CORALAGENT: FROM VISUAL RECOGNITION TO CONSTRAINED AGENTIC ECOLOGICAL ASSESSMENT FOR CORAL REEF CONSERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable coral reef ecological analysis requires more than accurate visual recognition; it demands strict adherence to sampling protocols, taxonomic standards, and precise measurement definitions. However, existing vision-language models and general-purpose agents lack explicit mechanisms to enforce these domain-specific requirements, a challenge compounded by the scarcity of reliable coral-specific models. To bridge this gap, we introduce CoralAgent, a multi-agent framework that translates ecological requirements into explicit rules for evidence collection, computation, and validation. Its core component, Coral Structured TAsk Representation (C-STAR), formalizes measurement objectives, evidence requirements, validation criteria, recovery policies, and admissible conclusions. Guided by C-STAR, CoralAgent coordinates reusable analytical primitives, validates intermediate results, and switches to workflows adapted from conventional coral surveys when coral-specific models are unavailable or their outputs fail validation. To evaluate quantitative ecological analysis beyond recognition and segmentation, we introduce Coral-Bench, comprising 1,986 question–answer pairs and 2,535 images across 23 tasks. These tasks cover coral recognition, visual knowledge, cover estimation, community composition and structure analysis, and ecological index computation, including cross-site and cross-year analysis. Experiments with seven multimodal baselines show that CoralAgent outperforms the GPT-5.4 baseline on 22 of 23 tasks and achieves the best or tied-best performance on 20. Compared with GPT-5.4, CoralAgent improves task accuracy for single-image coral cover estimation from 20.00% to 72.59% and for Gini–Simpson diversity calculation from 19.18% to 61.64%.

1 INTRODUCTION

While providing essential habitats for marine organisms to support marine biodiversity (Roberts et al., 2002; Graham & Nash, 2013), coral reefs face increasing environmental pressures from ocean warming (Hughes et al., 2017), acidification (Hoegh-Guldberg et al., 2007), pollution (Fabricius, 2005), and overexploitation (Bellwood et al., 2004). Effective conservation and restoration of coral reefs require accurate and sustained monitoring. An essential monitoring approach is image-based monitoring, which uses underwater imagery to estimate coral cover, characterize community composition, and quantify biodiversity. Estimating these ecological indicators requires more than accurate visual recognition, i.e, the analysis must follow consistent sampling procedures, apply uniform taxonomic standards, and use clearly defined measurement criteria. Conventional monitoring often relies on experts to manually interpret underwater images, making large-scale, long-term monitoring labor-intensive. Automated image-based monitoring must therefore combine visual analysis, ecological computation, and validation to check that the analysis follows ecological survey requirements and that the reported conclusions are supported by evidence.

Recent advances in marine visual models and vision–language models (VLMs) provide a foundation for automated underwater image analysis (Zheng et al., 2024a; Alawode et al., 2026). These developments support coral recognition (Han et al., 2025), segmentation (Zheng et al., 2024b), and visual question answering (Han et al., 2026). However, when VLMs are used for direct image-to-answer prediction, their outputs do not necessarily specify the procedures used to derive ecological

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

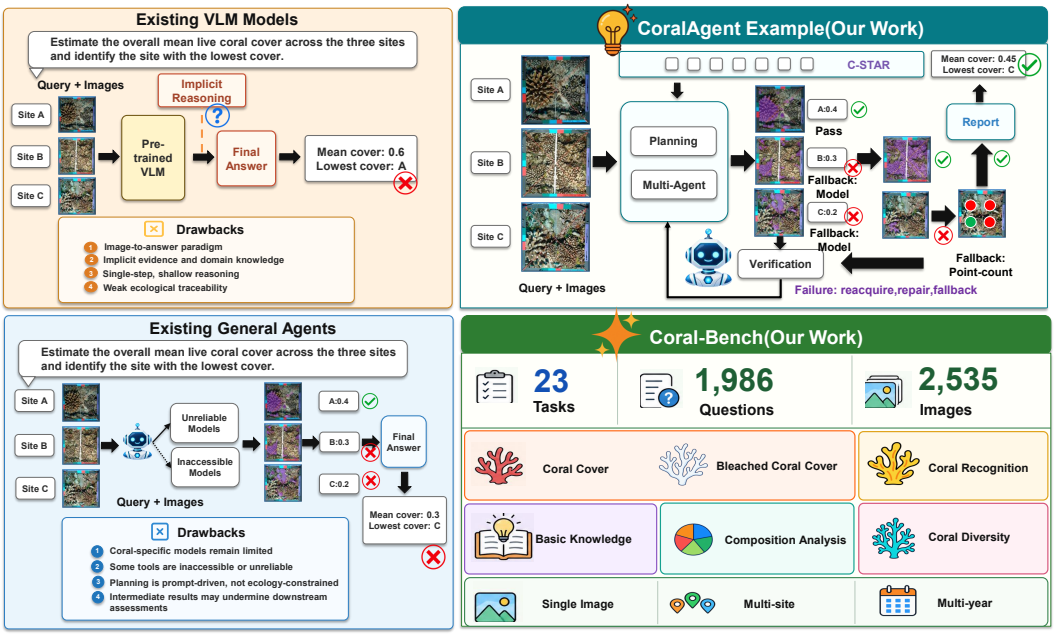

Figure 1: **Overview of CoralAgent and Coral-Bench. Left:** Existing VLMs typically follow a direct image-to-answer paradigm, while general-purpose agents provide tool-augmented reasoning but not enforce task-specific ecological requirements. **Right:** CoralAgent addresses these limitations with C-STAR, an evidence-constrained task representation that guides planning, multi-agent execution, validation, recovery, and reporting while incorporating expert knowledge as explicit execution constraints. The framework supports repair and fallback when intermediate evidence is insufficient and naturally extends to multi-site and multi-year ecological analysis. Coral-Bench evaluates these capabilities with 1,986 question–answer pairs and 2,535 images across 23 tasks.

measurements or provide the visual evidence supporting those measurements. Meanwhile, general-purpose agents extend visual analysis into multi-step workflows by decomposing user requests and coordinating visual models and statistical tools. However, two challenges remain in coral reef ecological analysis. First, agents may not consistently follow sampling protocols, taxonomic standards, and measurement definitions, even when these requirements are provided through prompts or retrieved knowledge. Second, coral-specific models may be unavailable or produce outputs unsuitable for the requested ecological tasks. An agent framework for coral reef analysis must therefore verify adherence to ecological standards and provide alternative analysis methods when specialized models are unavailable or cannot support the requested tasks.

For example, a typical monitoring task may be estimating the overall mean live coral cover across the three sites and identifying the site with the lowest cover (Figure 1). This task requires consistent sampling procedures, taxonomic standards, and cover definitions across three sites, yet a general-purpose agent may not reliably enforce these requirements. Moreover, coral segmentation models may produce incomplete masks for some images, leaving insufficient visual evidence to support the comparison. The CoralAgent example illustrates how these two challenges interact within a single workflow. Ecological requirements guide planning and provide criteria for verifying intermediate results. Segmentation outputs contribute to the final report only after passing validation. When a segmentation model is unavailable or its outputs fail validation, CoralAgent invokes an alternative workflow, such as image-based point counting. The alternative workflow preserves the survey scope and measurement definition. Point-count results are also validated and, where necessary, repaired and revalidated before acceptance. Across these steps, the agent must preserve the survey scope and cover definition while adapting its analysis methods, so that the final comparison uses consistent measurements and verified evidence.

We introduce CoralAgent, an evidence-constrained multi-agent framework inspired by the analytical workflows of marine ecologists. At its core is Coral Structured TAsk Representation (C-STAR), which links survey objectives and measurement constraints to evidence requirements, derivation rules, validation criteria, recovery policies, and admissible conclusions. This representation guides

visual evidence acquisition, ecological computation, and result validation. When suitable models are unavailable or their outputs fail validation, CoralAgent can fall back on analytical workflows designed by marine scientists based on conventional coral surveys. It organizes the corresponding sampling, visual identification, and statistical analysis procedures into executable steps governed by C-STAR. CoralAgent composes reusable analytical primitives and incorporates expert knowledge as checkable execution constraints rather than solely as prompt context. A systematic evaluation of CoralAgent therefore requires a benchmark covering both visual understanding and quantitative ecological analysis. Existing coral benchmarks primarily focus on recognition, segmentation, and visual question answering, offering limited coverage of quantitative ecological tasks such as coral cover estimation, community composition analysis, and diversity assessment. To address this evaluation gap, we introduce Coral-Bench, a unified benchmark for evaluating multimodal models and agents on coral reef ecological analysis. Coral-Bench comprises 1,986 question–answer pairs and 2,535 images across 23 tasks, organized into five groups: coral recognition, visual knowledge question answering, coral cover estimation, community composition and structure analysis, and diversity calculation and evenness assessment. These tasks span single-image analysis, cross-site comparisons, and changes across survey years. The benchmark combines educational resources with coral survey imagery, with questions and reference answers established by marine scientists. Quantitative ecological targets are derived from standardized point-count annotations, and benchmark quality is assessed through manual review, cross-annotator consistency checks, and expert auditing. Together, these components support systematic evaluation of visual understanding and quantitative ecological analysis.

Our main contributions are summarized as follows:

- We propose **CoralAgent**, a novel agentic framework dedicated to coral reef ecological analysis. Central to this architecture is the C-STAR representation, which formalizes expert knowledge into verifiable constraints for ecological analysis and validation. This mechanism enables the system to dynamically fall back to survey-based workflows when specialized models prove unavailable or unreliable.
- We introduce **Coral-Bench**, a unified benchmark designed to rigorously evaluate multimodal models and agents in coral reef ecology. The dataset comprises 1,986 question-answer pairs and 2,535 images across 23 tasks, encompassing both visual understanding and complex quantitative assessments in single-image, cross-site, and cross-year contexts.
- We systematically evaluate CoralAgent against seven multimodal baselines on Coral-Bench. CoralAgent outperforms the GPT-5.4 baseline on 22 of the 23 tasks, demonstrating substantial gains on critical tasks such as coral cover estimation and ecological index computation.

