
CoralVQA: A Large-Scale Visual Question Answering Dataset for Coral Reef Image Understanding

Hongyong Han¹, Wei Wang^{1*}, Gaowei Zhang¹, Mingjie Li^{2,3}, Yi Wang^{1,4,5}

¹Beijing University of Posts and Telecommunications

²Technology Innovation Center for South China Sea Remote Sensing,
Surveying and Mapping Collaborative Application, Ministry of Natural Resources

³South China Sea Development Research Institute, Ministry of Natural Resources

⁴Inspur Computer Technology Co., Ltd

⁵Shandong Key Laboratory of Advanced Computing

{hanhongyong, weiwang, zhanggaowei, yiwang}@bupt.edu.cn, lmj_21@163.com

Abstract

Coral reefs are vital yet vulnerable ecosystems that require continuous monitoring to support conservation. While coral reef images provide essential information in coral monitoring, interpreting such images remains challenging due to the need for domain expertise. Visual Question Answering (VQA), powered by Large Vision-Language Models (LVLMs), has great potential in user-friendly interaction with coral reef images. However, applying VQA to coral imagery demands a dedicated dataset that addresses two key challenges: domain-specific annotations and multidimensional questions. In this work, we introduce CoralVQA, the first large-scale VQA dataset for coral reef analysis. It contains 12,805 real-world coral images from 67 coral genera collected from 3 oceans, along with 277,653 question-answer pairs that comprehensively assess ecological and health-related conditions. To construct this dataset, we develop a semi-automatic data construction pipeline in collaboration with marine biologists to ensure both scalability and professional-grade data quality. CoralVQA presents novel challenges and provides a comprehensive benchmark for studying vision-language reasoning in the context of coral reef images. By evaluating several state-of-the-art LVLMs, we reveal key limitations and opportunities. These insights form a foundation for future LVLM development, with a particular emphasis on supporting coral conservation efforts.

1 Introduction

Coral reefs are among the most biodiverse ecosystems, supporting over a quarter of known marine species (Knowlton et al., 2010; Fisher et al., 2015). They provide ecological value, support coastal economies through fisheries and tourism, and protect shorelines from storms and erosion (Mumby et al., 2008; Bitterwolf et al., 2024). However, human activities and climate change are causing unprecedented global declines (Steffen et al., 2007; Hoegh-Guldberg et al., 2019). Effective global coral reef conservation depends on continuous monitoring of benthic communities, with coral reef images serving as a direct and essential source of information. For instance, experts interpret these images to identify species, evaluate their health, and extrapolate findings to downstream systems. However, the capacity to derive insights from coral images is largely restricted to specialists in marine science. Coral reef monitoring involves dynamic, multiple tasks—including coral recognition, health diagnostics (*e.g.*, bleaching severity), and habitat quality assessment (*e.g.*, algal symbiosis), all of which demand extensive domain expertise for reliable image interpretation (Beijbom et al., 2012).

*Corresponding author

This barrier restricts the range and variety of problems that can be addressed with coral images (*e.g.*, coral studies in developing countries), and limits the number of potential users (*e.g.*, conservation practitioners, science educators). Therefore, there is a pressing need for new automatic ways that can extract relevant information from coral images without requiring specialized expertise.

The Visual Question Answering (VQA) task—originally developed in computer vision as a user-friendly way for image-based queries—could help bridge this expertise gap (Antol et al., 2015; Wu et al., 2017; Kafle, Kanan, 2017; Lobry et al., 2020; Zheng et al., 2023; Li et al., 2024). For example, given a coral reef image (the corresponding coral reef image is located in the third row, fourth column in Figure 1) and a question (“Is the coral genus in the upper left corner susceptible to bleaching?”). VQA aims to give the correct answer (“Yes”). This transforms specialized ecological assessment into accessible, high-level semantic information. Recent advances in large vision-and-language models (LVLMs) (He et al., 2020; Liu et al., 2023; Bai et al., 2023; Li et al., 2023a; Chen et al., 2024) have demonstrated state-of-the-art performance across various VQA tasks for natural images. However, their extension to the coral reef domain remains limited, primarily due to the absence of comprehensive, high-quality VQA datasets tailored to coral conservation. The construction of coral reef VQA datasets presents substantially greater challenges compared to general-domain VQA, due to two critical factors. (1) **Domain-specific annotations.** While advances in autonomous underwater vehicles have enabled the large-scale collection of coral reef images, existing annotations are inadequate for supporting VQA dataset development. Key issues include: inconsistent label standards—some corals are labeled by morphology, others by genus, and some at the family level; and inclusion of non-coral categories (*e.g.*, sand and sediments). (2) **Multidimensional questions.** Current VQA datasets for natural images primarily focus on generic visual concepts (*e.g.*, animals, vehicles) and basic attributes (*e.g.*, quantity, color). This domain-general characteristic makes question-answer pair generation relatively straightforward for non-experts. However, creating coral VQA datasets requires interdisciplinary expertise in both visual content and textual questions. Additionally, to support coral reef conservation, such a dataset must include open-ended question-answer pairs that span multiple ecological and health-related dimensions, such as coral condition, growth status, and symbiotic relationships.

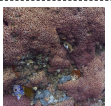
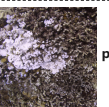
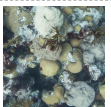
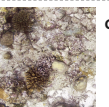
Coral Category	Coral Color	Coral Presence	Coral Position
 What coral genus is present in the lower left corner of the image? Montastrea	 What is the color of the coral at the left of the image? Red	 Is there any Acropora coral present on the left side of the image? Yes	 Where is Stylophora coral located in the image? Left
Coral Shape	Coral Size	Coral Quantity	Scene Type
 What is the morphology of the coral on the left side of the image? Branching	 What is the largest coral in the image? Lobophyllia	 How many coral species are present in the image? Three	 What scene is the image taken from? Underwater
Coral Growth Status	Coral Symbiotic Relationship	Coral Structural Integrity	Coral Health Status Existence
 Is coral growth evident in the image? Yes	 Can symbiotic zooxanthellae or fish be detected in the image? Yes	 Is the coral structure intact in the image? No	 Are there any bleached corals present in the image? Yes
Algal Existence	Coral Health Status Quantity	Coral Bleaching Susceptibility	Coral Habitat Environment
 Is the coral in the center covered by algae? Yes	 How many coral species have undergone bleaching in the image? One	 Is the coral genus in the upper left corner susceptible to bleaching? Yes	 Can the presence of trash be detected within the image? Yes

Figure 1: Example of coral reef image and corresponding question-answer pairs across 16 dimensions in CoralVQA dataset. The basic visual attributes of coral reef images are denoted by light orange boxes, while ecological and health-related attributes of coral reef images are indicated by light green boxes.

