a marked performance collapse for all evaluated models, where no models surpass 50% accuracy. Accurate inference of non-visual biological attributes requires connecting visual identification to a structured body of domain knowledge, a capability that present-day VLMs largely lack. The consistent shortfall across all architectures suggests a systemic knowledge gap rather than an optimization issue specific to any one model.

Future work could contain 1) incorporating curated ecological knowledge bases during multimodal pre-training, 2) exploring retrieval-augmented decoding to inject factual context at inference time, and 3) developing evaluation suites that disentangle visual recognition errors from knowledge-retrieval failures. Such directions are essential if VLMs are to support real-world biodiversity monitoring and conservation applications.

C.1.5. Inter-Species Relationship Reasoning

This evaluation dimension measures a model's ability to identify interspecific ecological interaction. The task covers the five canonical relationship types: *parasitism*, *predation*, *competition*, *commensalism*, and *mutualism*. Given one or two photographs of the organisms, the model must infer the most plausible interaction and choose the corresponding label. A visual sample is shown in Figure 14.



Q: What is the ecological relationship demonstrated in the images?

A) parasitism B) predation

C) competition D) commensalism

E) mutualism

[DeepSeek-VL]: A

Figure 14. Question sample of the Inter-Species Relationship Reasoning task.

Data acquisition. We first employed GPT-4 to generate candidate species pairs for each interaction category, then retrieved representative photographs through Google Image Search, retaining the original URLs to protect copyright. Marine biologists subsequently screened the material and curated a balanced set of 100 image–question–answer pairs in which 1) both organisms are clearly visible and 2) the annotated interaction is ecologically valid.

Performance. Although recent VLMs achieve encouraging results on general-purpose benchmarks, their accuracy on our marine benchmark remains modest: no model exceeds an average score of 70%, and several perform only marginally better than random guessing, as shown in Figure 15. The large performance room reveals that general-purpose pretraining is insufficient for reliably interpreting specialised marine imagery and reasoning about ecological relations. In particular, the models struggle with recognising less-frequent taxa and distinguishing subtle interaction cues. These limitations underscore the need for domain-specific data, expert supervision, and tailored fine-tuning before VLMs can be deployed with high confidence in real-world marine science applications.

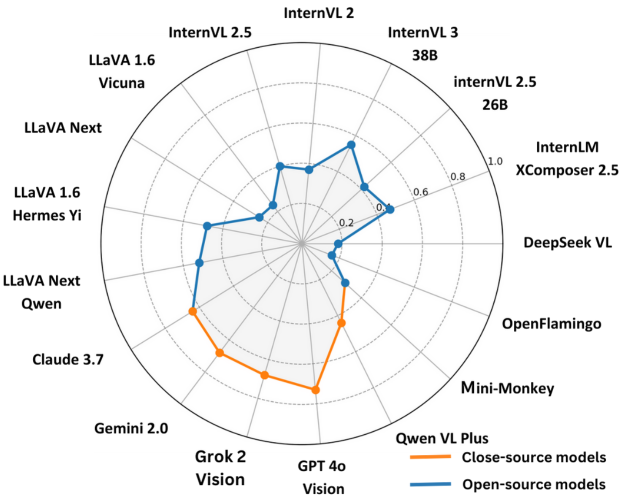
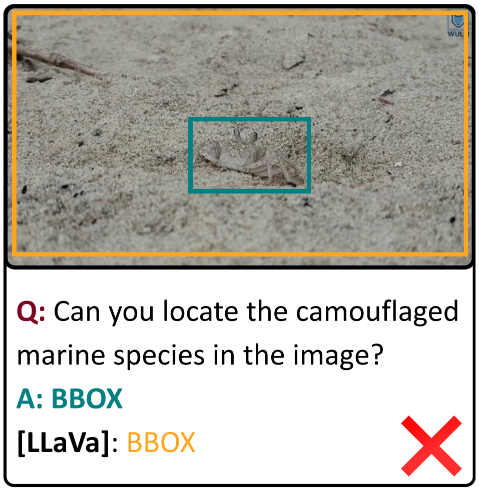


Figure 15. Model accuracy for the Inter-Species Relationship Reasoning task.



Q: Can you locate the camouflaged marine species in the image?

A: BBOX

[LLaVa]: BBOX

Figure 16. Question sample of the Camouflage Localization task.

C.1.6. Camouflage Localization

Camouflage localization is notoriously challenging, even for experienced human observers. Successful identification requires 1) knowledge of the species-specific chromatic patterns that allow the animal to blend seamlessly with its surroundings, and 2) the ability to delineate the organism's precise spatial extent. We introduce this task to probe the limits of current VLMs. The question is deemed correct if there is any one predicted bounding box that overlaps with the ground truth with $IOU > 0.3$. An illustrative example is shown in Figure 16.

Data collection. We curated 100 marine-camouflage images from the MOCA [17] and CAMO-FS [25] datasets. Non-marine instances were manually removed, and the bounding boxes provided by the original datasets were exhaustively re-verified. The resulting corpus comprises 100 image–question–answer triples.

Performance. The result is demonstrated in Figure 17. Some of the models achieve nearly zero accuracy because they cannot output bounding boxes in the required format. For the other models, the near-floor performance indicates that contemporary VLMs largely fail to perceive and localize camouflaged marine animals. In other words, the models do not yet possess the fine-grained color discrimination, texture sensitivity, and contour reasoning required for this localization task, highlighting a significant gap between current VLMs and human-level perception in complex natural scenes.

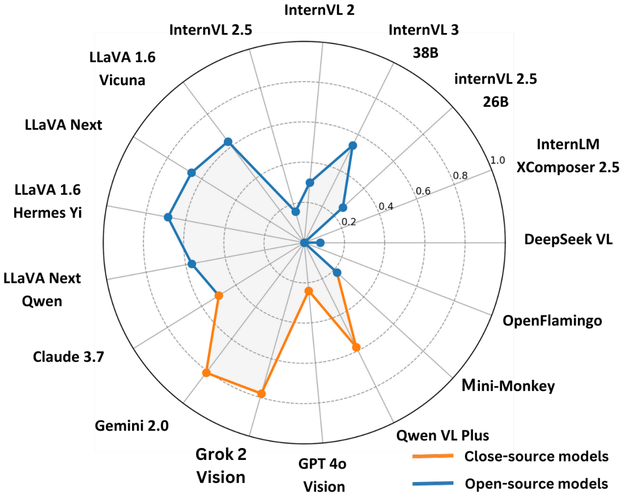


Figure 17. Model accuracy for the camouflage localization task.

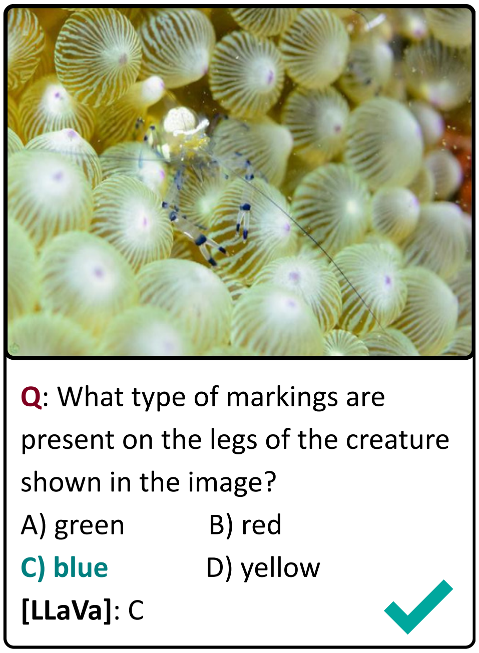
C.2. Behaviour & Trait Extraction

This capability focuses on extracting species-specific traits and classifying behaviors from images or sequences, aiding in understanding biological and ecological processes.