2 Related Work

2.1 Vision-language models and Tool-Augmented Agents

Vision-language models (VLMs) have achieved notable success across a range of general multimodal tasks, including image captioning (Li et al., 2022; 2023), visual grounding and visual question answering (Liu et al., 2023; Dai et al., 2023). Building on these advances, recent studies have applied VLMs to marine image analysis, including coral recognition and underwater scene understanding. CORAL-Adapter (Han et al., 2025) incorporates morphological and taxonomic knowledge into CLIP for fine-grained coral recognition, while AquaticCLIP (Alawode et al., 2026) learns aquatic image-text representations for underwater scene analysis. CoralVQA (Han et al., 2026) introduces a large-scale visual question answering dataset and evaluates existing VLMs, including Qwen2.5-VL and InternVL2.5, on questions covering coral visual attributes, ecological characteristics, and health conditions. Beyond interpreting visual images, the reef assessment tasks considered in our work require coordinating visual reasoning, quantitative computations, and validation across multiple analytical stages. Tool-augmented agents further extend the capabilities of foundation models by integrating specialized models and computational tools into task-oriented workflows. In remote sensing, RS-Agent (Xu et al., 2026) combines domain knowledge with image-processing tools to support tasks such as scene classification, object counting, and visual question answering. Earth-Agent (Feng et al., 2026) further integrates perception, index computation, parameter retrieval, and statistical analysis tools to support multi-step, cross-modal Earth observation tasks. Building on these

tool-augmented approaches, CoralAgent focuses on incorporating coral survey requirements into task planning and execution. Its evidence-constrained task representation links survey objectives to evidence requirements, analytical procedures, and validation criteria, guiding tool use, result verification, and additional evidence collection.

2.2 Related Datasets

Coral datasets provide a foundation for advancing automated reef monitoring and ecological assessment. A series of coral datasets and benchmarks have been introduced over the years, including TasCPC (Meyer et al., 2011), MLC (Beijbom et al., 2012), RSMAS (Shihavuddin et al., 2013), EILAT (Shihavuddin et al., 2013), Benthoz15 (Bewley et al., 2015), ATCRC (Rashid & Chennu, 2020), CoralMask (Zheng et al., 2024b), HKCoral (Zheng et al., 2024c), Coralscapes (Sauder et al., 2025). However, most existing datasets and benchmarks focus on single task evaluation, primarily coral recognition or coral segmentation. For example, dataset TasCPC, EILAT, RSMAS, Benthoz15, and ATCRC mainly support coral recognition, whereas CoralMask, HKCoral, and Coralscapes focus on coral segmentation. Moreover, these benchmarks are largely restricted to a single input modality, typically images. More recent efforts have begun to extend coral reef analysis toward vision–language understanding. HSCR16K (Han et al., 2025) introduces 16,659 stony coral images with hierarchical taxonomic labels and morphology-specific textual descriptions, primarily supporting fine-grained coral recognition. CoralVQA (Han et al., 2026) further extends coral understanding to visual question answering, providing 277,653 question–answer pairs across 16 question dimensions. Nevertheless, these vision–language benchmarks primarily emphasize recognition and image-based question answering, with limited coverage of quantitative ecological tasks such as coral community composition and structure analysis, and coral diversity calculation and evenness assessment. In contrast, we build a coral benchmark spanning 23 coral reef monitoring tasks, including coral cover estimation, coral community composition and structure analysis, and coral diversity calculation and evenness assessment. By extending task coverage beyond recognition and visual question answering to temporal analysis, domain-specific knowledge, and quantitative ecological assessment, our dataset enables a broader evaluation of multimodal models and agents for real-world coral reef monitoring and assessment.

3 CoralAgent Framework

CoralAgent is the first agent framework specifically designed for coral reef ecological analysis. It decouples ecological task requirements from the models and analytical procedures used to satisfy them. We encode these requirements in Coral Structured TAsk Representation (C-STAR) as explicit constraints on planning, execution, validation, recovery, and reporting. This representation allows execution strategies to change while preserving the survey scope and measurement definition.

3.1 Overall Framework

Given a user request q and survey images $\mathcal{X}$, CoralAgent proceeds through three stages: task planning, multi-agent execution, and verification and recovery (Figure 2). The Router first composes reusable analytical primitives into a base C-STAR specification. Task-relevant expert rules are then compiled into its fields as explicit execution constraints. The resulting specification guides the construction of an executable workflow using registered tools. During execution, the Vision, Statistics, and Knowledge agents acquire visual evidence, perform sampling and ecological computations, and supply domain rules, respectively. Agents coordinate through a shared state that records evidence, provenance, intermediate measurements, and validation outcomes.

The Validation agent applies expert rules designed by marine scientists to check intermediate evidence and ecological measurements. These rules are tailored to the specific requirements of each ecological task and analytical method. For coral cover, it compares the segmentation-based estimate with repeated image-based point-count estimates using absolute differences. If these comparisons fail the task-specific acceptance criteria, or no suitable model is available, the Router invokes the survey-based fallback specified by C-STAR. Other validation failures may trigger evidence reacquisition or repair. The Report agent generates conclusions only after the applicable validation checks pass; unresolved issues are flagged for review.

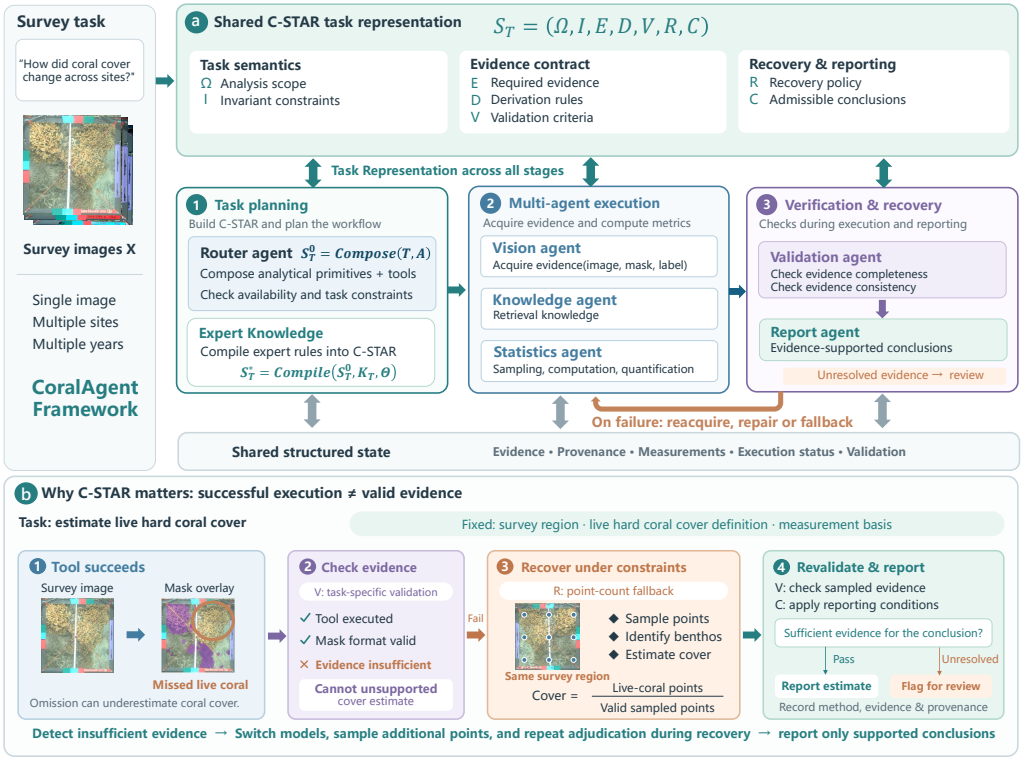

Figure 2: **CoralAgent framework.** **(a)** C-STAR links ecological task requirements to planning, multi-agent execution, and verification and recovery. Expert rules constrain the task specification, while a shared state records evidence, provenance, measurements, and validation outcomes. **(b)** A successful tool call may still provide insufficient measurement evidence. CoralAgent can recover through point counting within the same survey region and under the same live hard-coral definition; reporting requires revalidation of the recovered evidence.

3.2 Evidence-Constrained Task Representation

A central requirement in coral reef ecological analysis is that the *meaning of an ecological task should remain fixed even when its execution strategy changes*. We therefore separate task-level ecological requirements from the particular models and tools used to satisfy them. We introduce **Coral Structured TAsk Representation (C-STAR)**, an evidence-constrained task representation that makes these requirements explicit throughout planning, execution, validation, recovery, and reporting. Rather than leaving ecological requirements implicit in prompts or tool descriptions, C-STAR specifies what is being analyzed, what evidence is required, how the target result is derived from that evidence, how the result should be validated, and which conclusions are supported by the resulting evidence state. This representation allows CoralAgent to change models or analytical procedures during execution while preserving the task-specific survey and measurement requirements.

Task-level specification. For an ecological analysis task T, we define its C-STAR representation as

$$S_T = (\Omega, I, E, D, V, R, C), \tag{1}$$

with three groups of fields. *Task semantics* (Ω, I) define what the analysis measures: Ω specifies the objective, survey scope, observation units, and sampling requirements, while I records invariant constraints such as ecological definitions, the measurement basis, and taxonomic standards. *Evidence and derivation* (E, D, V) connect observations to valid results: E specifies required evidence and its sources, D defines how measurements or statistics are derived, and V specifies checks on evidence completeness, derivation consistency, and result validity. *Recovery and reporting* (R, C) govern responses to the evidence state: R specifies permitted recovery strategies, including additional acquisition, correction, adjudication, alternative analysis, and human review; C identifies the conclusions supported by the current evidence and validation status. These fields jointly constrain

a workflow rather than define seven sequential steps. In addition, successful tool execution alone does not imply that a conclusion can be reported; the available evidence must also satisfy the task-specific validation and reporting conditions. We provide an illustrative example in Appendix A.2.1 showing how C-STAR represents the task of live coral cover estimation.

Reusable analytical primitives. Complex ecological tasks require multiple operations, including evidence acquisition, sampling, measurement, aggregation, and validation. We therefore design a registry of 130 reusable analytical primitives to support all 23 tasks in Coral-Bench. Each analytical primitive implements a local analytical operation using the same seven-field schema, specifying its evidence dependencies, derivation, validation, recovery, and permitted downstream use. A locally validated output need not establish the final task-level conclusion. For example, PIXELAREAMEASUREMENT produces traceable pixel counts within a predefined valid survey region; reporting coral cover additionally requires cover-ratio calculation and task-level validation. The detailed local schema and primitive registry are described in Appendix A.2.2, and the complete example is given in Appendix A.2.3.

Compositional task expressions. A single primitive defines only a local operation. A complete task representation must additionally express how local operations depend on one another, when complementary evidence is required, when validation failures activate alternative procedures, and how unit-level analyses are extended across images, sites, or survey years. We therefore compose analytical primitives into task expressions using a grammar. The full grammar and operator semantics are provided in Appendix A.2.4.

3.3 COMPILING EXPERT KNOWLEDGE INTO C-STAR

Providing domain knowledge to an agent as textual context does not ensure that the knowledge is respected during execution. An agent may retrieve a correct ecological rule yet still acquire inappropriate evidence, apply an inconsistent measurement procedure, or invoke a recovery strategy that changes the quantity being estimated. CoralAgent therefore treats expert knowledge as operational constraints rather than auxiliary prompt context. We construct a library of 153 structured expert rules for coral reef ecological analysis. Given a task-specific C-STAR representation, CoralAgent identifies the expert rules applicable to the current ecological analysis and compiles them into the corresponding C-STAR fields. The resulting constraints govern evidence acquisition, result derivation, validation, recovery, and conclusion generation throughout execution. In this way, domain knowledge is not merely consulted during planning or response generation; it becomes part of the task specification against which intermediate evidence and analytical decisions can be checked.