To address the above challenges, our interdisciplinary team introduces CoralVQA, a large-scale dataset for vision-language understanding of coral reef images, developed through close collaboration with marine biologists possessing extensive expertise in coral conservation. Through data collection,

label cleaning and re-annotating, CoralVQA consists of 12,805 high-quality real coral images from multiple marine regions across 3 oceans, including the Atlantic, Indian Ocean, and Pacific Ocean. All coral instances are re-annotated at the genus level, in accordance with the biological taxonomy hierarchy (*Kingdom-Phylum-Class-Order-Family-Genus*), covering 67 distinct genera from 20 families. CoralVQA encompasses a comprehensive and diverse collection of questions, systematically organized into two key groups: **basic visual interpretation; ecological and health-oriented assessment**. Each group comprises eight distinct dimensions. We benchmark several state-of-the-art LVLMs and find that CoralVQA presents novel challenges for coral-specific vision-language reasoning. Notably, we observe a significant drop in model performance when answering questions involving images from previously unseen ocean regions or those requiring complex reasoning, such as assessing bleaching coverage. Some examples from CoralVQA are shown in Figure 1. The key contributions of our work are summarized as follows.

- To the best of our knowledge, CoralVQA is the first large-scale VQA dataset dedicated to coral reef understanding. It contains 12,805 real coral images across 67 genera from 3 oceans, and 277,653 question-answer pairs from 16 dimensions.
- We design a semi-automatic vision-language coral data collection pipeline, which includes six steps: dataset collection, label cleaning and re-annotating, attribute extraction, question generation, and human verification. Our pipeline can be widely applied to other marine domains.
- We systematically evaluate the performance of several state-of-the-art LVLMs on CoralVQA across multiple coral-related tasks, serving as a baseline and highlighting opportunities for future research.

2 Related Works

Coral datasets are fundamental for studying coral reef conservation and assessing coral reef health. Existing works on coral datasets mainly focus on classification or segmentation. For example, Tasmania Coral Point Count (TasCPC) (Meyer et al., 2011) contains 1,258 AUV-captured benthic images across 13 underwater object categories (*e.g.*, corals, sand, rock), though coral taxonomic details are unavailable. RSMAS (Shihavuddin et al., 2013) is composed of 766 image patches (256×256 pixels) from 8 coral genera, collected by divers from the University of Miami. Benthos15 (Bewley et al., 2015), collected by AUVs from Australian, includes 9,874 labeled images. Despite having 148 categories, only two coral genera follow the Linnaean system. EILAT (Shihavuddin, 2017) provides 1,123 Red Sea image patches (64×64 pixels), which represent four morphological coral classes, along with categories for favid coral, dead coral, sand, and urchin. ATCRC (Rashid, Chennu, 2020), collected from Curaçao, contains 147 hyperspectral images representing 6 coral genera. HSCR16K (Han et al., 2025) contains 16,659 image patches (224×224 pixels) of 10 genera with rich text knowledge. MLC (Beijbom et al., 2012), collected from the island of Moorea, consists of 2,055 coral reef images, including 5 coral genera and 4 non-coral classes. CoralSCOP (Zheng et al., 2024), an important step in coral segmentation, contains 41,297 images from sources like YouTube but focuses on segmentation and lacks coral taxonomic details. In this work, we present CoralVQA, a multi-regions, multi-genera, and multi-dimensions VQA dataset, accompanied by comprehensive benchmarks under diverse VQA scenarios.

Table 1: The comparisons of existing coral datasets.

Task	Dataset	Images	Genera	QA pairs
Cls.	TasCPC	1,258	*	×
Cls.	RSMAS	766	8	×
Cls.	Benthos15	9,874	2	×
Cls.	EILAT	1,123	*	×
Cls.	ATCRC	147	6	×
Cls.	HSCR16K	16,659	10	×
Seg./Cls.	MLC	2,055	5	×
Seg.	CoralSCOP	41,297	*	×
VQA	CoralVQA	12,805	67	277,653

*: Genera are not following the standard Linnaean system.

3 Pipeline

To build CoralVQA, we design a semi-automatic vision-language data pipeline ensuring utility and expert-level quality. It includes six stages (see Figure 2): collecting real-world coral reef images; label

cleaning and re-annotation; extracting enriched visual and ecological attributes; designing question generation prompts; auto-generating question-answer pairs and human verification.

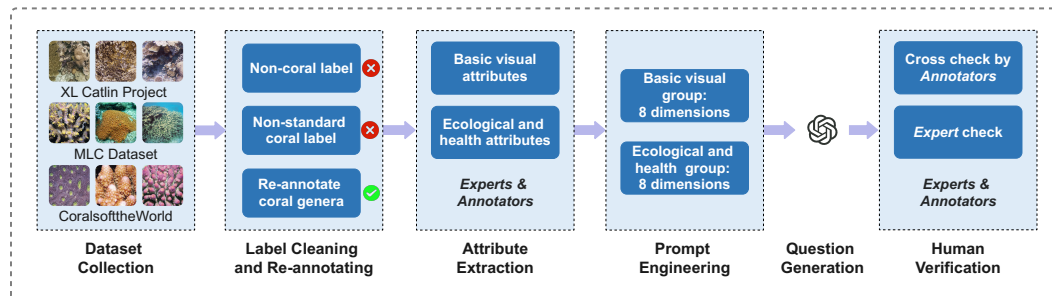


Figure 2: Pipeline for dataset creation. CoralVQA is developed through a six-stage semi-automated vision-language workflow: dataset collection, label cleaning and re-annotating, attribute extraction, prompt engineering, question generation, and human verification.

3.1 Dataset Collection

Our CoralVQA collects the real-world coral reef images from 3 sources: MLC dataset (Beijbom et al., 2012), XL Catlin Seaview Survey Project (González-Rivero et al., 2019), and the CoralsoftheWorld database (available at coralsoftheworld.org). The XL Catlin Seaview Survey Project (recording the health of coral reefs worldwide), launched in 2012, includes 11,387 images from the Atlantic, Indian, and Pacific Oceans, providing extensive geographic coverage. To enhance both image quantity and taxonomic coverage, we crawled additional 3,420 images from CoralsoftheWorld, a globally recognized source for scientifically validated coral taxonomy and distribution. This resulted in an initial dataset of 16,862 images. After the quality filter using the Underwater Color Image Quality Evaluation metric (Yang, Sowmya, 2015), the final dataset contains 12,805 high-quality images.