C.2.1. Traits Extraction

This dimension tests VLMs' ability to align visual trait appearance and textual description, which focuses on the biological traits of the species. Models are asked to answer multiple-choice questions to determine the trail (*e.g.,* color, shape, texture, or pattern *etc*) of the target species. A visual testing example is shown in Figure 18.

Data collection. Question-answer pairs are collected from a self-collected image captioning dataset, where each caption description the physical appearance of the target species, which is written by marine ecologists. We input the image and the image caption to GPT to generate multiple-choice questions, with one correct answer and other misleading answers.



Q: What type of markings are present on the legs of the creature shown in the image?

A) green B) red

C) blue D) yellow

[LLaVa]: C ✓

Figure 18. Model accuracy for the Traits Extraction task.

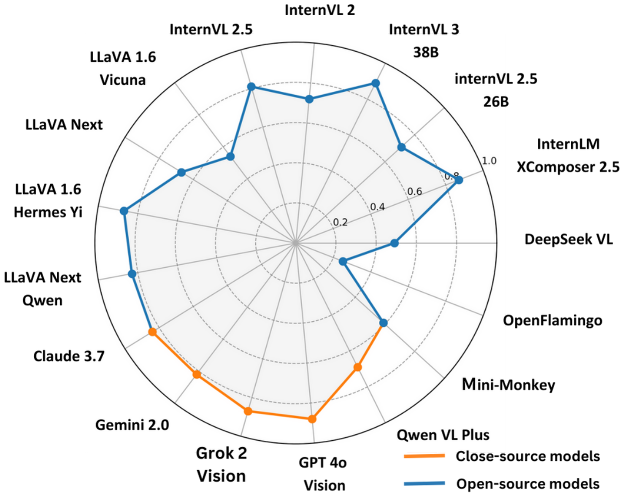
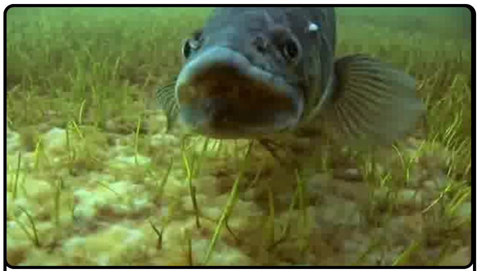


Figure 19. Model accuracy for the Traits Extraction task.

Performance. As shown in Figure 19, all open-source models perform equally well on the Traits Extraction tasks. Among the closed-source models, InternVL [3] and InternVL2.5 [3] outperform the others, demonstrating higher accuracy and better consistency in extracting detailed species traits.

C.2.2. Behavioural Classification

This dimension evaluates the VLMs' ability to infer species-specific behavioral characteristics from a sequence of frames. For example, models will be asked to identify behaviors such as attacking prey, escaping habitats, and surfacing for air or hiding. A visual sample is shown in Figure 20.



Q: What is the fish doing in this image?
A) Camouflaging in the grass
B) Attacking prey
C) Building a nest
D) Resting on the seabed
[Claude]: B ✓

Figure 20. Question sample of the Behavioural Characteristics task.

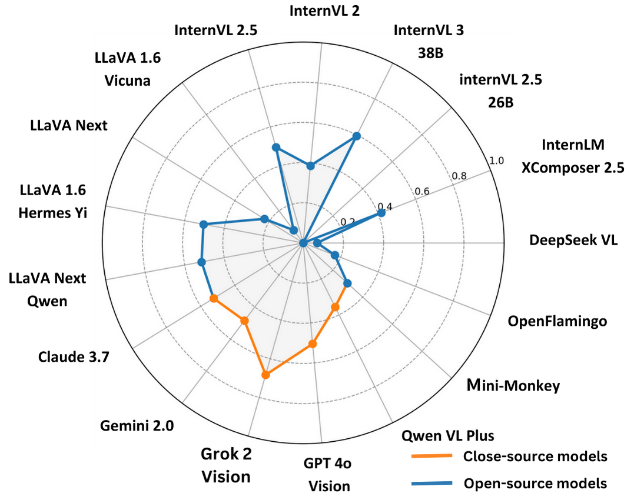


Figure 21. Model accuracy for the Behavioural Classification task.

Data collection. Data source for this dimension are drawn from the Animal Kingdom dataset [24], a large-scale resource designed for animal behavior understanding. Verified behavior and species labels are adapted to construct new question-answer pairs tailored to our task. We extract 10 consecutive frames at uniform intervals from 0 to 1, with a step size of 0.1 to capture dynamic behaviors. To ensure meaningful behavioral distinctions, we filter out the generic action label "swimming", as it is common to most marine species underwater.

Performance. Figure 21 presents the performance of the evaluated VLMs on the behavioural classification task. Overall, the closed-source models performed worse than the open-source ones, possibly due to suboptimal multi-frame processing. Among all the closed-source models, Grok-2-Vision achieved the highest accuracy, exceeding 60%. For the open-source models, InternVL-2.5 attained the best result with an accuracy of 50%.

C.3. Document Interpretation

This capability systematically evaluates a model's capacity to comprehend and interpret complex scientific documents, including detailed figures, tables, equations, and extensive technical diagram, thereby enabling accurate and contextually informed knowledge extraction to support advanced applications in education, research, and scientific discovery.

C.3.1. Figure Understanding

This dimension evaluates the VLMs' ability to derive insights from scientific figures or graphs. For example, models are asked to interpret trends, compare values, or summarize key information presented in charts, plots, or annotated diagrams. From a given figure, the model should infer relevant conclusions, describe relationships between variables, or identify anomalies and significant patterns. A visual testing sample is shown in Figure 22.

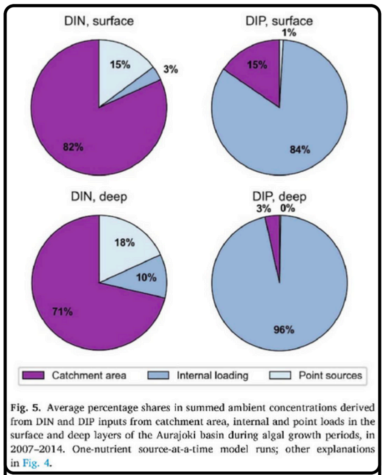


Fig. 5. Average percentage shares in summed ambient concentrations derived from DIN and DIP inputs from catchment area, internal and point loads in the surface and deep layers of the Aurajoki basin during algal growth periods, in 2007–2014. One-nutrient source-at-a-time model runs; other explanations in Fig. 4.

Q: In the surface layer, the largest percentage of DIN input comes from the ____ area.
GT: catchment
[Claude]: catchment ✓

Figure 22. Question sample of the Figure Understanding task.