Let $\mathcal{A}$ denote the reusable library of analytical primitives. For a task T, CoralAgent first constructs a base C-STAR representation by composing the primitives required for the requested analysis:

$$S_T^0 = \mathrm{Compose}(T, \mathcal{A}). \tag{2}$$

Let $\mathcal{K}$ denote the expert-rule library and $\mathcal{K}_T \subseteq \mathcal{K}$ the subset of rules applicable to task T. CoralAgent then compiles these task-relevant rules into the base representation:

$$S_T^* = \mathrm{Compile}\left(S_T^0, \mathcal{K}_T; \Theta\right), \tag{3}$$

where S_T^* is the resulting knowledge-augmented C-STAR representation and Θ contains calibration policies such as sampling limits, confidence settings, and retry limits. During compilation, each applicable expert rule is mapped to the C-STAR fields and analytical primitives that it constrains. Compatible constraints are incorporated into the task specification, while conflicts, omissions, and unsatisfiable conditions are detected as compilation diagnostics. The resulting representation therefore specifies not only which analytical operations should be executed, but also the scientific conditions under which their inputs, intermediate outputs, and downstream conclusions are considered valid.

For live coral cover estimation, these constraints preserve the survey region and target definition when switching from segmentation to point counting. Appendix A.2.1 illustrates how C-STAR governs evidence acquisition, validation, and recovery for this task.

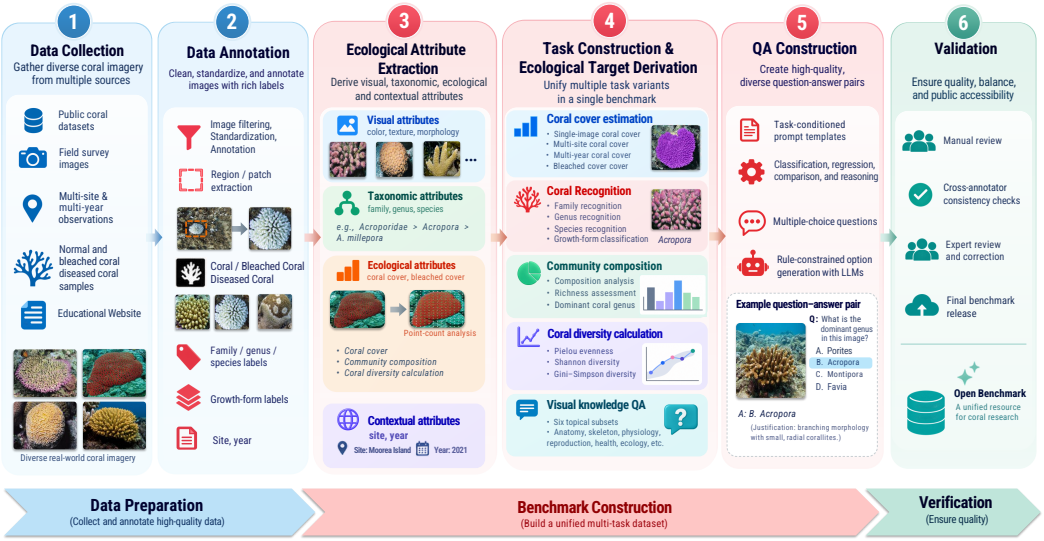

Figure 3: **Coral-Bench construction pipeline.** We construct Coral-Bench through six stages: data collection, data annotation, ecological attribute extraction, task construction and ecological target derivation, QA construction, and validation.

4 BENCHMARK

4.1 OVERVIEW

We introduce CoralAgent Benchmark (Coral-Bench), the first comprehensive benchmark for evaluating model and agent capabilities in coral reef ecological monitoring and visual analysis. Coral-Bench comprises 1,986 question-answer pairs and 2,535 images, covering 23 tasks across five task groups: coral cover estimation, coral recognition, visual knowledge question answering, coral diversity calculation and evenness assessment, and coral community composition and structure analysis. Representative examples from the five task groups are provided in Figures 5 – 9 in the appendix. By bringing together domain knowledge, visual perception, and quantitative ecological analysis, Coral-Bench extends evaluation beyond coral recognition and segmentation to ecological indicator computation and community-level assessment.

4.2 PIPELINE

To build Coral-Bench, we design the benchmark construction and verification pipeline. As shown in Figure 3, our pipeline comprises six stages: data collection, data annotation, ecological attribute extraction, task construction and ecological target derivation, QA construction, and validation. We combine authoritative educational resources with public coral survey imagery and standardize the patch-level annotation and point-based ecological annotation. For quantitative ecological tasks, sampled-point records capture live-coral presence, bleaching status, and genus identity. These records are subsequently aggregated to derive reference values for coral cover, genus composition, genus richness, dominance, diversity, and evenness. Exact annotation procedures, ecological definitions, and target derivations are provided in Appendix A.1.3 and Appendix A.1.4.

4.3 QA CONSTRUCTION AND QUALITY CONTROL

To improve the quality and reliability of Coral-Bench, three marine scientists formulate the questions and establish reference answers from educational sources, standardized annotations, or derived ecological targets; GPT-5.4 is used only to generate candidate multiple-choice distractors. The resulting benchmark undergoes manual annotation review, inter-annotator consistency checks, and expert auditing by a team comprising three marine scientists and six students majoring in marine science. Details of the data sources, annotation procedures, and validation protocols are provided in Appendix A.1.3.

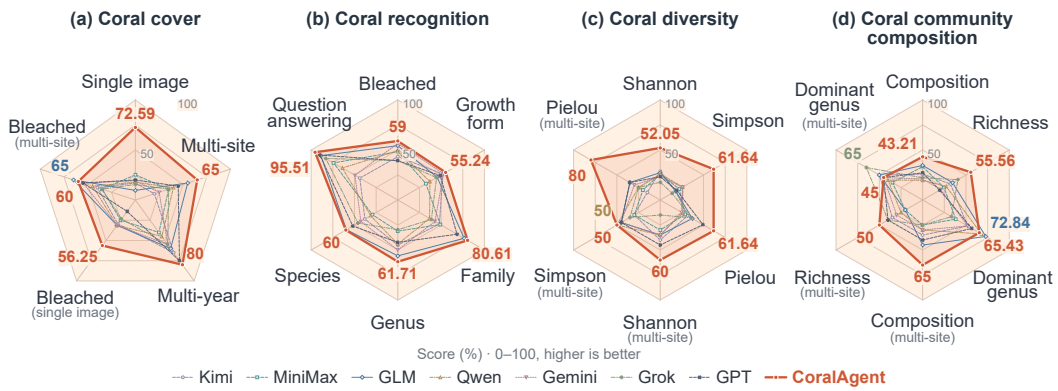

Figure 4: **Comparison of CoralAgent and seven VLM baselines across 23 tasks in Coral-Bench.** The figure covers (a) coral cover estimation, (b) recognition and visual knowledge question answering, (c) diversity calculation and evenness assessment, and (d) community composition and structure analysis. Orange curves show CoralAgent's performance, while curves in other colors show the performance of the VLM baselines. Detailed results are provided in Table 1 in the appendix.

5 Experiments

5.1 Experimental Setup

We evaluate seven MLLMs, including four closed-source models—GPT-5.4 (GPT), Gemini 3.1 Flash (Gemini), Grok 4.20 (Grok), and Qwen3.6-Plus (Qwen)—and three open-source models—Kimi K2.6 (Kimi), MiniMax-M3 (MiniMax), and GLM-5.3-Flash (GLM). CoralAgent uses GPT-5.4, accessed through the API, as its backbone model for task processing. We evaluate all methods using accuracy, defined as the percentage of correctly answered multiple-choice questions within each task. We evaluate 23 tasks across the five Coral-Bench groups introduced in Section 4, covering visual recognition, biological understanding, and quantitative ecological analysis. Detailed task descriptions and statistics are provided in Appendix A.1, with the full task list in Table 2 and representative examples in Figures 5– 9. Additional implementation details are provided in Appendix A.3.1. Our code and dataset are publicly available at `https://huggingface.co/datasets/CoralAgent/CoralAgent`.

5.2 Results

Table 1 reports the performance of CoralAgent and seven multimodal baselines across five task groups. CoralAgent outperforms GPT on 22 of the 23 tasks and matches its performance on the remaining task, multi-site dominant genus recognition, where both achieve 45.00%. On coral cover estimation, CoralAgent improves performance from 20.00% with GPT to 72.59%, yielding a gain of 52.59 percentage points. Substantial gains are also observed in Gini–Simpson diversity calculation, which improves from 19.18% to 61.64% (+42.46 percentage points), and multi-site Pielou's evenness assessment, which improves from 35.00% to 80.00% (+45.00 percentage points). Overall, these results highlight the benefits of augmenting GPT with CoralAgent for coral reef analysis, particularly on tasks requiring quantitative ecological assessment.

Coral Cover Estimation. CoralAgent achieves the best performance on four cover estimation tasks, with particularly large improvements in single image coral cover assessment. CoralAgent achieves 72.59% on coral cover estimation, exceeding the strongest baseline, MiniMax (25.19%), by 47.40 percentage points. Improvements also extend to spatial and temporal comparisons. CoralAgent reaches 65.00% on multi-site coral cover estimation and 80.00% on multi-year coral cover estimation, surpassing GLM (55.00%) and GPT (75.00%) by 10.00 and 5.00 percentage points, respectively.

Coral Recognition and Visual Knowledge QA. CoralAgent obtains the highest performance on all six recognition and visual knowledge tasks. For family-level, genus-level, and species-level

recognition, CoralAgent achieves 80.61%, 61.71%, and 60.00%, respectively, outperforming GPT by 12.24, 19.37, and 8.00 percentage points. CoralAgent also leads on bleached coral recognition, scoring 59.00 % against GPT's 39.33%, a gain of 19.67 percentage points. On visual knowledge QA, CoralAgent achieves 95.51%, exceeding GPT's 89.89% by 5.62 percentage points. Overall, CoralAgent delivers substantial gains over the GPT baseline across coral recognition tasks, indicating that the effectiveness of our framework extends beyond quantitative ecological assessment.

Coral Diversity Calculation and Evenness Assessment. In the single-image setting, it obtains 52.05% on Shannon diversity index calculation, 61.64% on Gini–Simpson diversity index calculation, and 61.64% on Pielou's evenness index assessment. These results exceed the strongest respective baselines—Kimi (28.77%), Qwen and Gemini (both 26.03%), and GPT (49.32%)—by 23.28, 35.61, and 12.32 percentage points. In the multi-site setting, CoralAgent achieves 60.00% on Shannon diversity index calculation, outperforming GPT-5.4's 45.00% by 15.00 percentage points. It also reaches 80.00% on Pielou's evenness index calculation, exceeding the best baseline score of 35.00%, shared by GLM and GPT, by 45.00 percentage points. Overall, CoralAgent consistently outperforms GPT across all six diversity and evenness index calculation tasks, highlighting its strength in quantitative coral diversity and evenness assessment across the evaluated indicators and survey settings.