3.2 Label Cleaning and Re-annotating

Despite the large volume of collected coral reef images, inconsistent annotation standards and non-coral categories pose challenges for direct downstream use. Therefore, we apply a systematic cleaning and re-annotating process guided by the following criteria: 1) removing the coral labels without following the standard coral biological taxonomic system (e.g., branching *Acroporidae* and hispidose *Acroporidae*); 2) eliminating non-coral labels such as sand, fishing gear, and sediments; 3) re-annotating coral instances based on a genus-level coral biological taxonomy hierarchy (from *Kingdom* and *Phylum* to *Family* and *Genus* and *Species*). Marine biologists, led by the fourth author, conducted this step. During the process, 50,200 non-coral annotations and location entries were removed, while 97,395 coral-related annotations and their corresponding locations were re-labeled. In this step, the major issue is inconsistent annotation standards and incorrect coral annotations. To address the problem, we have every image re-examined by three marine scientists after the label cleaning and re-annotating step, and adopt a majority voting approach to determine the final annotations. In fact, only a very small set of images (203 out of 12,805, <2 %) contains errors. The final curated dataset contains 12,805 images spanning 20 families and 67 genera.

3.3 Textual Attribute Extraction

We extract enriched attributes from the coral images to facilitate the subsequent automatic generation of question-answer pairs. Guided by the practical needs of coral reef monitoring and conservation, the extracted attributes are organized into two key groups: 1) basic visual attributes and 2) ecological and health-related attributes. Following the paradigm of general VQA tasks, we extract basic visual attributes—such as coral genera, position, and quantity—using automated scripts based on the existing annotation files. In addition, ecological and health-related attributes—including coral health status, growth condition, and symbiotic relationships—are manually annotated for each image.

3.4 Prompt Engineering

We carefully develop an automated prompt template to guide GPT-4o in generating detailed question-answer pairs. To enable a comprehensive evaluation and analysis of coral reef images, we formulate questions from two groups spanning 16 distinct dimensions. **Basic visual group** (1) coral category: identify the coral genus based on image position; (2) coral color: identify the coral color at a specific image position; (3) coral presence: determine if coral is present at a given location or if a specific coral exists in the image; (4) coral position: output the position of a specific coral in the image; (5) coral quantity: count the total number of coral genera represented in the image; (6) coral size: identify the dominant or minimal coral genus present in the image; (7) coral shape: describe the coral morphology at a specified position of image; (8) scene type: identify the benthic habitat type in the image (sandy, rocky, or other substrates). **Ecological and health-related group** (1) coral growth condition: determine if the coral exhibits active growth; (2) algal presence: determine if algae cover surrounds the coral (corals and algae exhibit a competitive relationship); (3) coral symbiotic relationship: determine if the coral maintains a stable symbiotic relationship with zooxanthellae; (4) health status existence: determine if coral bleaching or coral disease has occurred at a specific position in the image; (5) coral structural integrity: analyze whether the coral's skeletal structure is intact or if there are signs of damage. (6) health status quantity: count the number of coral genera showing signs of bleaching or disease; (7) bleaching susceptibility: identify if the coral genus belongs to a bleaching-susceptible genus; (8) coral habitat environment: evaluate coral environmental conditions: water clarity, presence of debris, and potential contaminants. To enhance the diversity of the generation question, we carefully design multiple sets of prompts with varying language styles and semantic content for coral images from different data sources.

3.5 Question-Answer Generation

We utilize the GPT-4o API to automatically generate question-answer pairs. Existing VQA data generation approaches typically rely on feeding textual inputs into large language models. In contrast, we leverage OpenAI's multimodal capabilities by uploading both textual attributes and coral images through the GPT-4o image API, enabling more authentic and context-aware question-answer generation grounded in visual content. Specifically, based on the image coordinate system, we first divide each image into a 3×3 grid. Using the coral coordinate annotations, we determine the corresponding grid region for each coral instance. The GPT-4o API then automatically assigns directional indicators (*e.g.*, upper left, center, lower right) to support the generation of location-related questions in downstream tasks. To ensure diversity in question generation, we constrain each question type to appear no more than three times per image. Additionally, we employ the GPT-4o API to perform diversity-oriented paraphrasing, generating questions that differ in linguistic expression while retaining similar semantic intent. We further iteratively refine the prompt templates to better capture domain-specific terminology and improve the overall quality of generated question-answer pairs in the context of coral reef understanding. In this step, the major issue is about the reproducibility issues of GPT-4o generated question-answer pairs. To address the issue, we set the lower temperature parameter of 0.3 and incorporate targeted prompt phrases that explicitly require definitive answers. We randomly sampled 1,000 questions, with each question being answered 10 times by GPT-4o. The average number of consistent answers was 8.2 times. Although GPT-4o generated answers exhibit minor inconsistencies, since all question answers undergo multiple rounds of manual verification, the impact of answer inconsistencies is further mitigated.

3.6 Human Verification

Even with carefully designed prompts, errors persist in the GPT-4o API's question-answer pairs. To address these concerns, we perform a three-stage process: 1) manual verification, 2) cross-checking, and 3) expert sampling inspection. In the first stage, twelve students with backgrounds in marine science are recruited to verify the coral image and question-answer pairs manually and fix the errors resulted from using GPT-4o. Each of them needed to deal with roughly 1,000 images. In total, there were 12 subsets of images and corresponding question-answer pairs. Corrections are made based on the following criteria: removing hallucinated questions unrelated to the image; fixing incorrect spatial references within questions; revising uncertain and inaccurate answers. After the first stage, most issues are effectively resolved. To further enhance dataset quality, we implement a cross-checking procedure wherein each verified question-answer pair is reviewed by a second annotator. Each student

cross-checked the subset of images verified by another student. If the consistency of cross-checking is lower than 95% in any specific subset, a third inspector was assigned to check these inconsistencies again and perform necessary fixes. In the third phase, domain experts inspected 10% images and corresponding question-answer pairs randomly chosen from each subset. If the accuracy of a subset's sample was lower than 95% (judged by the assigned expert), that subset had to be re-verified again. We compiled statistics on the manual pruning of GPT-4o responses. During the human verification stage, we modified 13.4% of the questions generated by GPT-4o and 56.8% of the answers generated by GPT-4o required manual pruning.