Data collection. 15 Open-access marine science papers [2, 4–7, 9, 10, 16, 18, 20, 27, 28, 31, 34, 36], released between mid-2024 and 2025, are crawled from reputable journals and institutional sources. Only statistical figures are extracted and cropped, excluding surrounding text, to isolate visual content that demands careful interpretation.

Performance. For the Figure Understanding performance as shown in Figure 23, all close-source models achieved high performance, each exceeding 60% accuracy. Interestingly, within the open-source models, InternVL-2.5 [3], InternVL-2[3], and InternVLM-XComposer [37] also reached satisfactory results, performing on par with the closed-source models.

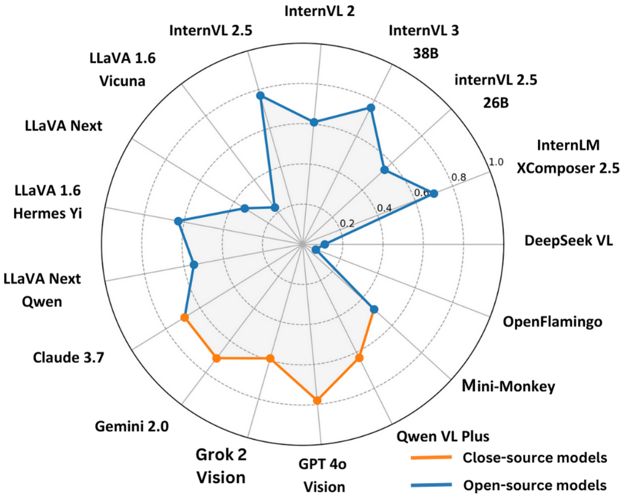


Figure 23. Model accuracy for the Figure Understanding task.

C.3.2. Book Understanding

This dimension assesses the VLMs' capability to comprehend and interpret information presented on a book page that combines both textual and visual elements. Models are prompted to read descriptive paragraphs, analyze accompanying figures or diagrams, and integrate these sources to answer questions.

Data collection. We collect pages from the three marine identification books: "Reef Fish Identification - Tropical Pacific", "Reef Creature Identification - Tropical Pacific", and "Nudibranch & Sea Slug Identification - Indo-Pacific" [1, 11, 14]. From these collected materials, we generate question–answer pairs based on the book content.

Performance. All models perform poorly on this task, with accuracy scores remaining below 40%, highlighting significant limitations in their ability to handle fine-grained or domain-specific reasoning.

C.3.3. Paper Understanding

This dimension assesses the VLMs' ability to comprehend and extract information from scientific papers. For example, models are asked to interpret overall architecture diagrams, summarize experimental results, or explain methods and conclusions based on texts, figures, and tables within the paper. From a given section or schematic, the model should infer the main ideas, describe the flow or design of the architecture, and clarify how its components interact. A visual testing sample is shown in Figure 25.

Data collection. 15 Recent open-access papers [2, 4–7, 9, 10, 16, 18, 20, 27, 28, 31, 34, 36], published between mid-2024 and 2025, are crawled from reputable marine science journals, websites, and institutional repositories. Documents are selected to combine complex textual content with scientific figures, diagrams, and tables. Pages with high visual and contextual complexity are prioritized to ensure that answering correctly demands genuine comprehension of the scientific material, rather than relying on pattern matching or random guessing by VLMs.

Performance. Figure 26 show that the open-source models InternVL [3], InternVL-2.5 [3], and InternVLM-XComposer

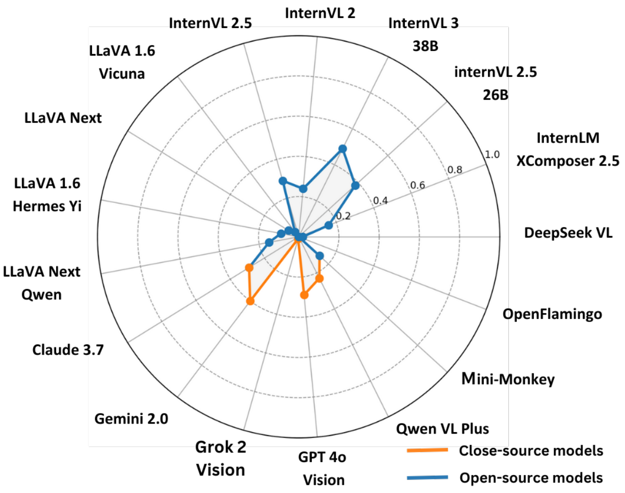
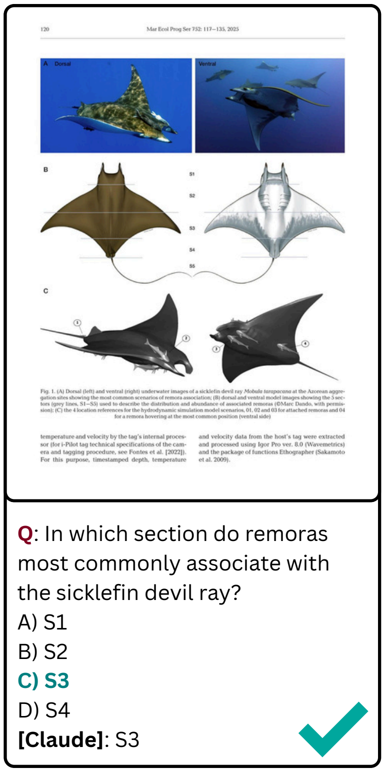


Figure 24. Model accuracy for the Book Understanding task.



120 Mar Ecol Prog Ser 752: 117–135, 2025

Fig. 1. (A) Dorsal (left) and ventral (right) underwater images of a sicklefin devil ray *Mobula tarapacana* at the Azorean aggregation sites showing the most common scenarios of remora association; (B) dorsal and ventral model images showing the 5 sectors (grey lines, S1–S5) used to describe the distribution and abundance of associated remoras (©Marc Dando, with permission); (C) the 4 location references for the hydrodynamic simulation model scenarios, 01, 02 and 03 for attached remoras and 04 for a remora hovering at the most common position (ventral side)

temperature and velocity by the tag's internal processor (for i-Pilot tag technical specifications of the camera and tagging procedure, see Fontes et al. [2022]). For this purpose, timestamped depth, temperature and velocity data from the host's tag were extracted and processed using Igor Pro ver. 8.0 (Wavemetrics) and the package of functions Ethographer (Sakamoto et al. 2009).

Q: In which section do remoras most commonly associate with the sicklefin devil ray?
A) S1
B) S2
C) S3
D) S4
[Claude]: S3 ✓

Figure 25. Question sample of the Paper Understanding task.