Coral Community Composition and Structure Analysis. CoralAgent achieves the highest score on four coral community composition and structure analysis tasks, with improvements concentrated in composition analysis and richness assessment. On the single-image tasks, it scores 43.21% for coral composition analysis and 55.56% for coral richness assessment, exceeding GLM (34.57%) and Grok (40.74%) by 8.64 and 14.82 percentage points, respectively. Its advantage also extends to cross-site comparisons: CoralAgent achieves 65.00% on multi-site coral composition analysis, outperforming GLM's 45.00% by 20.00 percentage points, and 50.00% on multi-site richness assessment, exceeding the best baseline score of 40.00%, shared by Grok and GPT, by 10.00 percentage points.

5.3 Ablation Studies

Effect of C-STAR-guided execution. To assess the benefit of C-STAR-guided execution, we compare GPT-Agent and CoralAgent using the same backbone and tool set. CoralAgent achieves the highest accuracy on all eight evaluated ecological analysis tasks, outperforming GPT-Agent by 17.28–44.44 percentage points. GPT-Agent does not consistently outperform GPT without tools, indicating that tool access alone does not ensure better ecological assessment. These results support the value of C-STAR-guided execution in coordinating evidence acquisition, computation, validation, and recovery. Detailed experimental configurations, results, and analysis are provided in Appendix A.3.2.

Effects of validation and recovery in C-STAR. We further evaluate the effectiveness of C-STAR's validation and recovery mechanisms on the Gini–Simpson diversity index calculation task. With the same GPT-5.4 backbone and available tools, CoralAgent achieves 61.64% accuracy, compared with 27.40% for GPT-Agent. Removing validation or recovery reduces accuracy by 12.32 and 23.28 percentage points, respectively. These results support the importance of integrating validation and recovery into C-STAR to improve ecological assessment accuracy beyond tool and expert-rule availability. Detailed experimental configurations, results, and analysis are provided in Appendix A.3.2.

6 Conclusion

We presented CoralAgent, a multi-agent framework that uses C-STAR to encode ecological requirements and expert knowledge as checkable constraints on evidence acquisition, computation, and validation. It supports fallback to workflows adapted from conventional coral surveys while preserving measurement definitions when specialized models are unavailable or their outputs fail validation. We also introduced Coral-Bench, covering 23 tasks spanning visual understanding and quantitative ecological analysis in single-image, cross-site, and cross-year settings. CoralAgent outperformed the GPT-5.4 baseline on 22 of 23 tasks and achieved the best or tied-best performance on 20 against seven multimodal baselines. These findings support C-STAR-guided execution for quantitative coral reef assessment and suggest executable measurement constraints as a promising design principle for scientific agents.

AI USE STATEMENT

We used generative AI tools as part of the experimental workflow and for limited assistance in manuscript preparation. GPT-5.4 was accessed through its API as the backbone model of CoralAgent for task processing, including workflow planning and visual reasoning. During Coral-Bench construction, GPT-5.4 was used only to generate candidate multiple-choice distractors under predefined task constraints. Benchmark questions and reference answers were established by marine scientists from curated sources, standardized annotations, or derived ecological targets, and were not generated by the model. AI-generated answer options were manually reviewed, cross-checked, and subject to expert auditing before inclusion in the benchmark. Generative AI tools were also used for language editing, wording refinement, and organization of parts of the manuscript. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is guided by the ICLR Code of Ethics. The study involves computational analysis of existing coral imagery and educational materials, with no human-subject research or experiments involving live animals. Coral-Bench draws on publicly available resources, including Moorea Labeled Corals, Flower Garden Banks National Marine Sanctuary survey imagery, the XL Catlin Seaview Survey Project, Corals of the World, and institutional educational resources. We acknowledge the original data providers and are committed to respecting applicable licenses and usage restrictions. Expert review and cross-annotator consistency checks are employed to reduce annotation errors. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the transparency and reproducibility of our work. The CoralAgent framework, including the C-STAR representation and the integration of structured expert knowledge, is described in Section 3.1 and Appendix A.2. The construction of Coral-Bench, including data sources, annotation procedures, quality control, and ecological target derivation, is detailed in Section 4 and Appendix A.1. The evaluated models are presented in Section 5.1, with implementation details provided in Appendix A.3.1. Our code and Coral-Bench are available through the repository linked in Section 5.1.

REFERENCES

Basit Alawode, Iyyakutti Iyappan Ganapathi, Sajid Javed, Mohammed Bennamoun, and Arif Mahmood. Aquaticclip: A vision-language foundation model and dataset for underwater scene analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Oscar Beijbom, Peter J Edmunds, David I Kline, B Greg Mitchell, and David Kriegman. Automated annotation of coral reef survey images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1170–1177. IEEE, 2012.

David R Bellwood, Terry P Hughes, Carl Folke, and Magnus Nyström. Confronting the coral reef crisis. *Nature*, 429(6994):827–833, 2004.

Michael Bewley, Ariell Friedman, Renata Ferrari, Nicole Hill, Renae Hovey, Neville Barrett, Ezequiel M Marzinelli, Oscar Pizarro, Will Figueira, Lisa Meyer, et al. Australian sea-floor survey data, with images and expert annotations. *Scientific Data*, 2(1):1–13, 2015.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Katharina E Fabricius. Effects of terrestrial runoff on the ecology of corals and coral reefs: review and synthesis. *Marine pollution bulletin*, 50(2):125–146, 2005.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, Lei Bai, Conghui He, et al. Earth-agent: Unlocking the full landscape of earth observation with agents. In *International Conference on Learning Representations*, volume 2026, pp. 55700–55764, 2026.

Nicholas AJ Graham and Kirsty L Nash. The importance of structural complexity in coral reef ecosystems. *Coral reefs*, 32(2):315–326, 2013.

Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li, and Yi Wang. Enhancing vision-language models with morphological and taxonomic knowledge: Towards coral recognition for ocean health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28052–28060, 2025.

Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li, and Yi Wang. Coralvqa: A large-scale visual question answering dataset for coral reef image understanding. *Advances in Neural Information Processing Systems*, 38, 2026.

Ove Hoegh-Guldberg, Peter J Mumby, Anthony J Hooten, Robert S Steneck, Paul Greenfield, Edgardo Gomez, C Drew Harvell, Peter F Sale, Alasdair J Edwards, Ken Caldeira, et al. Coral reefs under rapid climate change and ocean acidification. *science*, 318(5857):1737–1742, 2007.

Terry P Hughes, James T Kerry, Mariana Álvarez-Noriega, Jorge G Álvarez-Romero, Kristen D Anderson, Andrew H Baird, Russell C Babcock, Maria Beger, David R Bellwood, Ray Berkelmans, et al. Global warming and recurrent mass bleaching of corals. *Nature*, 543(7645):373–377, 2017.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Lisa Meyer, Nicole Hill, Peter Walsh, and Neville Barrett. Methods for the processing and scoring of auv digital imagery from south eastern tasmania. *Institute for Marine and Antarctic Studies Internal Report*, 2011.

Ahmad Rafiuddin Rashid and Arjun Chennu. A trillion coral reef colors: Deeply annotated underwater hyperspectral images for automated classification and habitat mapping. *Data*, 5(1):19, 2020.

Callum M Roberts, Colin J McClean, John EN Veron, Julie P Hawkins, Gerald R Allen, Don E McAllister, Cristina G Mittermeier, Frederick W Schueler, Mark Spalding, Fred Wells, et al. Marine biodiversity hotspots and conservation priorities for tropical reefs. *Science*, 295(5558): 1280–1284, 2002.

Jonathan Sauder, Viktor Domazetoski, Guilhem Banc-Prandi, Gabriela Perna, Anders Meibom, and Devis Tuia. The coralscapes dataset: Semantic scene understanding in coral reefs, 2025. URL https://arxiv.org/abs/2503.20000.

ASM Shihavuddin, Nuno Gracias, Rafael Garcia, Arthur CR Gleason, and Brooke Gintert. Image-based coral reef classification and thematic mapping. *Remote Sensing*, 5(4):1809–1841, 2013.

Wenjia Xu, Zijian Yu, Boyang Mu, Jiuniu Wang, Zhiwei Wei, and Mugen Peng. Rs-agent: Automating remote sensing tasks through intelligent agent. *Science China Information Sciences*, 69(8):180302, 2026.

Ziqiang Zheng, Yiwei Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. Marineinst: A foundation model for marine image analysis with instance visual description. In *European Conference on Computer Vision*, pp. 239–257. Springer, 2024a.

Ziqiang Zheng, Haixin Liang, Binh-Son Hua, Yue Him Wong, Put Ang, Apple Pui Yi Chui, and Sai-Kit Yeung. Coralscop: Segment any coral image on this planet. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28170–28180. IEEE, 2024b.

Ziqiang Zheng, Haixin Liang, Fong Hei Wut, Yue Him Wong, Apple Pui Yi CHUI, and Sai-Kit Yeung. Hkcoral: Benchmark for dense coral growth form segmentation in the wild. *IEEE Journal of Oceanic Engineering*, 2024c.

Ziqiang Zheng, Yuk-Kwan Wong, Binh-Son Hua, Jianbo Shi, and Sai-Kit Yeung. Coralsrt: Revisiting coral reef semantic segmentation by feature rectification via self-supervised guidance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19967–19977. IEEE, 2025.

A Appendix

A.1 Benchmark

A.1.1 Benchmark Task

Coral-Bench comprises 23 task variants, which we organize into five evaluation groups: coral cover estimation, coral recognition, visual knowledge question answering, coral diversity calculation and evenness assessment, and coral community composition and structure analysis. Together, these tasks evaluate visual recognition, biological understanding, and quantitative ecological analysis. Beyond single-image assessment, selected tasks incorporate observations from multiple sites or survey years to evaluate spatial comparison and temporal change assessment. Coral cover estimation contains five tasks: single image coral cover estimation, multi-site coral cover estimation, multi-year coral cover estimation, single image bleached coral cover estimation, and multi-site bleached coral cover estimation. The group of coral recognition comprises five recognition tasks. Family-level, genus-level, and species-level coral recognition require identifying corals at the specified taxonomic resolution. Coral growth-form classification identifies the growth form depicted in an image, whereas bleached coral recognition requires genus-level identification of bleached coral image. Visual knowledge QA evaluates the ability to interpret biological structures, processes, and ecological relationships using educational images and their associated questions. It covers six thematic subsets: polyp anatomy and tissue organization; skeletal architecture, calcification, and growth records; physiology, feeding, defense, and symbiosis; reproduction, early development, and life history; coral health, bleaching, and disease; and reef biotic interactions, food webs, and ecosystem functions. Together, these tasks assess taxonomic and morphological recognition alongside the ability to connect visual evidence with coral-specific biological knowledge. The group of coral diversity calculation and evenness assessment includes six tasks: Shannon diversity index calculation, Gini–Simpson diversity index calculation, Pielou's evenness index calculation, and a multi-site variant of each. Coral community composition and structure analysis comprises six tasks: coral composition analysis, coral richness assessment, dominant genus recognition, and their respective multi-site variants.