4 Dataset Statistics

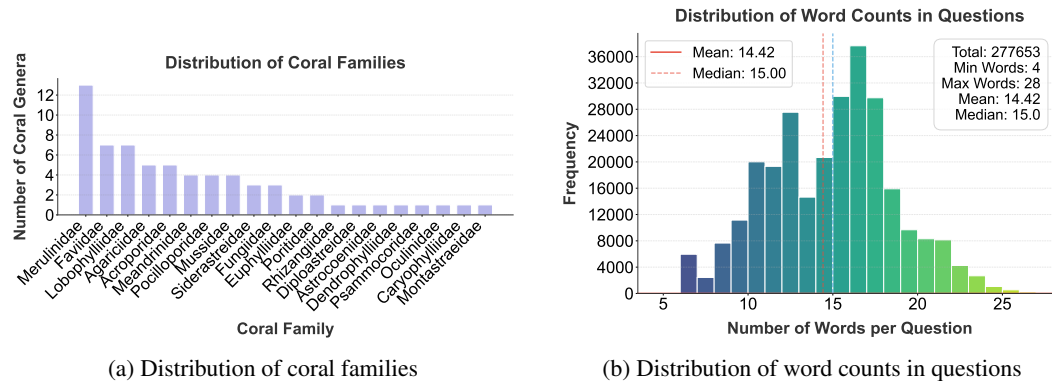


Figure 3: (a) The distribution of coral genera with different coral families in the CoralVQA dataset. (b) Distribution of the word length of questions.

4.1 Image Statistics

The CoralVQA dataset contains 12,805 underwater coral reef images and Figure 1 shows some examples. Most of the images are high resolution, with an average pixel width of 1,350 and a pixel height of 1,280. Based on the locations of coral reef survey points, the coral reef images are collected from multiple marine regions across three oceans. These regions include Island of Moorea (MLC), the Atlantic (ATL), the Indian Ocean and Chagos Archipelago (IND_CHA), the Indian Ocean and Maldives (IND_MDV), the Pacific Ocean and USA (PAC_USA), the Pacific Ocean, Indonesia and Philippines (PAC_IDN_PHL), the Pacific Ocean and Solomon Islands (PAC_SLB), the Pacific Ocean and Taiwan (PAC_TWN), and the Pacific Ocean and Timor-Leste (PAC_TLS). The geographic diversity of marine regions significantly enhances the taxonomic richness of coral genera within the dataset, while also enabling cross-regional evaluation of large multimodal models. CoralVQA, comprising 67 coral genera across 20 families, represents the most taxonomically diverse coral dataset to date. The distribution of these genera across their corresponding families is illustrated in Figure 3a. Our code and dataset are publicly available at <https://huggingface.co/datasets/CoralReefData/CoralVQA/tree/main>.

4.2 Questions-Answers Statistics

CoralVQA contains 277,653 question-answer pairs with 16 question types. The question types in the dataset are categorized into two main groups: (1) basic visual attributes and (2) ecological and health-related attributes. On average, each coral reef image has 21.6 questions. Figure 3b illustrates the distribution of question lengths in terms of words, with each question containing an average of 14.4 words. Out of the 16 question types, eight are open-ended, including coral category, color, position, quantity, size, shape, scene type, and health status quantity. These open-ended questions make up 42.1% of all question-answer pairs, totaling 116,960 questions. The diversity of answers to open-ended questions and the variety of visual features in coral reefs present significant challenges for visual language models in coral reef VQA tasks. The remaining question types consist of closed-ended “yes/no” questions, with 73,095 questions having “yes” responses and 87,598 with “no” responses. The distribution of different question types is shown in Figure 4.

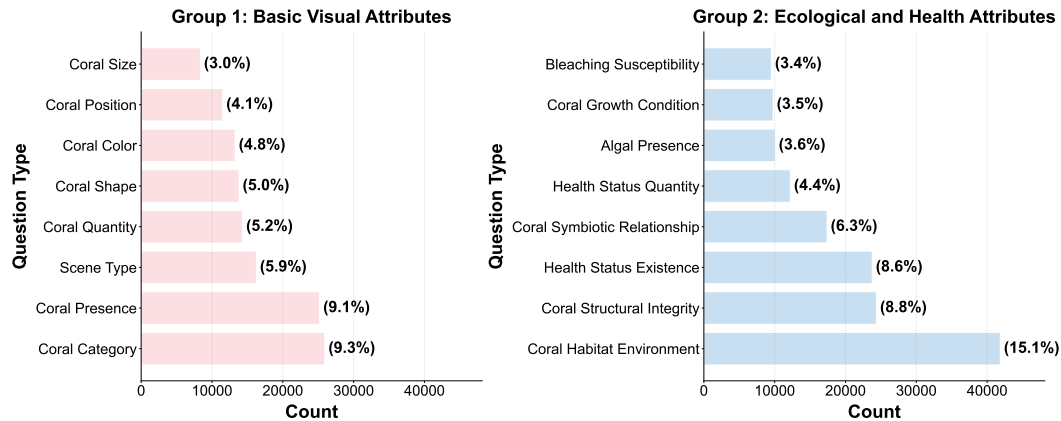


Figure 4: Distribution of two question groups.

5 VQA Benchmark Performances

5.1 Implementation Details

5.1.1 Models

To establish benchmark results, we conduct evaluations on six representative large vision-language models (LVLMs). Among them, the open-source models are Mini-Gemini(7B) (Li et al., 2023b), Qwen2.5VL(7B) (Bai et al., 2023), BLIP3 (Le Xue, 2024), and InternVL2.5(8B) (Chen et al., 2024). In addition, we also include two leading closed-source models, GPT-4o (Hurst et al., 2024) and Claude3.5 Haiku (Anthropic, 2024). For the Mini-Gemini, we utilize the OpenCLIP ConvNext (Ilharco et al., 2021) as the auxiliary vision encoder and use a 2-layer MLP with GeLU activation for projecting visual features. Following the BLIP3, we utilize Microsoft’s Phi3 (Abdin et al., 2024) as the base language model and use Google’s SigLIP-SO400M (Zhai et al., 2023) as the vision encoder. Other architectures adhere to the original models. For a fair comparison, we initialize all open-source models with their own pre-trained weights. Additional training details are provided in the appendix.