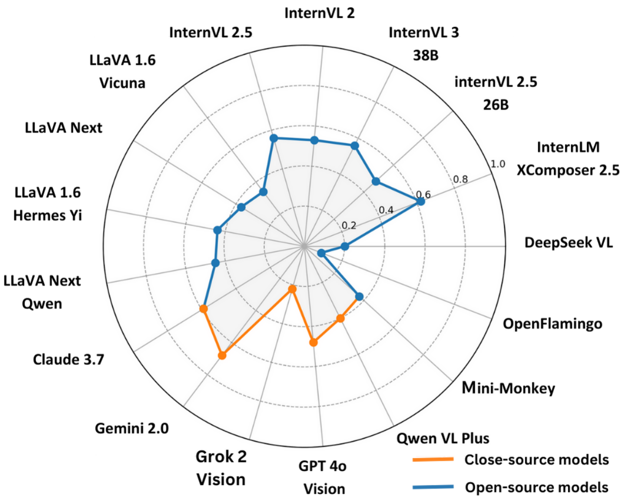
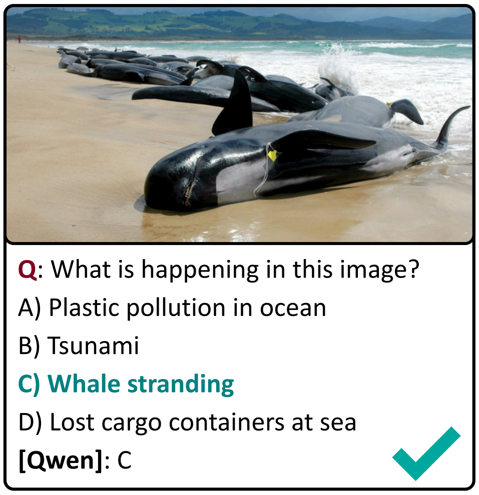


Figure 26. Model accuracy for the Paper Understanding task.



Q: What is happening in this image?
A) Plastic pollution in ocean
B) Tsunami
C) Whale stranding
D) Lost cargo containers at sea
[Qwen]: C ✓

Figure 27. Question sample of the Disaster Diagnosis task.

[37] outperform all of the close-source models. This advantage may come from the fact that these open-source models are trained with more specialized scientific or document-focused data. Their architecture is also better suited for aligning figures and text, which is critical for the paper understanding task.

C.4. Conservation & Threat Analysis

This capability focuses on identifying environmental threats and assessing the conservation status of organisms, aiding in ecological preservation efforts.

C.4.1. Disaster Diagnosis

This dimension evaluates the capability of VLMs to recognize and interpret marine disasters, such as shipwrecks, underwater volcanic eruptions, and tsunamis. We provide a visual sample shown in Figure 29.

Data collection. To construct a comprehensive Disaster Diagnosis Visual Question Answering dataset, a diverse range of real-world disaster scenarios is curated to ensure broad coverage of both natural and human-induced oceanic and coastal hazards. The dataset targets 15 representative categories: Oil tanker explosion, Sinking cruise ship, Capsized ferry, Lost cargo containers at sea, Plastic pollution in ocean, Coral bleaching, Melting polar ice caps, Underwater volcanic eruption, Toxic waste in ocean, Whale stranding, Ship collision, Coastal flooding disaster, Algae bloom disaster (red tide), Tsunami, and Fishing net entanglement.

All images are sourced from publicly available online repositories or reputable scientific news and environmental websites. To ensure copyright compliance and source transparency, the corresponding URL for each image is recorded as metadata. This guarantees traceability, proper attribution, and enables future auditing or updates of the dataset. The resulting image collection provides a robust foundation for generating visually grounded questions and answers that diagnose, describe, and reason about various marine and coastal disaster scenarios.

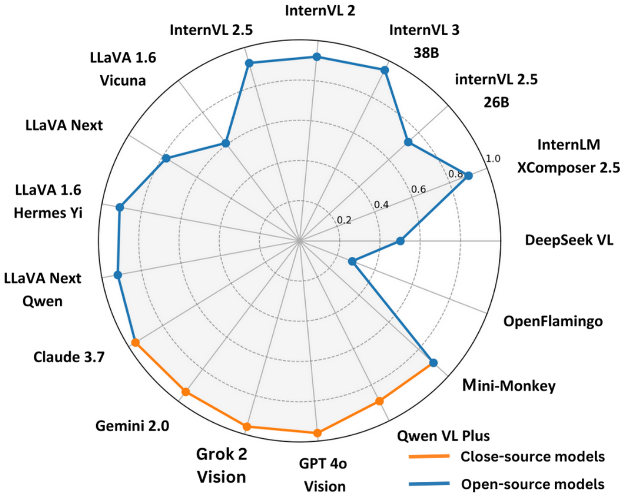


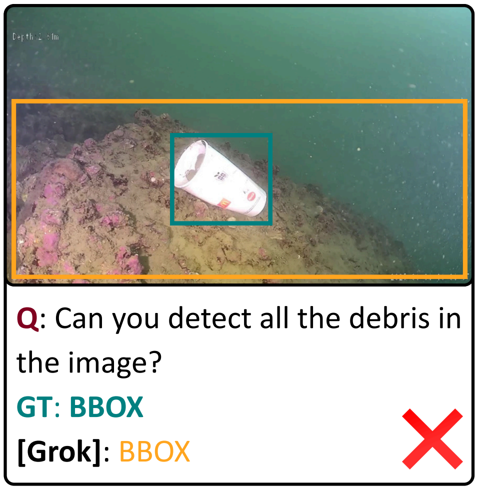
Figure 28. Model accuracy for the Disaster Assessment task.

Performance. Figure 28 illustrates the performance of various VLMs on the Disaster Diagnosis task. Notably, all closed-source models demonstrate exceptional results, with accuracy consistently exceeding 90%. This indicates that proprietary models maintain strong capabilities in recognizing and interpreting disaster-related scenarios. In contrast, some open-source models still show a large room for improvement, highlighting the performance gap that remains between commercial and community-driven solutions.

C.4.2. Pollutant Localization

This dimension evaluates the capability of VLMs to detect and localize anthropogenic pollutants, specifically trash (*e.g.*, plastics, debris) and oil on the ocean surface or trash under the sea.

Data collection. The dataset is self-collected from internet sources and split into tasks for both underwater trash



Q: Can you detect all the debris in the image?

GT: BBOX

[Grok]: BBOX

Figure 29. Question sample of the Pollutant Localization task.

localization and classification. The visual questions ask the model to identify the location of the trash, specify its type, and determine whether there is an oil spill present. The oil spill images are additionally sourced from the Oil Spill dataset [32].