A.1.2 Benchmark Statistics

Table 2 summarizes the composition of our benchmark, which comprises 1,986 questions across 23 tasks organized into five task groups, with 2,535 associated images. These groups cover coral cover estimation, coral recognition, visual knowledge question answering, community composition and structure analysis, and diversity calculation and evenness assessment, spanning visual identification, domain knowledge, and quantitative ecological analysis. Coral recognition constitutes the largest group with 1,056 questions (53.2%), followed by community composition and structure analysis (303; 15.3%), diversity calculation and evenness assessment (279; 14.0%), coral cover estimation (259; 13.0%), and visual knowledge QA (89; 4.5%). The visual knowledge QA task further encompasses six topical subsets, covering coral biology and reef ecology. In addition to 14 single-image tasks comprising 1,806 questions, the benchmark includes nine multi-image tasks in multi-site or multi-year settings. These multi-image tasks account for 180 questions (9.1%), averaging 4.05 images per question.

Table 1: Performance comparison of CoralAgent and VLM baselines across four task groups in Coral-Bench.

Groups	Tasks	Kimi	MiniMax	GLM	Qwen	Gemini	Grok	GPT	CoralAgent
Coral cover	Coral cover estimation	17.04	25.19	9.63	17.78	19.26	15.56	20.00	**72.59**
	Multi-site coral cover estimation	20.00	35.00	55.00	40.00	25.00	35.00	45.00	**65.00**
	Multi-year coral cover	65.00	40.00	60.00	45.00	55.00	55.00	75.00	**80.00**
	Bleached coral cover estimation	29.69	25.00	31.25	31.25	31.25	23.44	14.06	**56.25**
	Multi-site bleached coral cover	40.00	35.00	**65.00**	45.00	55.00	35.00	55.00	60.00
Coral recognition	Bleached coral recognition	43.33	48.33	54.33	49.33	56.67	39.33	39.33	**59.00**
	Coral growth-form classification	52.10	32.52	45.45	37.06	52.80	44.76	48.95	**55.24**
	Family-level coral recognition	47.96	48.98	76.53	36.73	42.86	38.78	68.37	**80.61**
	Genus-level coral recognition	45.05	31.08	55.86	44.14	50.00	29.73	42.34	**61.71**
	Species-level coral recognition	38.00	29.33	59.33	31.33	42.67	46.00	52.00	**60.00**
	Visual knowledge QA	43.82	74.16	88.76	64.04	49.44	83.15	89.89	**95.51**
Coral composition	Coral composition analysis	25.93	33.33	34.57	19.75	20.99	22.22	27.16	**43.21**
	Coral richness assessment	32.10	24.69	34.57	27.16	20.99	40.74	19.75	**55.56**
	Dominant genus recognition	51.85	38.27	**72.84**	67.90	51.85	33.33	56.79	65.43
	Multi-site composition analysis	35.00	25.00	45.00	30.00	30.00	25.00	40.00	**65.00**
	Multi-site richness assessment	35.00	20.00	20.00	25.00	30.00	40.00	40.00	**50.00**
	Multi-site dominant genus	50.00	35.00	50.00	40.00	45.00	**65.00**	45.00	45.00
Coral diversity	Shannon diversity index	28.77	26.03	27.40	26.03	23.29	17.81	23.29	**52.05**
	Simpson diversity index	17.81	21.92	17.81	26.03	26.03	13.70	19.18	**61.64**
	Pielou's evenness index	31.51	28.77	36.99	26.03	24.66	35.62	49.32	**61.64**
	Multi-site Shannon diversity	40.00	30.00	35.00	35.00	35.00	15.00	45.00	**60.00**
	Multi-site Simpson diversity	30.00	30.00	45.00	**50.00**	45.00	35.00	45.00	**50.00**
	Multi-site Pielou's evenness	15.00	20.00	35.00	25.00	30.00	25.00	35.00	**80.00**

Note: Bold indicates the highest score in each row, including ties. QA stands for question answering.
Group labels: Coral cover denotes coral cover estimation; Coral recognition denotes coral recognition and visual knowledge question answering; Coral composition denotes coral community composition and structure analysis; Coral diversity denotes coral diversity calculation and evenness assessment.
Shortened task names: Multi-year coral cover denotes multi-year coral cover estimation; Multi-site bleached coral cover denotes multi-site bleached coral cover estimation; Visual knowledge QA denotes visual knowledge question answering; Multi-site composition analysis denotes multi-site coral composition analysis; Multi-site richness assessment denotes multi-site coral richness assessment; Multi-site dominant genus denotes multi-site dominant genus recognition; Shannon diversity index denotes Shannon diversity index calculation; Simpson diversity index denotes Gini–Simpson diversity index calculation; Pielou's evenness index denotes Pielou's evenness index calculation; Multi-site Shannon diversity denotes multi-site Shannon diversity index calculation; Multi-site Simpson diversity denotes multi-site Gini—Simpson diversity index calculation; Multi-site Pielou's evenness denotes multi-site Pielou's evenness index calculation.



Figure 5: **Representative examples of coral cover estimation tasks in Coral-Bench.** The examples span five task variants: single-image, multi-site, and multi-year coral cover estimation, together with single-image and multi-site bleached coral cover estimation. Each example includes a question, one or more survey images, and multiple-choice answers. Header colors distinguish task variants, while highlighted options with checkmarks indicate the reference answers.

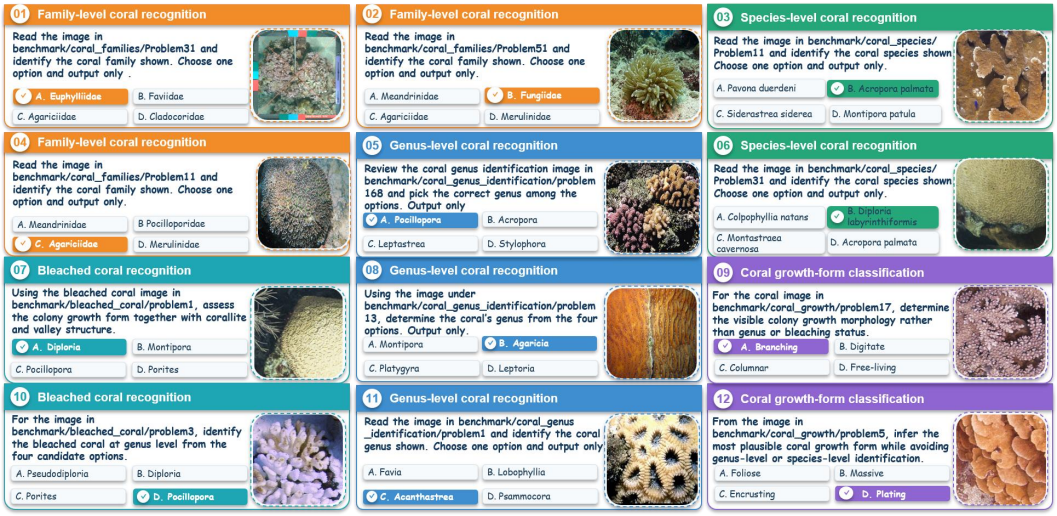

Figure 6: **Representative examples of coral recognition tasks in Coral-Bench.** The examples span five task variants: family-level, genus-level, and species-level coral recognition, bleached coral recognition, and coral growth-form classification.

A.1.3 Benchmark Creation

Figure 3 summarizes the six-stage construction of Coral-Bench: data collection, data annotation, ecological attribute extraction, task construction and ecological target derivation, question-answer (QA) construction, and validation.

Data Collection Our benchmark draws on two source categories: authoritative educational resources and publicly available coral datasets. Educational materials support the assessment of foundational biological knowledge, while coral reef survey image supports visual analysis and quantitative ecological assessment under field conditions.

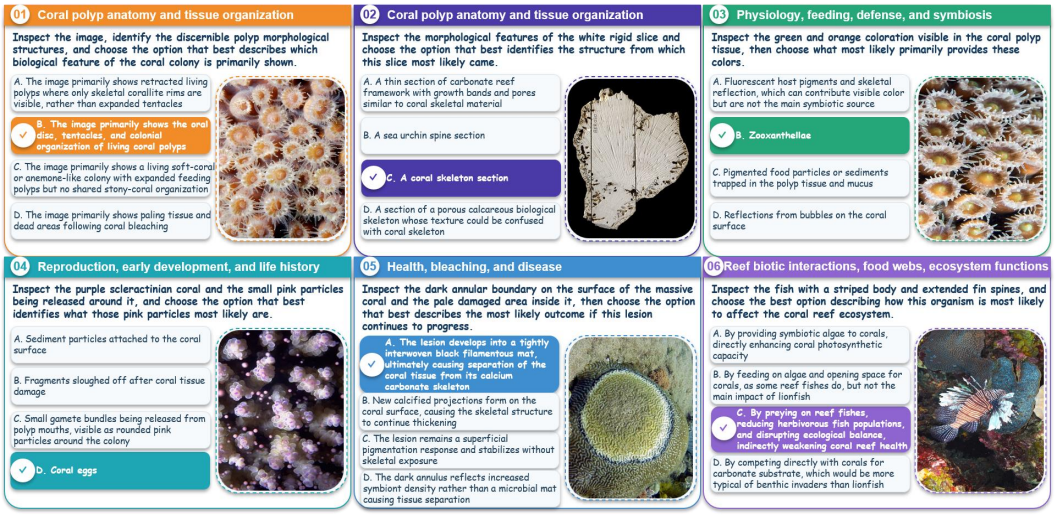

Figure 7: **Representative examples of visual knowledge question answering in Coral-Bench.** The examples cover coral polyp anatomy, skeletal structure, physiology and symbiosis, reproduction, coral health and disease, and reef ecological interactions. These questions assess the ability to connect visual evidence with biological knowledge to interpret coral structures, life processes, health conditions, and ecological relationships.