5.1.2 Coral VQA Tasks

We divide CoralVQA into three non-overlapping subsets to enable comprehensive evaluation: a training dataset, a testing dataset, and a cross-region dataset (marine region near Hawaii). Detailed statistics are presented in Table 2. We then establish three tasks related to coral reef conservation. (1) **VQA on test dataset**: Evaluation on the test dataset to establish standard performance benchmark. (2) **VQA on cross-region dataset**: Evaluation on the cross-region dataset, designed to examine model generalization to unseen regions. (3) **VQA on bleaching-coverage dataset**: Evaluation on our novel small-scale bleaching-coverage dataset (detailed in Section 5.4), which tests the model’s ability to perform a complex reasoning task. We fine-tune all open-source models on the training set and subsequently evaluate their performance on different tasks.

Table 2: Statistics of train, test, and cross-region dataset. “Q” and “A” denote question and answer.

Items	Train	Test	Cross
Images	10,537	1,274	994
QA pairs	226,726	27,984	22,943
Avg words per Q	14.45	13.25	15.58
Avg words per A	1.12	1.01	1.04

5.2 VQA Results on Test Dataset

Settings and Evaluation Based on the question types in the CoralVQA dataset, we divide the testing set into 16 subsets and measure the model’s performance using the accuracy as the metric. To gain a more complete understanding of the model’s performance, we calculate the average accuracy separately from two groups: basic visual attributes, and ecological and health attributes. We utilize the GPT-4o API to evaluate whether the semantics of the ground-truth answer from CoralVQA match

those of answers from the models. For a comprehensive description of the calculation methods, please refer to the appendix.

Results As shown in Table 3, the following three conclusions can be drawn. First, compared to other methods, InternVL2.5 achieves superior average performance across both groups. In addition, we further perform the experiments with two closed-source models: GPT-4o and Claude-3.5 Haiku. Although existing closed-source models have achieved good performance on general tasks, they perform poorly on coral-related tasks. Second, the accuracy of all models on open-ended questions is significantly lower than on closed-ended questions, with a performance gap exceeding 10%. For instance, the accuracy of InternVL2.5 for scene-type questions is 14.74% higher than for category-related questions. Similarly, Qwen2.5VL shows a 39.70% higher accuracy in addressing scene-type questions compared to category-related questions. These results suggest that open-ended questions often require domain-specific knowledge of coral categories and morphological characteristics, while current vision-language models pre-trained on general image-text datasets lack specialized knowledge in this area. Third, all methods show lower accuracy on questions regarding coral shape and quantity. In coral reef images, multiple coral genera are often distributed in interspersed patterns, with some genera being partially occluded. The limited ability of current visual-language models to extract textural, edge, and morphological features from coral reef images leads to enumeration errors, including both missed and duplicate counts.

Table 3: Visual question answering performance on test dataset. Boldface indicates the best performance.

Methods	Category	Presence	Quantity	Color	Position	Size	Shape	Scene	All
GPT-4o	71.35	50.38	7.04	70.09	43.13	54.26	57.47	85.83	54.94
Claude3.5 Haiku	69.27	52.26	5.92	43.08	48.37	62.63	38.36	47.54	45.93
Mini-Gemini(FT)	8.72	28.86	6.98	2.01	7.23	4.47	3.18	40.65	12.76
BLIP3(FT)	69.43	62.66	14.90	87.72	36.76	1.30	34.72	95.47	50.37
Qwen2.5VL(FT)	55.00	70.43	14.90	70.68	46.18	59.45	57.47	94.70	58.60
InternVL2.5(FT)	81.69	79.45	35.54	78.27	62.18	74.46	62.85	96.43	71.35
Methods	Growth	Algal	Presence	Quantity	Integrity	Susceptibility	Environment	Symbiosis	All
GPT-4o	52.81	42.88	25.21	14.78	33.68	28.68	61.84	87.48	43.42
Claude3.5 Haiku	44.00	38.48	42.96	34.11	18.05	30.64	62.00	30.64	37.61
Mini-Gemini(FT)	38.74	27.88	30.61	20.96	37.26	38.71	24.82	53.34	34.04
BLIP3(FT)	88.09	61.67	94.51	63.47	96.54	80.92	74.83	82.60	80.32
Qwen2.5VL(FT)	78.52	58.18	90.56	60.50	72.56	74.37	72.00	72.45	72.39
InternVL2.5(FT)	86.96	65.45	88.73	65.25	95.90	80.04	80.20	88.81	81.42

Table 4: Visual question answering performance on cross-region dataset. Boldface indicates the best performance. The “-” denotes that missing question-answer data for this dimension.

Methods	Category	Presence	Quantity	Color	Position	Size	Shape	Scene	All
Mini-Gemini(FT)	4.93	16.88	9.59	1.83	5.58	2.37	2.33	23.36	8.36
BLIP3(FT)	25.21	29.26	1.37	12.82	9.88	17.00	8.35	30.71	16.83
Qwen2.5VL(FT)	20.33	26.17	1.11	8.41	12.80	25.49	12.48	28.84	16.95
InternVL2.5(FT)	24.09	26.45	2.14	10.07	14.43	23.12	14.45	35.25	18.75
Methods	Growth	Algal	Presence	Quantity	Integrity	Susceptibility	Environment	Symbiosis	All
Mini-Gemini(FT)	18.65	15.48	18.84	17.13	17.38	-	11.78	20.64	17.13
BLIP3(FT)	32.35	26.03	32.06	8.80	31.07	-	38.03	33.80	28.88
Qwen2.5VL(FT)	31.82	24.10	30.12	2.78	26.04	-	33.50	31.54	25.70
InternVL2.5(FT)	31.93	26.21	31.20	6.25	29.89	-	36.58	35.71	28.25

5.3 VQA Results on Cross-Region Dataset

Settings and Evaluation Coral genera exhibit considerable intra-class variation in their composition, color, and morphology from different marine regions. Following the previous evaluation metrics, we adopt the accuracy as the metric to assess the model’s generalization performance on the cross-region dataset. Similar to the previous setting, we employ the GPT-4o API to assess the semantic consistency between cross-region dataset ground-truth answers and model-generated answers.

Results As shown in Table 4, compared to the results on the standard test dataset, the performance of nearly all models decreased by more than 30% in both groups. For instance, InternVL2.5 achieves only 2.14% accuracy on coral quantity-related questions and shows more than 10% drops in accuracy

for questions related to coral position, shape, and color. These results indicate that existing visual-language models exhibit limitations in generalization when handling coral images from diverse ecological or geographical environments.