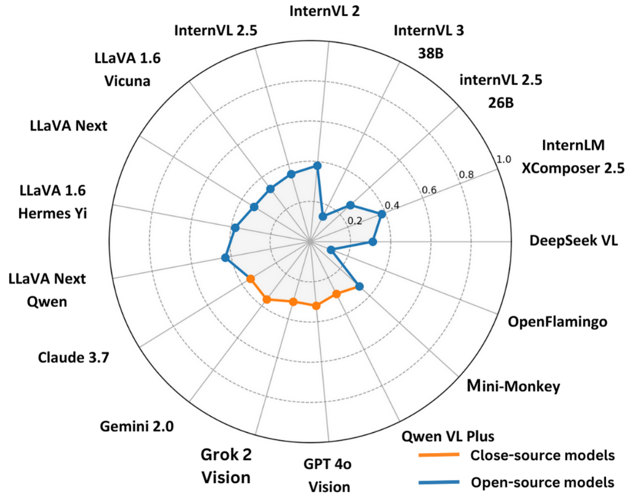
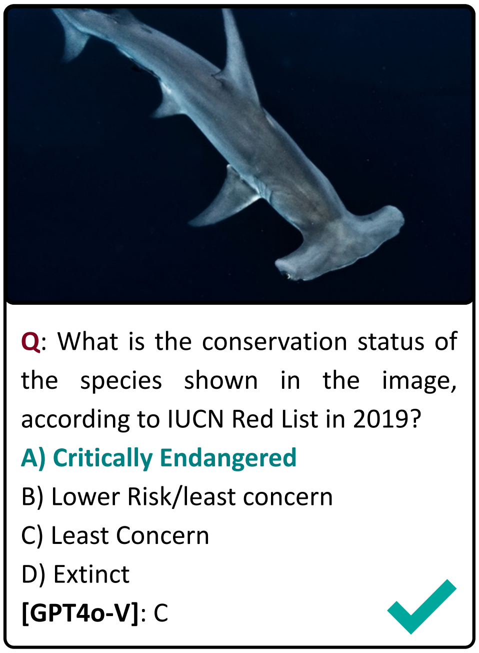


Figure 30. Model accuracy for the Pollutant Localization task.

Performance. The performance of all the evaluated VLMs is notably poor, with accuracy scores remaining below 40%. This result suggests that the models struggle significantly with the combined challenges of underwater trash localization and classification. A primary factor contributing to this low accuracy is likely the failure to correctly localize the scattered or partially occluded trash objects in complex underwater environments. Furthermore, the presence of oil spills adds additional visual noise, making it even more difficult for the models to distinguish different types of debris. These findings highlight the need for more robust region-level grounding and fine-grained recognition capabilities in current VLMs when applied to real-world marine pollution scenarios.

C.4.3. Threat-Status Determination



Q: What is the conservation status of the species shown in the image, according to IUCN Red List in 2019?

A) Critically Endangered

B) Lower Risk/least concern

C) Least Concern

D) Extinct

[GPT4o-V]: C

Figure 31. Question sample of the Threat-Status Determination task.

This dimension evaluates the capability of VLMs to identify marine organisms and ascertain the IUCN Red List [15] conservation status, including categories such as *Endangered*, *Critically Endangered*, *Vulnerable*, *Near Threatened*, or *Least Concern*. A visual sample is shown in Figure 31.

Data collection. To construct the testing data, images of marine species were collected from publicly available resources and the official IUCN Red List website [15]. Each image is paired with its verified conservation status according to the IUCN Red List categories [15]. The testing data covers a broad spectrum of species, ranging from commonly observed marine life to rare and critically endangered organisms, ensuring diversity in both species representation and conservation categories.

The categories include: “Extinct”, “Critically Endangered”, “Endangered”, “Vulnerable”, “Near Threatened”, “Least Concern”, “Lower Risk/conservation dependent”, “Lower Risk/near threatened”, “Lower Risk/least concern”.

Performance. Overall, most closed-source models struggled to accurately identify marine organisms and correctly assign their IUCN Red List conservation status [15]. This underperformance may be attributed to a lack of exposure to specialized biodiversity and conservation data during pretraining, which limits their ability to recognize rare or less visually distinct species and to link them with the appropriate conservation category. Interestingly, one closed-source model, LLAVA-NExt [19], stood out by achieving an accuracy exceeding 90% on this dimension. This suggests

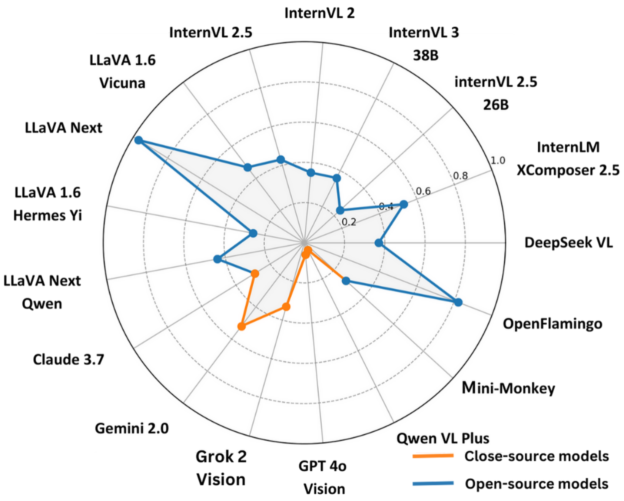


Figure 32. Model accuracy for the Threat-Status Determination task.

that LLAVA-NExt [19] may have benefited from more extensive fine-tuning on scientific or ecological datasets, or more advanced reasoning capabilities that help bridge visual identification with external taxonomic or conservation knowledge.

C.5. Marine Technology Understanding

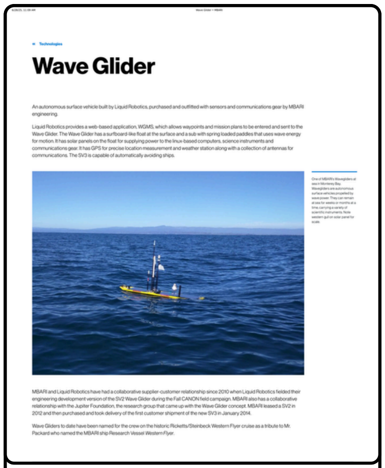
This capability assesses the ability of VLMs to recognize and infer the purpose of marine technology based on visual inputs.

C.5.1. Instrument Function Identification

This dimension assesses the VLMs' ability to recognize and infer the purpose of marine technology based solely on visual inputs. For example, models may be asked to identify the type of equipment, explain its function, or describe how it is used in marine exploration or monitoring tasks. Given an image of marine equipment, infer its intended operational purpose. A visual sample is shown in Figure 33

Data collection. For marine technology, images and articles published in 2024 and 2025 are crawled from marine technology magazines [22] and websites [21] to capture the latest technologies that are unlikely to overlap with the training data of vision-language models. GPT-4V [26] is first used to analyze each image and automatically generate candidate questions. A human-in-the-loop process was involved to verify and refine these question-answer pairs to ensure their validity and relevance for evaluation.

Performance. According to Figure 34, describing the Instrument Function Identification performance, Gemini-2.0, Claude-3.7, and QwenVL-2.5 achieve the highest scores, each exceeding 80%. Among the open-source models, InternVL [3], InternVL-2.5 [3], Internlm-XComposer [37], and Mini-Monkey [13] surpass two well-known closed-source models, GPT-4o [26] and Grok-2-Vision [33].



6/28/25, 11:09 AM Wave Glider • MBARI

« Technologies

Wave Glider

An autonomous surface vehicle built by Liquid Robotics, purchased and outfitted with sensors and communications gear by MBARI engineering.

Liquid Robotics provides a web-based application, WGMS, which allows waypoints and mission plans to be entered and sent to the Wave Glider. The Wave Glider has a surfboard-like float at the surface and a sub with spring loaded paddles that uses wave energy for motion. It has solar panels on the float for supplying power to the linux-based computers, science instruments and communications gear. It has GPS for precise location measurement and weather station along with a collection of antennas for communications. The SV3 is capable of automatically avoiding ships.