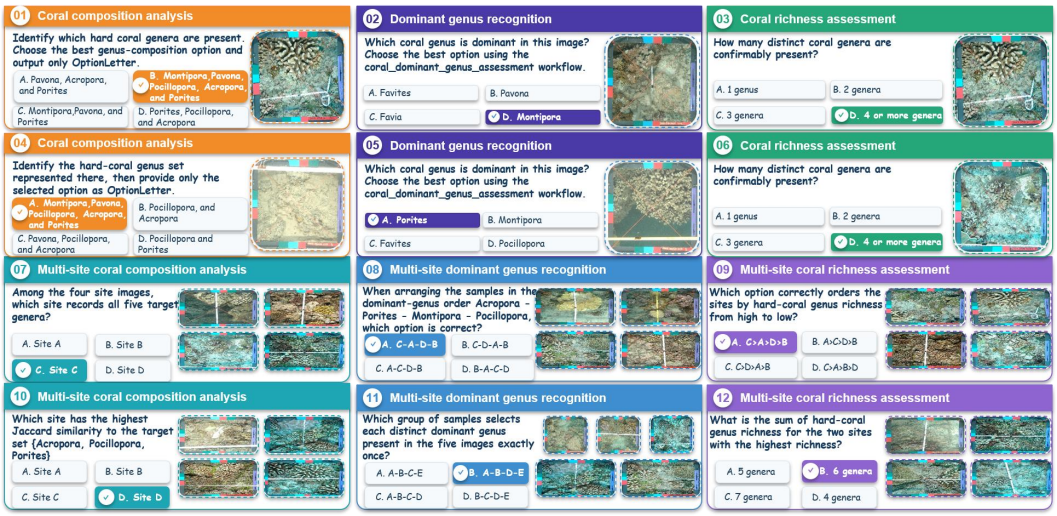

Figure 8: **Representative examples of coral community composition and structure analysis tasks in Coral-Bench.** The examples span six task variants: coral composition analysis, dominant genus recognition, and coral richness assessment, together with their multi-site variants.

Specifically, we curate 89 educational images from resources maintained by the NOAA Ocean Service, NOAA Coral Disease and Health Consortium, Flower Garden Banks National Marine Sanctuary, Smithsonian Ocean, Woods Hole Oceanographic Institution (WHOI), the Khaled bin Sultan Living Oceans Foundation, and the Caribbean Coral Reef Institute (CCRI) at the University of Puerto Rico at Mayagüez. These materials include microscopy images, anatomical and skeletal diagrams, schematics of biological processes and reef ecosystems, coral-core X-ray images, and animated GIFs. They support six thematic subsets: (1) coral polyp anatomy and tissue organization; (2) skeletal architecture, calcification, and growth records; (3) physiology, feeding, defense, and symbiosis; (4) reproduction, early development, and life history; (5) coral health, bleaching, and disease; and (6) biotic interactions, food webs, and ecosystem functions in coral reefs. Detailed source webpages and their corresponding thematic categories are provided in Table 4.

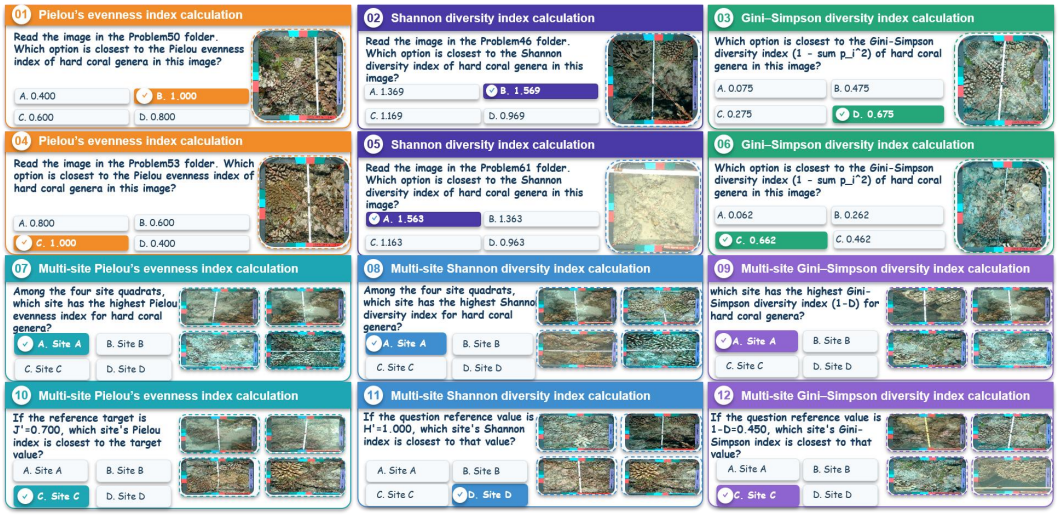

Figure 9: **Representative examples of coral diversity calculation and evenness assessment tasks in Coral-Bench.** The examples span six task variants: Shannon diversity index calculation, Gini–Simpson diversity index calculation, and Pielou's evenness index calculation, together with their multi-site variants. Questions assess genus-level diversity and evenness within individual survey images, compare indices across sites, and identify sites closest to specified reference values.

Table 2: Composition and data statistics of the benchmark. The benchmark contains five task groups, 23 tasks, 1,986 questions, and 2,535 images.

Task group	Task	Questions	Image
Coral cover estimation	Coral cover estimation	135	135
	Multi-site coral cover estimation	20	75
	Multi-year coral cover estimation	20	85
	Bleached coral cover estimation	64	64
	Multi-site bleached coral cover estimation	20	80
Coral recognition	Family-level coral recognition	98	98
	Genus-level coral recognition	222	222
	Species-level coral recognition	150	150
	Bleached coral recognition	300	300
	Coral growth-form classification	286	286
Visual knowledge QA	Coral visual knowledge QA (six topical subsets)[a]	89	89
Coral community composition and structure analysis	Coral composition analysis	81	81
	Dominant genus recognition	81	81
	Coral richness assessment	81	81
	Multi-site coral composition analysis	20	80
	Multi-site dominant genus recognition	20	89
	Multi-site coral richness assessment	20	80
Coral diversity calculation and evenness assessment	Pielou's evenness index calculation	73	73
	Shannon diversity index calculation	73	73
	Gini–Simpson diversity index calculation	73	73
	Multi-site Pielou's evenness index calculation	20	80
	Multi-site Shannon diversity index calculation	20	80
	Multi-site Gini–Simpson diversity index calculation	20	80
Total (5 groups and 23 tasks)		**1,986**	**2,535**

[a] The visual-knowledge task covers six topical subsets: (1) coral polyp anatomy and tissue organization; (2) skeletal architecture, calcification, and growth records; (3) physiology, feeding, defense, and symbiosis; (4) reproduction, early development, and life history; (5) health, bleaching, and disease; and (6) reef biotic interactions, food webs, and ecosystem functions.

Table 3: Comparison of existing coral datasets. "–" indicates that the information is not provided.

Dataset	Cross Modality	Images	Tools	Steps	Tasks
TasCPC	No	1,258	–	1	Cls.
RSMAS	No	766	–	1	Cls.
Benthoz15	No	9,874	–	1	Cls.
EILAT	No	1,123	–	1	Cls.
ATCRC	No	147	–	1	Cls.
MLC	No	2,055	–	1	Cls./Seg.
CoralMask	No	41,297	–	1	Seg.
HKCoral	No	2,515	–	1	Seg.
Coralscapes	No	2,075	–	1	Seg.
HSCR16K	Yes	16,659	–	1	Cls.
CoralVQA	Yes	12,805	–	1	–
Coral-Bench	**Yes**	**2,535**	**Yes**	≥ 5	**23**

For survey imagery, we draw on the Moorea Labeled Corals (MLC) dataset (Beijbom et al., 2012), derived from the Moorea Coral Reef Long-Term Ecological Research (MCR-LTER) program in French Polynesia. MLC contains 2,055 benthic survey images collected six monitoring sites. This multi-site coverage supports the construction of cross-site ecological analysis tasks in Coral-Bench. In addition, we additionally collect coral reef survey images acquired from 2002 to 2011 at three monitoring sites (PIN26, PIN37, and PIN40) in Flower Garden Banks National Marine Sanctuary (FGBNMS). These multi-year observations of individual sites further support the construction of interannual ecological analysis tasks in Coral-Bench. For the coral taxonomy and morphology component, we collect additional coral images to address the limited taxonomic coverage of MLC. These images are sourced from the XL Catlin Seaview Survey Project and the Corals of the World database.

Data Annotation Despite the large number of coral reef images collected, inconsistent annotation standards across sources limit their direct use for coral recognition tasks in Coral-Bench. We therefore develop a five-step data preparation and annotation pipeline: (1) extract coral patches from the original survey images; (2) retain patches containing hard corals; (3) annotate coral health status as healthy, bleached, or diseased; (4) standardize taxonomic annotations at the family, genus, and species levels; and (5) annotate coral growth forms. This pipeline yields 98 images with family-level annotations, 222 with genus-level annotations, and 150 with species-level annotations. The resulting data also include 300 images of bleached corals annotated at the genus level and 286 images with growth-form annotations. These annotated images can support five coral recognition tasks in Coral-Bench: family-level recognition, genus-level recognition, species-level recognition, genus-level recognition of bleached corals, and growth-form recognition.

Ecological Attribute Extraction While patch-level annotations support coral recognition, they do not quantify coral cover or community structure within a surveyed area. To address incomplete annotations and inconsistent labeling standards among the multiple data sources, we develop a unified point-counting pipeline for the original benthic survey images, generating new sampling locations and annotating all sampled points. The pipeline comprises five steps: (1) distribute spatially random 100 points for each original benthic survey image; (2) assign standardized benthic-category labels to the sampled points; (3) annotate each sampled point with separate labels for live-coral presence, bleaching status, and genus identity (where identifiable); (4) aggregate point-level annotations into image-level count records; and (5) associate these records with available site and year metadata. This pipeline yields a point level annotation record for each image. For every sample point, the record indicates whether it corresponds to live hard coral and, for coral sample points, further records bleaching status and genus level label. These structured records provide a quantitative basis for supporting subsequent ecological target derivation of coral cover, genus composition, dominant genus, genus richness, and diversity indices.

Table 4: Thematic subsets, representative visual content, and primary educational sources used in the foundational coral biology component.