5.4 VQA Results on Bleaching-Coverage Dataset

Settings and Evaluation Evaluating the proportion of coral bleaching is essential for determining the extent of impacted ecosystems and the number of affected coral populations. Existing methods of coral bleaching coverage assessment typically rely on image segmentation to obtain the bleached coral area and calculate the coral bleaching coverage ratio. However, these methods heavily rely on domain knowledge and involve multiple steps, making them limited in application and time-consuming. In this context, VQA offers a promising alternative. The evaluation of coral bleaching rates through VQA presents significant challenges. This process requires not only pixel-level semantic segmentation and quantitative analysis of coral regions in images, but also a comprehensive understanding of both the spatial distribution and the visual details within coral images. In this section, we introduce a new VQA dataset called Bleaching-Coverage. Specifically, we collect 309 bleached coral reef images from the CoralVQA dataset and build the 309 question-answer pairs about coral bleaching coverage. We employ the LabelMe tool to segment the regions of bleached coral genera and calculate the proportion of the bleached area to the entire surveyed area, thereby obtaining the ground-truth. We utilize the MAE (Mean Absolute Error) and MASE (Mean Absolute Scaled Error) to evaluate the error between the output from VQA models and the ground truth.

Results The mean absolute error (MAE) of Qwen2.5VL is 0.1124, whereas that of InternVL2.5 is 0.0818. The MASE for Qwen2.5VL and InternVL2.5 are 1.2326 and 0.8967, respectively. Compared with Qwen2.5VL, InternVL2.5 has lower values of MAE and MASE, indicating better generalize ability for complex reasoning tasks. However, all methods still exhibit substantial prediction errors. Furthermore, Mini-Gemini and BLIP3 fail to effectively comprehend the task-specific questions, often generating unrelated answers. These results suggest that existing visual-language models still face significant challenges in these complex tasks. In addition, we visualize the bleaching regions of corals predicted by InternVL2.5 in the appendix, which further illustrates the limitations of existing vision-language models in assessing bleaching coverage.

5.5 Further Discussion

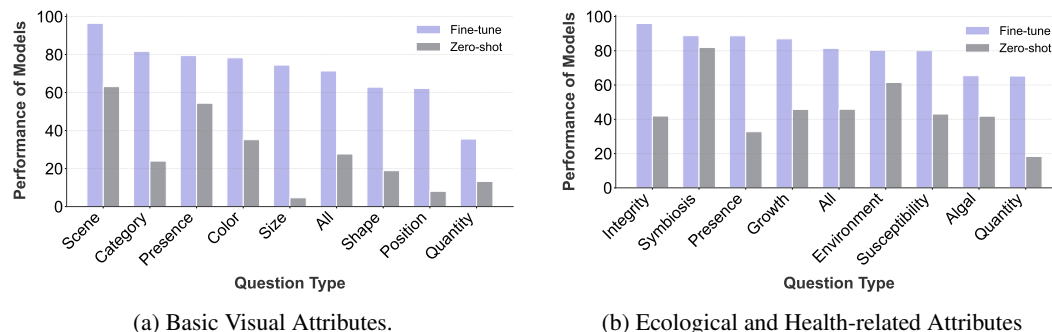


Figure 5: Performance comparison between fine-tuned and zero-shot evaluation on InternVL2.5 model. “All” represents the average accuracy across all question types.

Zero-shot Evaluation To further evaluate the real performance of existing visual-language models pre-trained on general image-text datasets in coral reef image understanding, we perform zero-shot evaluation on the InternVL2.5 models. Following the same evaluation method as before, we utilized the GPT-4o API to evaluate answer accuracy. As shown in Figure 5, two conclusions can be drawn. First, compared to the fine-tuning strategy, the average performance of zero-shot evaluation for InternVL2.5 shows varying degrees of degradation. For instance, InternVL2.5’s zero-shot evaluation shows performance decreases of 43.68 % and 35.54 % in the basic visual attributes, and ecological and health-related attributes, respectively. The result also indicated the effectiveness of fine-tuning strategies from the vision-language model. Second, for questions related to coral size, position, and quantity, the InternVL2.5’s question-answering accuracy falls below 20 %. The result indirectly

reflects the limited capability of current vision-language models in extracting texture, edge, and morphological features from coral reef images. More detailed results are provided in the appendix. **Case Study** To further investigate the reason behind the model's lower performance on questions related to coral size and shape, we conduct a case study. Focusing on InternVL2.5, as shown in Figure 6, we perform a visual analysis of its decision regions when answering questions about coral size and quantity. We design the following prompts to guide the model's responses: "What is the genus of the largest hard coral in the image? Provide a brief response and output the location coordinates." "How many coral genera are present in the image? Give a concise answer and specify their respective location coordinates." For coral size-related questions, the InternVL2.5 model incorrectly identifies non-target categories as the largest coral genus—mistakenly selecting regions labeled as other organisms, while the correct answer is the genus *Porites*, located at the top of the image. This mistake is mainly due to the model's lack of domain-specific knowledge, which hinders its ability to distinguish between coral genera and similar-looking categories such as algae. For coral quantity-related questions, the InternVL2.5 model is unable to identify the positions of individual coral genera, leading to incorrect counts. These results highlight the model's limitations in spatial reasoning, indicating that it lacks the fine-grained ecological understanding necessary for accurate differentiation among visually similar coral types. Furthermore, they demonstrate the necessity of developing novel frameworks for integrating marine biological knowledge into visual-language models to enhance their performance on coral analysis tasks.

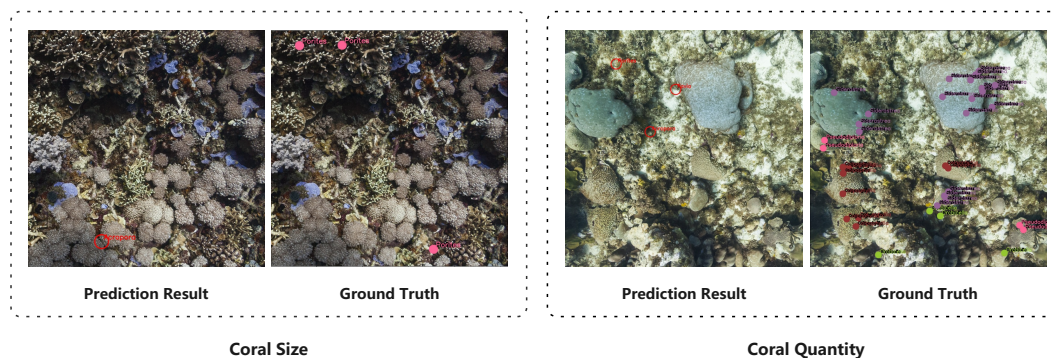


Figure 6: Visualization results. The hollow circles on the left indicate predicted positions of the model, while the solid circles on the right represent positions of the ground truth.