One of MBARI's Wavegliders at sea in Monterey Bay. Wavegliders are autonomous surface vehicles propelled by wave power. They can remain at sea for weeks or months at a time, carrying a variety of scientific instruments. Note western gull on solar panel for scale.

MBARI and Liquid Robotics have had a collaborative supplier-customer relationship since 2010 when Liquid Robotics fielded their engineering development version of the SV2 Wave Glider during the Fall CANON field campaign. MBARI also has a collaborative relationship with the Jupiter Foundation, the research group that came up with the Wave Glider concept. MBARI leased a SV2 in 2012 and then purchased and took delivery of the first customer shipment of the new SV3 in January 2014.

Wave Gliders to date have been named for the crew on the historic Ricketts/Steinbeck Western Flyer cruise as a tribute to Mr. Packard who named the MBARI ship Research Vessel *Western Flyer*.

Q: What is the primary propulsion method used by the Wave Glider shown in the image?
A) Solar power
B) Wave energy
C) Wind turbines
D) Diesel engine
[Claude]: Wave energy ✓

Figure 33. Question sample of the Instrument Function Identification task.

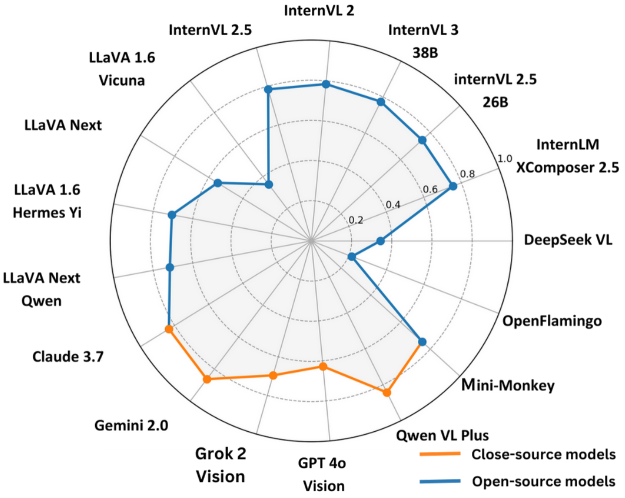


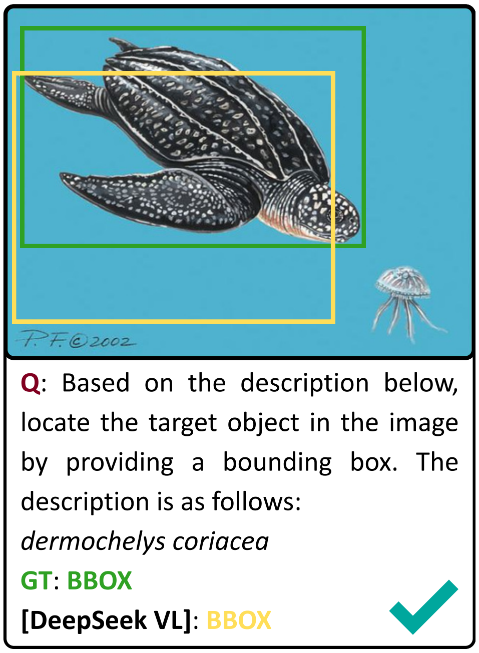
Figure 34. Model accuracy for the Instrument Function Identification task.

C.6. Spatial Reasoning

This capability examines visual-spatial reasoning skills to localize, count, or infer relative positions of objects within the image.

C.6.1. Visual Grounding

This dimension evaluates the model's ability to accurately locate and identify the species within an image that is referenced by a given textual query. The model must correctly match the description to the visual region containing the species and, if applicable, draw bounding boxes or highlight the relevant area to demonstrate precise localization. A visual testing sample is illustrated in Figure 35.



Q: Based on the description below, locate the target object in the image by providing a bounding box. The description is as follows: *dermochelys coriacea*

GT: BBOX

[DeepSeek VL]: BBOX ✓

Figure 35. Question sample for the Visual Grounding task.

Data collection. The data for this task comes from a private image captioning dataset that specifies detailed species traits and races. Each image is annotated with fine-grained textual descriptions linked to specific regions, allowing the evaluation to test whether models can accurately localize the specified species by matching traits, races, and visual cues within complex scenes.

Performance. Figure 36 shows the performance of different models in visual grounding tasks. Most of the closed-source models struggle to localize objects referred to in text, likely due to limited grounding-specific training or insufficient alignment between vision and language components.

The closed-source VLMs also show 0.00% average accuracy across the board. This surprising result suggests that even the strongest proprietary models lack robustness for our grounding benchmark, possibly because the task requires fine-grained spatial reasoning beyond their generic multimodal pretraining. These results reveal a critical need for specialized grounding datasets and model enhancements to improve marine object localization accuracy.

C.6.2. Numerosity Estimation

This dimension assesses the model's ability to count all relevant entities present in an image, such as fish and vessels. The model must accurately detect, enumerate, and report the total number of instances, demonstrating robust object detection and counting capabilities even in complex or cluttered scenes. A visual testing sample is shown in Figure 37.

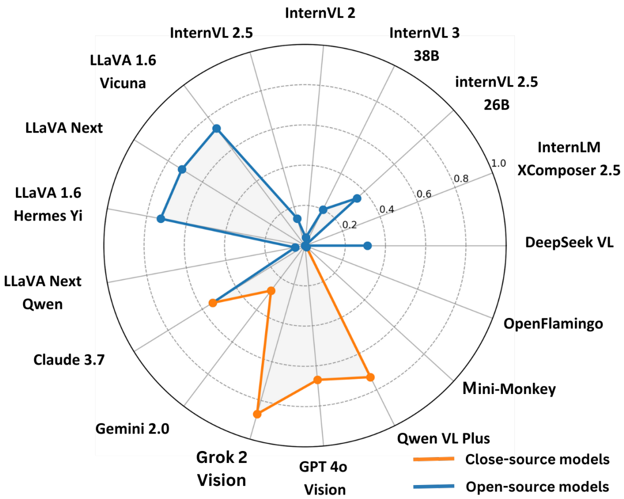
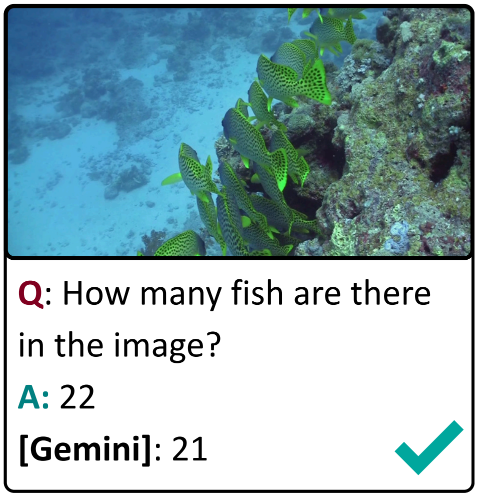


Figure 36. Model accuracy for the visual grounding task.

Difficulty	Description
Simple	Question is correct if the error percentage is less than 0.2.
Medium	Question is correct if the error percentage is less than 0.1
Hard	Question is correct if the error percentage is less than 0.05

Table 6. Definition of each difficulty in Species Identification.



Q: How many fish are there in the image?