Thematic subset	**Representative visual content**	**Primary educational sources**
Coral polyp anatomy and tissue organization	Polyp morphology, tentacles, gastrovascular cavity, tissue layers, corallites, and colonial organization	NOAA Coral Disease & Health Consortium; NOAA Ocean Service; Smithsonian Ocean; Flower Garden Banks NMS; Living Oceans Foundation
Skeletal architecture, calcification, and growth records	Calcium carbonate skeletons, corallites, skeletal growth, density bands, and coral cores	Woods Hole Oceanographic Institution (WHOI); NOAA Ocean Service; NOAA Coral Disease & Health Consortium; Flower Garden Banks NMS
Physiology, feeding, defense, and symbiosis	Tentacle feeding, nematocysts, symbiotic algae, photosynthesis, mucus production, and defense mechanisms	Smithsonian Ocean; NOAA Ocean Service; NOAA Coral Disease & Health Consortium; Flower Garden Banks NMS
Reproduction, early development, and life history	Spawning, gametes, larvae, settlement, budding, and colony formation	NOAA Coral Disease & Health Consortium; NOAA Ocean Service; Smithsonian Ocean
Coral health, bleaching, and disease	Healthy versus bleached corals, disease lesions, tissue loss, yellow-band disease, and other disease signs	NOAA Ocean Service; NOAA Coral Disease & Health Consortium; University of Puerto Rico Mayagüez Coral Reef Research Institute (UPRM CCRI)
Biotic interactions, food webs, and ecosystem functions	Coral–fish interactions, reef food webs, habitat provision, predator–prey relationships, and invasive species	NOAA Ocean Service; Flower Garden Banks NMS; Living Oceans Foundation; Smithsonian Ocean

Task Construction And Ecological Target Derivation Building on the curated educational materials, standardized coral annotations, and point-level annotation records, we construct tasks across the five areas defined in the benchmark overview. For foundational biological knowledge, we construct image-grounded biological knowledge tasks spanning the six thematic subsets and derive reference answers from the curated educational images and their accompanying source descriptions. These tasks evaluate models' ability to identify coral structures and interpret ecological relationships from visual evidence. For coral taxonomy and morphology, we construct five recognition tasks: family-level recognition, genus-level recognition, species-level recognition, genus-level recognition of bleached corals, and growth-form recognition. Their reference targets are obtained directly from the corresponding standardized annotations.For quantitative ecological tasks, we derive reference targets directly from the point-level annotation record of each image. We first aggregate the sampled points to obtain the number of valid benthic points, live hard-coral points, bleached live hard-coral points, and live hard-coral points assigned to each identified genus. Coral cover and bleached-coral cover are derived from the corresponding coral-point counts relative to the number of valid benthic points. For the coral community composition analysis, genus composition is defined by the set

of genera represented among the sampled coral points. The genus with the largest point count is identified as the dominant genus, while observed genus richness is given by the number of distinct recorded genera. For coral reef biodiversity assessment, genus-specific point counts are converted into relative genus frequencies, from which we derive Shannon diversity, Gini–Simpson diversity, and Pielou's evenness.

QA Construction Marine scientists construct the questions and reference answers, while GPT-5.4 is used exclusively to assist with answer-option generation. For each question–answer pair, the experts specify the input image or image set, formulate the question, and establish the reference answer using the educational source descriptions, standardized annotations, or ecological targets obtained in the preceding stages. We organize these question–answer pairs into multiple-choice formats, with reference answers expressed as coral categorical labels, numerical values, or short explanations.

For foundational biological knowledge, questions connect visible structures or processes to their biological significance. For example, anatomical diagrams support questions such as " Which labeled structure is responsible for [a specified function]? ", while images illustrating biological processes or coral reef interactions support questions such as " What process is depicted? " and " What ecological relationship is illustrated between the organisms? ". Marine scientists determine the answers from the corresponding educational materials. For taxonomy and morphology, related questions (such as " To which family does the coral belong? ", " Which genus does this bleached coral belong to? ", or " What growth form is shown? ") explicitly state the requested annotation level. The reference answer is the corresponding standardized family, genus, species, or growth-form label, without requiring a finer-grained recognition than the available annotation supports.

For quantitative ecological tasks, marine scientists construct task-specific questions for individual images, with reference answers drawn from previously computed ecological targets. For example, an image with a computed coral cover of 35% is paired with the question (" What is the coral cover in this survey quadrat? ") and the reference answer (" 35% "). For cross-site and cross-year tasks, experts formulate questions using image sets labeled with the corresponding sites or years. These questions compare task-specific ecological estimates (*e.g.*, identifying the site with the highest coral cover) and assess the direction and magnitude of differences between specified sites or changes between specified years. Experts establish reference answers from the corresponding ecological targets and specify task-appropriate option formats, including percentages, site labels, directional judgments (increase or decrease), and percentage-point differences. GPT-5.4 is used solely to generate candidate multiple-choice options according to these specifications, while the expert-authored questions and reference answers remain fixed.

Validation We adopt a three-stage validation procedure comprising manual review, cross-annotator consistency checks, and expert check. First, six student annotators manually inspect coral-patch labels, point-level annotation records, and LLM-generated answer options. For coral patches, they verify the taxonomic, health-status, and growth-form labels against the images. For sample points, they check live hard-coral presence, bleaching status, and genus-level label. They also assess the plausibility of the generated distractors and verify that each multiple-choice question has a unique correct option consistent with the expert-established reference answer. Second, each annotator cross-checks the coral-patch labels and point-level annotations reviewed by the other two annotators. These cross-reviews assess inter-annotator agreement and identify labeling discrepancies. When agreement falls below 95%, the disputed sampled points are re-examined against the original survey images and re-annotated. Finally, marine scientists audit 30% of the data to assess the accuracy of coral-patch labels and point-level annotations. Samples containing annotation errors identified during this audit undergo re-annotation.

A.1.4 Ecological Target Derivation

We derive all quantitative ecological targets from the standardized point-count annotations described above. For image i, let N_i denote the number of valid sampled benthic points, L_i the number of points classified as live hard coral, and B_i the number of live hard-coral points additionally annotated as bleached. Let $\mathcal{G}$ denote the set of standardized coral genera and n_{ig} the number of live hard-coral points in image i assigned to genus $g \in \mathcal{G}$. Thus,

$$0 \le B_i \le L_i \le N_i, \qquad \sum_{g \in \mathcal{G}} n_{ig} \le L_i,$$

where the second inequality allows live hard-coral points whose genus cannot be resolved.

Coral cover estimation. For $N_i > 0$, we define live hard-coral cover and bleached live hard-coral cover as

$$C_i^{\text{live}} = 100\,\frac{L_i}{N_i}, \qquad C_i^{\text{bl}} = 100\,\frac{B_i}{N_i}. \tag{4}$$

Both quantities are expressed as percentages of the valid sampled benthos. Bleaching is treated as a condition of living coral rather than a mutually exclusive benthic category; hence bleached live-coral points contribute to both B_i and L_i. Consequently, C_i^{bl} measures the benthic cover of bleached live coral, not the fraction of live coral that is bleached. The latter quantity would instead be B_i/L_i.

Coral community composition and structure analysis. Let

$$M_i = \sum_{g \in \mathcal{G}} n_{ig} \tag{5}$$

denote the number of genus-identified live hard-coral points in image i. For $M_i > 0$, the genus-composition vector $\mathbf{p}_i = (p_{ig})_{g \in \mathcal{G}}$ is defined by

$$p_{ig} = \frac{n_{ig}}{M_i}, \qquad \sum_{g \in \mathcal{G}} p_{ig} = 1. \tag{6}$$

The dominant genus and observed genus richness are then

$$g_i^\star \in \arg\max_{g \in \mathcal{G}} n_{ig}, \qquad S_i = \sum_{g \in \mathcal{G}} \mathbb{1}[n_{ig} > 0]\,. \tag{7}$$

Unresolved live hard-coral points remain part of the coral-cover target through L_i, but are excluded from $\mathbf{p}_i$, $g_i^\star$, and S_i. Accordingly, these quantities describe the sampled genus-level composition of identified live hard coral and should not be interpreted as colony-abundance estimates.

Coral diversity calculation and evenness assessment. Using the composition vector in Eq. equation 6, we compute Shannon diversity, Gini–Simpson diversity, and Pielou's evenness as

$$\begin{aligned} H_i &= -\sum_{g:\, p_{ig} > 0} p_{ig} \log p_{ig}, \\ D_i &= 1 - \sum_{g \in \mathcal{G}} p_{ig}^2, \\ J_i &= \frac{H_i}{\log S_i}, \qquad S_i > 1. \end{aligned} \tag{8}$$

We use natural logarithms throughout. H_i measures the entropy of the sampled genus distribution, while D_i is the probability that two independent draws from this distribution belong to different genera. J_i normalizes Shannon diversity by its maximum value at the observed richness S_i. These indices therefore quantify *genus-level* diversity among genus-identified live hard-coral points, rather than species-level diversity or total reef biodiversity.

Table 5: C-STAR specification for single-image live coral cover estimation. The fields constrain evidence acquisition, derivation, recovery, and reporting while preserving the measurement target.

C-STAR field	Task specification
Ω: Analysis scope	Estimate live hard-coral cover as a percentage of the valid survey region in one image.
I: Invariant constraints	Preserve the survey region, target definition, and measurement basis. Include bleached but living hard coral, and retain non-coral substrate in the survey-area denominator.
E: Required evidence	Retain the image, valid-region mask, segmentation outputs, point-count records, and provenance for the procedures used.
D: Derivation rules	Derive cover from pixel-area ratios or point-count estimates under the same survey scope and live hard-coral definition.
V: Validation criteria	Check evidence completeness and calculation consistency. Assess agreement between evidence sources and consistency across repeated measurements where applicable.
R: Recovery policy	Permit targeted label review, alternative segmentation, and point-count fallback within a bounded recovery sequence.
C: Admissible conclusions	Report estimates with their method, scope, and validation status. Flag unresolved evidence for review and restrict conclusions to the surveyed image region.

Aggregation across images, sites, and years. For a multi-image analysis unit $\mathcal{I}$, such as all images associated with a given monitoring site or survey year, we first pool the underlying point counts:

$$\begin{aligned} N_{\mathcal{I}} &= \sum_{i \in \mathcal{I}} N_i, & L_{\mathcal{I}} &= \sum_{i \in \mathcal{I}} L_i, \\ B_{\mathcal{I}} &= \sum_{i \in \mathcal{I}} B_i, & n_{\mathcal{I}g} &= \sum_{i \in \mathcal{I}} n_{ig}. \end{aligned} \tag{9}$$

The corresponding cover, composition, richness, dominance, and diversity targets are then obtained by applying Eq. equation 4 - Eq. equation 8 to these pooled counts. Thus, multi-image targets are computed from aggregated point evidence rather than by averaging image-level ecological indices. For cross-site tasks, $\mathcal{I}$ is defined separately for each monitoring site; for interannual tasks, it is defined separately for each survey year within the same site. The resulting site- or year-level targets are subsequently compared to construct the corresponding spatial and temporal benchmark questions.

A.2 Framework

A.2.1 Example: live coral cover estimation

Live hard-coral cover estimation illustrates how C-STAR preserves a measurement target while allowing changes in the analytical procedure. For a single image x, the task estimates live hard-coral cover within a fixed valid survey region. Table 5 instantiates the seven fields of $S_T = (\Omega, I, E, D, V, R, C)$ for this task.

Evidence acquisition and derivation (E, D). CoralAgent initially obtains a coral segmentation mask and computes the proportion of the valid survey region labeled as live hard coral. Image-based point counting provides complementary evidence for checking the segmentation-derived estimate.The workflow retains masks, sampling records, counts, and provenance to support inspection and recomputation.

Validation and recovery (V, R). The Validation agent checks evidence completeness and calculation consistency, and compares segmentation-derived and image-based point-count estimates. When these checks indicate insufficient or inconsistent evidence, the workflow attempts alternative segmentation and, if necessary, falls back to image-based point counting. Repeated point-count

measurements are assessed for consistency, and unresolved discrepancies remain flagged for human review. These checks assess operational consistency.