6 Conclusion and Future Works

In this paper, we introduce Visual Question Answering as an accessible way of extracting high-level semantic information from coral image data. To this end, we build a coral VQA dataset that includes 12,805 real coral reef images of 67 coral genera and 277,653 question-answer pairs of 16 dimensions. We further introduce a semi-automatic data collection pipeline, empowering researchers to extend and customize the coral dataset, and even generalize the framework to other marine ecosystems. Our benchmarking of several state-of-the-art LVLMs reveals significant challenges in coral VQA tasks, particularly when handling unseen marine regions and complex logical reasoning questions. These findings provide critical insights and establish new research directions for advancing vision-language models in coral research and protection.

7 Acknowledgments

This work is partially supported by National Natural Science Foundation of China under grants 62076232 and 62172049, BUPT Excellent Ph.D. Students Foundation under grant CX20251006, Science and Technology Development Foundation of South China Sea Bureau, Ministry of Natural Resources under grant 230208, Guangdong S&T Programme under grant 2025B1111130002, and Key Program of Marine Economy Development Special Foundation of Department of Natural Resources of Guangdong Province under grant GDNRC[2020]012. We sincerely thank the marine biologists in the fourth author's research team (Nansha Islands Coral Reef Ecosystem National Observation and Research Station) for their help on data collection and processing.

References

- Abdin Marah, Aneja Jyoti, Awadalla Hany, Awadallah Ahmed, Awan Ammar Ahmad, Bach Nguyen, Bahree Amit, Bakhtiari Arash, Bao Jianmin, Behl Harkirat, others* . Phi-3 technical report: A highly capable language model locally on your phone // arXiv preprint arXiv:2404.14219. 2024.
- Anthropic Sonnet*. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet // URL <https://api.semanticscholar.org/CorpusID.2024.273639283>.
- Antol Stanislaw, Agrawal Aishwarya, Lu Jiasen, Mitchell Margaret, Batra Dhruv, Zitnick C Lawrence, Parikh Devi*. Vqa: Visual question answering // Proceedings of the IEEE International Conference on Computer Vision. 2015. 2425–2433.
- Bai Jinze, Bai Shuai, Chu Yunfei, Cui Zeyu, Dang Kai, Deng Xiaodong, Fan Yang, Ge Wenbin, Han Yu, Huang Fei, Hui Binyuan, Ji Luo, Li Mei, Lin Junyang, Lin Runji, Liu Dayiheng, Liu Gao, Lu Chengqiang, Lu Keming, Ma Jianxin, Men Rui, Ren Xingzhang, Ren Xuancheng, Tan Chuanqi, Tan Sinan, Tu Jianhong, Wang Peng, Wang Shijie, Wang Wei, Wu Shengguang, Xu Benfeng, Xu Jin, Yang An, Yang Hao, Yang Jian, Yang Shusheng, Yao Yang, Yu Bowen, Yuan Hongyi, Yuan Zheng, Zhang Jianwei, Zhang Xingxuan, Zhang Yichang, Zhang Zhenru, Zhou Chang, Zhou Jingren, Zhou Xiaohuan, Zhu Tianhang*. Qwen Technical Report // arXiv preprint arXiv:2309.16609. 2023.
- Beijbom Oscar, Edmunds Peter J, Kline David I, Mitchell B Greg, Kriegman David*. Automated annotation of coral reef survey images // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012. 1170–1177.
- Bewley Michael, Friedman Ariell, Ferrari Renata, Hill Nicole, Hovey Renae, Barrett Neville, Marzinelli Ezequiel M, Pizarro Oscar, Figueira Will, Meyer Lisa, others* . Australian sea-floor survey data, with images and expert annotations // Scientific Data. 2015. 2, 1. 1–13.
- Bitterwolf Stephan A, Reguero Borja G, Storlazzi Curt D, Beck Michael W*. Shifting sands: The influence of coral reefs on shoreline erosion from short-term storm protection to long-term disequilibrium // Nature-Based Solutions. 2024. 6. 100174.
- Chen Zhe, Wu Jiannan, Wang Wenhai, Su Weijie, Chen Guo, Xing Sen, Zhong Muyan, Zhang Qinglong, Zhu Xizhou, Lu Lewei, others* . Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2024. 24185–24198.
- Fisher Rebecca, O’Leary Rebecca A, Low-Choy Samantha, Mengersen Kerrie, Knowlton Nancy, Brainard Russell E, Caley M Julian*. Species richness on coral reefs and the pursuit of convergent global estimates // Current Biology. 2015. 25, 4. 500–505.
- González-Rivero Manuel, Rodriguez-Ramirez Alberto, Beijbom Oscar, Dalton Peter, Kennedy Emma V, Neal Benjamin P, Vercelloni Julie, Bongaerts Pim, Ganase Anjani, Bryant Dominic EP, others* . Seaview Survey Photo-quadrat and Image Classification Dataset. 2019.
- Han Hongyong, Wang Wei, Zhang Gaowei, Li Mingjie, Wang Yi*. Enhancing Vision-Language Models with Morphological and Taxonomic Knowledge: Towards Coral Recognition for Ocean Health // Proceedings of the AAAI Conference on Artificial Intelligence. 39, 27. 2025. 28052–28060.
- He Xuehai, Zhang Yichen, Mou Luntian, Xing Eric, Xie Pengtao*. Pathvqa: 30000+ questions for medical visual question answering // arXiv preprint arXiv:2003.10286. 2020.
- Hoegh-Guldberg Ove, Jacob Daniela, Taylor M, Guillén Bolaños Tania, Bindi Marco, Brown Sally, Camilloni Ines Angela, Diedhiou Arona, Djalante Riyanti, Ebi Kristie, others* . The human imperative of stabilizing global climate change at 1.5 C // Science. 2019. 365, 6459. eaaw6974.
- Hurst Aaron, Lerer Adam, Goucher Adam P, Perelman Adam, Ramesh Aditya, Clark Aidan, Ostrow AJ, Welihinda Akila, Hayes Alan, Radford Alec, others* . Gpt-4o system card // arXiv preprint arXiv:2410.21276. 2024.
- Ilharco Gabriel, Wortsman Mitchell, Wightman Ross, Gordon Cade, Carlini Nicholas, Taori Rohan, Dave Achal, Shankar Vaishaal, Namkoong Hongseok, Miller John, Hajishirzi Hannaneh, Farhadi Ali, Schmidt Ludwig*. OpenCLIP. VII 2021.