A: 22

[Gemini]: 21 ✓

Figure 37. Question sample of the Numerosity Estimation task.

To fairly evaluate the performance, all the questions are split into three different difficulties, where the definitions are provided in Table 6.

Data collection. This dimension comprises two tasks: fish counting and ship counting. For the fish counting task, we derive our question-answer pairs directly from the IOCFish5k dataset [29], which provides precise annotations for fish counting. For the ship counting task, we generate question-answer pairs using the ShipRSImageNet dataset [38], which includes bounding box annotations for ships in satellite images. These annotations are transformed by counting the number of bounding boxes.

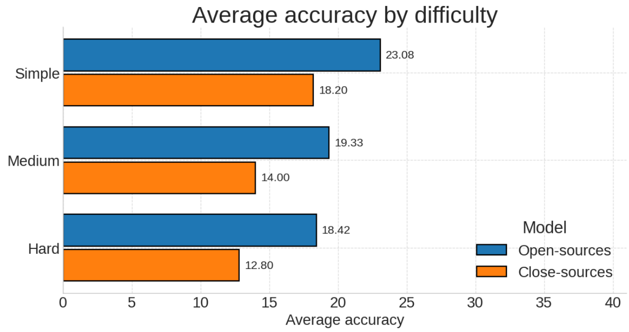


Figure 38. Model performance for the Numerosity Estimation task.

Performance. As depicted in Figure 38, the model's accuracy declines significantly as the difficulty level increases. This trend highlights that existing models struggle to deliver precise counting outputs, particularly in complex scenes or when objects are densely packed. Even open-source models trained on extensive datasets exhibit limited improvement, suggesting that current data-driven approaches alone are insufficient to address the challenges of accurate object counting.

C.6.3. Depth Ordering

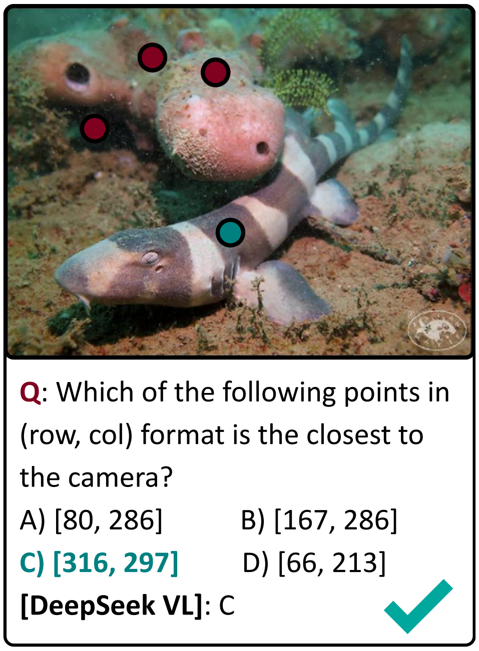
This dimension evaluates the model's ability to infer spatial relationships by determining the relative depth of objects within an image. Specifically, the model must identify which of the four designated points is closest to the camera, demonstrating an understanding of perspective, occlusion, and spatial cues. A visual sample is shown in Figure 39.

Data collection. Data for this task is collected from a private underwater dataset focusing on depth ordering. For each image, four points are manually labeled to represent different objects or reference spots. The VLMs are then challenged to determine which of these points is closest to the camera, testing their ability to reason about spatial depth and perspective in complex marine environments.

Performance. All models, illustrated in Figure 40, fail at this depth ordering task, and no model achieves over 20% of accuracy. This poor performance is likely due to the fact that models lack explicit training on fine-grained spatial reasoning and depth cues, especially when only 2D visual inputs are available without the dedicated 3D context or depth supervision.

C.6.4. Spatial Relation Assessment

This dimension tests the model's ability to infer and describe the relative spatial positions of two organisms or objects within an image. For example, the model must determine whether one organism is to the left/right of, above/below, or in front/back of another, demonstrating an understanding of spatial layout and positional relationships. The model is are to tell the spatial relationship of one species with respect to another. A visual sample is shown in Figure 41.



Q: Which of the following points in (row, col) format is the closest to the camera?

A) [80, 286] B) [167, 286]

C) [316, 297] D) [66, 213]

[DeepSeek VL]: C ✓

Figure 39. Question sample of the Depth Ordering task.

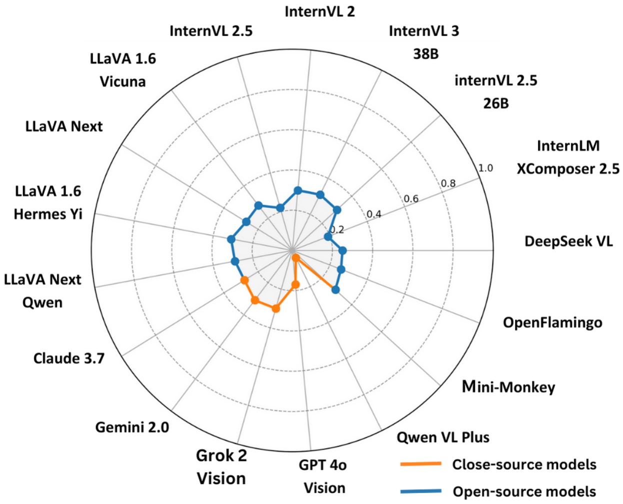
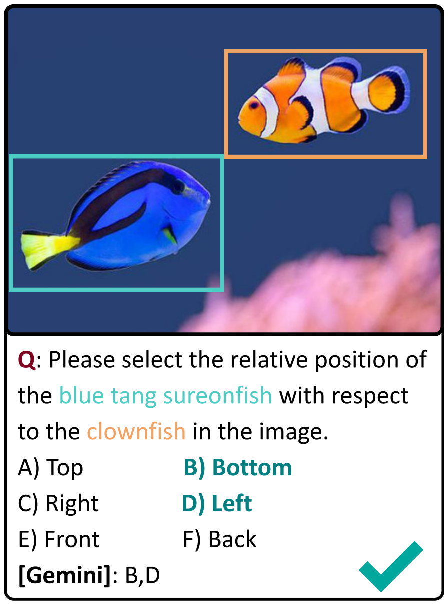


Figure 40. Model accuracy for the Depth Ordering task.

Data collection. The images are self-collected by the human annotators, who meticulously construct annotations for 100 question-answer pairs.

Performance. As illustrated in Figure 42, it is challenging for all models to perform accurately in tasks involving reasoning about relative positions. This observation highlights a significant gap in the spatial understanding capabilities of current VLMs. It suggests that the existing training strategies and datasets commonly used for mainstream architectures



Q: Please select the relative position of the blue tang sureonfish with respect to the clownfish in the image.

A) Top B) Bottom

C) Right D) Left

E) Front F) Back

[Gemini]: B,D

Figure 41. Question sample of the Spatial Relation Assessment task.