- Kafle Kushal, Kanan Christopher.* Visual question answering: Datasets, algorithms, and future challenges // *Computer Vision and Image Understanding*. 2017. 163. 3–20.
- Knowlton Nancy, Brainard Russell E, Fisher Rebecca, Moews Megan, Plaisance Laetitia, Caley M Julian, others .* Coral reef biodiversity // *Life in the world's oceans: diversity distribution and abundance*. 2010. 65–74.
- Le Xue Manli Shu.* xGen-MM (formerly BLIP-3): A Family of Open Large Multimodal Models // *arXiv preprint*. 2024.
- Li Xiang, Ding Jian, Elhoseiny Mohamed.* Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding // *Advances in Neural Information Processing Systems*. 2024. 37. 3229–3242.
- Li Yanwei, Zhang Yuechen, Wang Chengyao, Zhong Zhisheng, Chen Yixin, Chu Ruihang, Liu Shaoteng, Jia Jiaya.* Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models // *arXiv:2403.18814*. 2023a.
- Li Yanwei, Zhang Yuechen, Wang Chengyao, Zhong Zhisheng, Chen Yixin, Chu Ruihang, Liu Shaoteng, Jia Jiaya.* Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models // *arXiv:2403.18814*. 2023b.
- Liu Haotian, Li Chunyuan, Wu Qingyang, Lee Yong Jae.* Visual Instruction Tuning. 2023.
- Lobry Sylvain, Marcos Diego, Murray Jesse, Tuia Devis.* RSVQA: Visual question answering for remote sensing data // *IEEE Transactions on Geoscience and Remote Sensing*. 2020. 58, 12. 8555–8566.
- Meyer Lisa, Hill Nicole, Walsh Peter, Barrett Neville.* Methods for the processing and scoring of AUV digital imagery from South Eastern Tasmania // *Institute for Marine and Antarctic Studies Internal Report*. 2011.
- Mumby Peter J, Broad Kenneth, Brumbaugh Daniel R, DAHLGREN P CRAIG, Harborne Alastair R, Hastings Alan, Holmes Katherine E, Kappel Carrie V, Micheli Fiorenza, Sanchirico James N.* Coral reef habitats as surrogates of species, ecological functions, and ecosystem services // *Conservation Biology*. 2008. 22, 4. 941–951.
- Rashid Ahmad Rafiuddin, Chennu Arjun.* A trillion coral reef colors: Deeply annotated underwater hyperspectral images for automated classification and habitat mapping // *Data*. 2020. 5, 1. 19.
- Shihavuddin ASM.* Coral reef dataset. 2017.
- Shihavuddin ASM, Gracias Nuno, Garcia Rafael, Gleason Arthur CR, Gintert Brooke.* Image-based coral reef classification and thematic mapping // *Remote Sensing*. 2013. 5, 4. 1809–1841.
- Steffen Will, Crutzen Paul J, McNeill John R, others .* The Anthropocene: are humans now overwhelming the great forces of nature // *Ambio-Journal of Human Environment Research and Management*. 2007. 36, 8. 614–621.
- Wu Qi, Teney Damien, Wang Peng, Shen Chunhua, Dick Anthony, Van Den Hengel Anton.* Visual question answering: A survey of methods and datasets // *Computer Vision and Image Understanding*. 2017. 163. 21–40.
- Yang Miao, Sowmya Arcot.* An underwater color image quality evaluation metric // *IEEE Transactions on Image Processing*. 2015. 24, 12. 6062–6071.
- Zhai Xiaohua, Mustafa Basil, Kolesnikov Alexander, Beyer Lucas.* Sigmoid loss for language image pre-training // *Proceedings of the IEEE International Conference on Computer Vision*. 2023. 11975–11986.
- Zheng Lianmin, Chiang Wei-Lin, Sheng Ying, Zhuang Siyuan, Wu Zhanghao, Zhuang Yonghao, Lin Zi, Li Zhuohan, Li Dacheng, Xing Eric, others .* Judging llm-as-a-judge with mt-bench and chatbot arena // *Advances in Neural Information Processing Systems*. 2023. 36. 46595–46623.
- Zheng Ziqiang, Liang Haixin, Hua Binh-Son, Wong Yue Him, Ang Put, Chui Apple Pui Yi, Yeung Sai-Kit.* CoralSCOP: segment any coral image on this planet // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2024. 28170–28180.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: In this work, we introduce CoralVQA, the first large-scale VQA dataset for coral reef analysis. It contains 12,805 real-world coral images from 67 coral genera collected from 3 oceans, along with 277,653 question-answer pairs that comprehensively assess ecological and health-related conditions. To construct this dataset, we develop a semi-automatic data construction pipeline in collaboration with marine biologists to ensure both scalability and professional-grade data quality. CoralVQA presents novel challenges and provides a comprehensive benchmark for studying vision-language reasoning in the context of coral reef images. By evaluating several state-of-the-art LVLMs, we reveal key limitations and opportunities. These insights form a foundation for future LVLM development, with a particular emphasis on supporting coral conservation efforts.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of the work are addressed in the appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Detailed experimental settings are provided in both the Implementation Details section and the Appendix. Our code and dataset are publicly available at <https://huggingface.co/datasets/CoralReefData/CoralVQA/tree/main>.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and dataset are publicly available at <https://huggingface.co/datasets/CoralReefData/CoralVQA/tree/main>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Detailed experimental settings are provided in both the Implementation Details section and the Appendix. Our code and dataset are publicly available at <https://huggingface.co/datasets/CoralReefData/CoralVQA/tree/main>.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: For different tasks, we define distinct evaluation metrics and conduct corresponding experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide relevant information on the computer resources (type of computer workers, memory, time of execution) in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The paper conforms to the NeurIPS Code of Ethics. We have addressed data usage, possible societal impact, and limitations appropriately.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential benefits of the proposed datasets, such as improving ecological monitoring and coral conservation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All relevant data and code are cited in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: The new assets introduced in the paper are well documented, and the documentation is provided alongside the assets (e.g., with the code or dataset).

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#) .

Justification: N/A.

Guidelines: N/A.

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#) .

Justification: N/A.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: N/A.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.