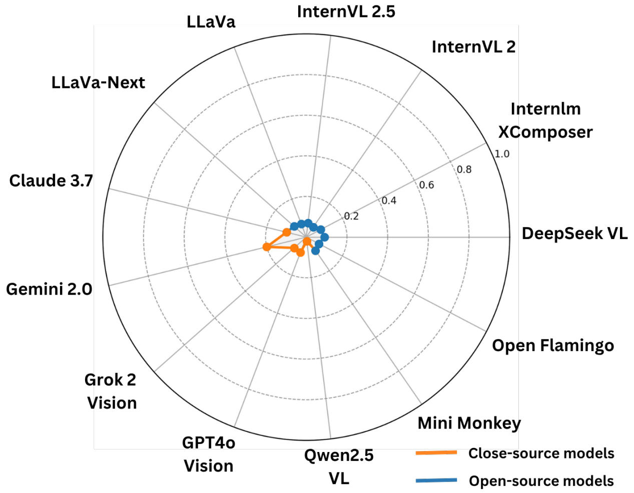
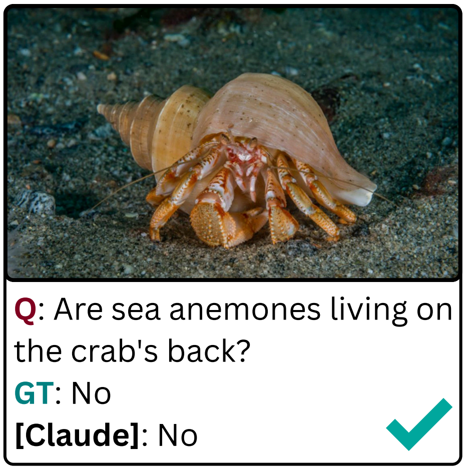


Figure 42. Model accuracy for the Spatial Relation Assessment task.

are insufficient for equipping models with robust spatial reasoning abilities. Addressing this limitation may require the development of specialized training frameworks, enhanced datasets that emphasize spatial relationships, or architectural modifications that prioritize spatial reasoning as a core capability.

C.7. Hallucination Resistance

This capability ensures that the model consistently refrains from generating unsupported, factually inaccurate, or hallucinatory content when responding to controlled and well-defined prompts, thereby upholding the reliability, factual integrity, and trustworthiness of its outputs in structured evaluation settings.



Q: Are sea anemones living on the crab's back?

GT: No

[Claude]: No

Figure 43. Question sample of the Hallucination Resistance task.

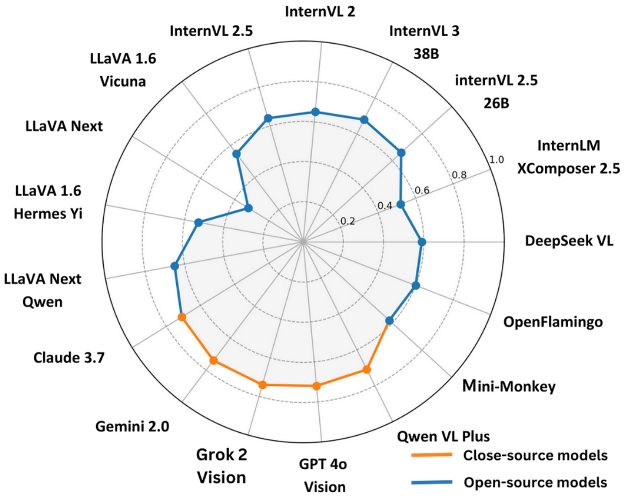


Figure 44. Model accuracy for the Hallucination Resistance task.

C.7.1. Hallucination Resistance

This dimension evaluates the VLMs' tendency to generate inaccurate or fabricated information when responding to questions about images or scientific content. For example, models are tested by deliberately posing misleading or trick questions, such as asking whether certain sea creatures are positioned next to each other due to mutualism or other ecological relationships, even when this is not visually evident. It evaluates whether the model generates unsupported or hallucinatory content under controlled prompts. A visual sample is shown in Figure 43.

Data collection. To build the hallucination test set, a list of ecological relationships among sea creatures is first collected, covering *mutualism*, *parasitism*, *commensalism*, *predation*, and *competition*. For each ecological relationship category, we selected ten distinct marine organism pairs, demonstrating that specific interaction. Images are gathered that contain only one species from each pair, and questions are generated to ask whether both species are present in the image. This experimental setup evaluates the model's tendency to hallucinate by testing whether it falsely claims the presence of the absent species based on the ecological pairing.

Performance. Figure 44 reports the accuracy of various VLMs. All closed-source models achieve approximately the same score of 70%, indicating that hallucination remains a common failure case for these well-known models. In addition, this demonstrates that the constructed dataset and benchmark effectively challenge these models and highlight this overlooked but important research area for further improvement.

D. Copyright And Dataset Usage

Copyright and Data Usage. MarineEval is compiled from a combination of openly available resources, including public datasets [12, 17, 23, 25, 30, 32, 35, 38], open-access journal articles [2, 4–7, 9, 10, 16, 18, 20, 27, 28, 31, 34, 36], search engine results, and self-collected samples. All third-party materials are limited to content that is publicly accessible and legally distributable under their respective licenses, and each source is explicitly cited or linked to its origin. To ensure **reproducibility** and maintain strict **licensing compliance**, the released dataset will include source information for all collected materials, along with evaluation prompts and scripts to facilitate transparent benchmarking and responsible community adoption. We release our dataset on the HuggingFace website.

References

[1] Gerald R. Allen. *Reef Fish identification: Tropical pacific*. New World Publications, 2015. 9

[2] 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. 8, 9, 16

[3] 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. 4, 7, 9, 12

[4] 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. 8, 9, 16

[5] 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.

[6] 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.

[7] 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. 8, 9, 16

[8] R. Froese and D. (Editors) Pauly. Fishbase. `https://www.fishbase.se`, 2025. Accessed: 2025-07-19. 5

[9] 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. 8, 9, 16

[10] 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. 8, 9, 16

[11] Terrence Goslinear, Angel Valdes, and David W. Behrens. *Nudibranch & Sea Slug Identification: Indo-Pacific*. New World Publications, 2019. 9

[12] 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, 4, 16

[13] 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. 12

[14] Paul Humann and Ned DeLoach. *Reef creature identification: Tropical pacific*. New World Publications, 2010. 9

[15] IUCN. The IUCN red list of threatened species. `https://www.iucnredlist.org`, 2022. Accessed on [day month year]. 11

[16] 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. 8, 9, 16

[17] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. *Asian Conference on Computer Vision*, 2020. 7, 16

[18] 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. 8, 9, 16

[19] 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. 11, 12

[20] 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. 8, 9, 16

[21] MBARI Technology News. Mbari technology news website. https://www.mbari.org/technology, . 12

[22] Marine Technology News. Marine technology news website. https://www.marinetechnologynews.com/magazine, . 12

[23] 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. 16

[24] 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. 8

[25] 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. 7, 16

[26] OpenAI. Gpt-4o system card, 2024. 12

[27] 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. 8, 9, 16

[28] 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. 8, 9, 16

[29] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13791–13801, 2023. 13

[30] 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. 16

[31] 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. 8, 9, 16

[32] 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. 11, 16

[33] xAI. Grok-2 beta release. https://x.ai/news/grok-2, 2024. 12

[34] 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. 8, 9, 16

[35] 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. 3, 4, 5, 16

[36] 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. 8, 9, 16

[37] 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. 4, 9, 10, 12

[38] 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. 14, 16