Reporting and admissible conclusions (C). The Report agent records the estimate together with its derivation method, supporting evidence, recovery history, and validation status. The result and its reporting status are stored separately, preserving the distinction between provisional estimates and results accepted under the workflow's criteria. Conclusions remain restricted to the surveyed image region and retain any unresolved limitations.

Throughout execution, Ω and I preserve the survey region and live hard-coral definition. Changes to evidence or derivation procedures require renewed validation, while C governs reporting under the resulting evidence state.

A.2.2 Reusable analytical primitives as local specifications.

CoralAgent instantiates C-STAR through a registry of reusable analytical primitives, where each primitive implements a well-defined local operation with explicit scope, invariant constraints, evidence dependencies, derivation rules, validation criteria, recovery behavior, and locally supported conclusions. Rather than associating each ecological task with a monolithic end-to-end workflow, the Router composes task-specific workflows from these registered primitives according to the corresponding C-STAR specification. This modularization allows the same analytical operation to be reused across tasks and enables individual models or tools to be replaced during execution without changing the task-level ecological semantics or measurement requirements.

A complete ecological task is rarely executed as a single operation. We therefore instantiate C-STAR using reusable analytical primitives, where each primitive represents a well-defined local analytical operation with explicit evidence dependencies, derivation rules, and validation conditions. An analytical primitive follows the same specification schema as the task-level representation:

$$a = (\Omega_a, I_a, E_a, D_a, V_a, R_a, C_a), \tag{10}$$

where the subscript a indicates that each field is local to the primitive. Ω_a and I_a define the scope and invariant constraints of the local operation; E_a, D_a, and V_a specify the evidence consumed or produced by the operation, the corresponding derivation rule, and its validation criteria; and R_a and C_a specify how local failures are handled and what downstream conclusion may rely on the validated output.

The conclusion field C_a should be interpreted locally: a primitive does not necessarily establish the final task-level conclusion. Instead, it specifies what subsequent analytical operations may infer from its validated outputs. For example, the PixelAreaMeasurement primitive used in live hard-coral cover analysis acquires a live-coral segmentation mask and counts live-coral pixels within a previously defined valid survey region. It preserves the corresponding masks, pixel counts, provenance, and validation status, but neither redefines the survey region nor independently establishes a validated coral-cover estimate.

Table 6 summarizes the coverage of the primitive registry across the five Coral-Bench task groups. In total, the registry contains 130 entries supporting all 23 benchmark tasks. Coral community composition and structure analysis has the largest number of registered primitives (34), followed by coral diversity calculation and evenness assessment (27), coral cover estimation (26), coral recognition (25), and visual knowledge QA (18). The distribution reflects the heterogeneous analytical operations required by different task families. In particular, community-level and quantitative ecological tasks require combinations of evidence acquisition, sampling, aggregation, metric computation, and validation, whereas recognition tasks are more directly centered on taxonomic and morphological inference. Overall, the registry provides reusable building blocks for composing workflows across visual recognition, ecological measurement, community analysis, and biodiversity assessment.

A.2.3 Illustrative Example: Pixel-Area Measurement

We illustrate the local C-STAR specification with PixelAreaMeasurement, an analytical primitive used in segmentation-based live hard-coral cover estimation. The example specifies the primitive's inputs, measurement derivation, validation, and permitted downstream use.

Table 6: Coverage of the analytical primitive registry across the five Coral-Bench task groups. Analytical primitives counts refer to registry entries.

Task groups	Tasks	Analytical primitives
Coral cover estimation	5	26
Coral recognition	5	25
Visual knowledge QA	1	18
Coral community composition and structure analysis	6	34
Coral diversity calculation and evenness assessment	6	27
Total	**23**	**130**

Scope and interface. The primitive takes a survey image x and the valid-region mask provided by DEFINEVALIDAREA. It acquires a live hard-coral segmentation mask and returns a pixel-count record together with the supporting masks, provenance, and local validation status. Valid-region definition and cover-ratio calculation are assigned to separate primitives; PIXELAREAMEASUREMENT neither changes the survey region nor independently produces a coral-cover estimate.

Measurement derivation. Let $\mathcal{P}_x$ be the set of pixels in image x, and let $M_{\text{valid}}, M_{\text{live}} : \mathcal{P}_x \to \{0, 1\}$ be the valid-region and live hard-coral masks, respectively. A value of one indicates membership in the valid survey region under the task-specified definition. The fixed valid-region count and the mask-based live-coral count are

$$\begin{aligned} N_{\text{valid}} &= \sum_{p \in \mathcal{P}_x} M_{\text{valid}}(p), \\ N_{\text{live}} &= \sum_{p \in \mathcal{P}_x} M_{\text{valid}}(p)\, M_{\text{live}}(p). \end{aligned} \tag{11}$$

Both quantities measure area in image pixels. The product in N_{live} restricts the coral count to the valid region. The denominator N_{valid} depends only on the supplied M_{valid} and must not change when the live-coral segmentation is updated.

Local C-STAR specification. Table 7 instantiates the seven C-STAR fields for the pixel-area measurement primitive. Specifically, E_a defines the required survey image, valid-region mask, live-coral mask, and provenance to be retained. D_a specifies how to compute the live-coral pixel count N_{live} within the valid region and combine it with the fixed valid-region pixel count N_{valid}, supporting masks, and provenance to form a traceable measurement record.

Validation Scope and Task-Level Sufficiency. A mask can be spatially aligned and yield exactly reproducible counts while omitting live hard-coral regions. Passing V_a therefore does not establish that the segmentation evidence is sufficient for a coral-cover conclusion. If task-level validation identifies such omissions, the task-level recovery policy may reacquire or repair the evidence, or invoke point-count fallback. These actions preserve the survey region and live hard-coral definition, and the resulting evidence must be re-validated before a conclusion is reported.

A.2.4 COMPOSITIONAL TASK EXPRESSIONS

A task expression specifies how local analytical operations depend on one another, when complementary evidence is required, when failures activate alternative procedures, and how analyses extend across survey units. Analytical primitives are composed using

$$\begin{aligned} e ::= a \mid e_1; e_2 \mid e_1 \parallel e_2 \\ \mid (g_1 \Rightarrow e_1) \oplus (g_2 \Rightarrow e_2) \mid \text{fallback}_f(e_1, e_2) \\ \mid \text{repeat}_{K,B}(e, g) \mid \text{map}_U(e) \mid \text{reduce}_\rho(e), \end{aligned} \tag{12}$$

Table 7: Local C-STAR specification of PIXELAREAMEASUREMENT for live hard-coral cover estimation. The primitive produces a traceable measurement record.

C-STAR field	Local specification
Ω_a: Analysis scope	Measure the pixel area labeled as live hard coral within the predefined valid survey region of a single image x.
I_a: Invariant constraints	Preserve M_{valid} and the task-specified definition of live hard coral. Non-coral pixels within the valid region remain in N_{valid}; segmentation updates must not change the denominator.
E_a: Required evidence	Use the survey image x and the predefined valid-region mask; acquire the live-coral mask and its provenance. Retain the supporting masks so that each count is traceable to the evidence from which it was computed.
D_a: Derivation rules	Compute N_{live} using Eq. equation 11. Associate it with the fixed N_{valid}, supporting masks, and provenance to form the measurement record.
V_a: Validation criteria	Check mask alignment with x, completeness of provenance, and agreement between recorded counts and counts recomputed from the retained masks. Require $N_{\text{valid}} > 0$ and $0 \leq N_{\text{live}} \leq N_{\text{valid}}$.
R_a: Recovery policy	Reacquire missing segmentation evidence, correct spatial misalignment without changing the valid survey region, and recompute inconsistent counts. Revalidate after correction; record unresolved failures and refer them to task-level recovery.
C_a: Admissible conclusions	Locally validated records support subsequent cover-ratio calculation and consistency checks, but do not independently establish a validated coral-cover estimate. Reporting a coral-cover conclusion requires task-level validation.

where a denotes an analytical primitive, while e_1 and e_2 may be individual primitives or previously composed subexpressions.

These operators serve three roles. First, $e_1; e_2$ and $e_1 \parallel e_2$ encode *evidence dependencies*. Sequential composition $e_1; e_2$ specifies that e_2 depends on the output of e_1. Complementary-evidence composition $e_1 \parallel e_2$ specifies that the analysis requires evidence from both expressions, without implying parallel execution or statistical independence between them. Second, the expression $(g_1 \Rightarrow e_1) \oplus (g_2 \Rightarrow e_2)$ selects an execution branch according to conditions evaluated against the current evidence and validation state. $\text{fallback}_f(e_1, e_2)$ switches from e_1 to a task-specified alternative e_2 when a failure of type f is detected. $\text{repeat}_{K,B}(e, g)$ performs bounded supplementary evidence acquisition, where K is the maximum number of repetitions, B is the cumulative evidence-acquisition budget, and g is the stopping condition. Repetition terminates when g is satisfied or either bound is reached. Third, map_U and reduce_ρ extend unit-level analysis to *multi-unit ecological assessment*. $\text{map}_U(e)$ applies the same C-STAR-constrained analysis to each survey unit in U, such as images, sites, or years. $\text{reduce}_\rho(e)$ then combines or compares the resulting unit-level measurements according to aggregation or comparison rule ρ, for example by computing differences or assessing temporal changes. These operators allow the same local analytical requirements to be reused for cross-site and cross-year tasks while preserving the corresponding measurement and validation constraints.

Example: live coral cover estimation. Consider live coral cover estimation for a survey image. C-STAR first fixes the analysis scope, including the valid survey region and the task-specific definition of live hard coral. A segmentation-based primitive may then produce a live-coral mask, from which a pixel-area measurement is derived and subsequently converted into a cover estimate. The resulting evidence is checked against the corresponding validation requirements.

A segmentation output may nevertheless be incomplete even when the model executes successfully and returns a correctly formatted mask. When the resulting evidence fails task-specific validation, CoralAgent can acquire supplementary evidence, repair the intermediate result, or invoke an alternative analytical procedure. In particular, image-based point counting can be used as a fallback procedure when segmentation-based evidence does not satisfy the measurement requirements.

Crucially, fallback changes the *procedure* used to obtain measurement evidence but does not change the *estimand*: the valid survey region, the definition of live hard coral, and the associated measurement and reporting requirements remain fixed. C-STAR therefore serves as an interface between ecological

Table 8: Coverage of the structured expert rule registry across the five task groups.

Task groups	Tasks	Expert rules
Coral cover estimation	5	62
Coral recognition	5	23
Visual knowledge QA	1	22
Coral community composition and structure analysis	6	37
Coral diversity calculation and evenness assessment	6	9
Total	**23**	**153**

requirements and executable workflows, constraining which execution paths can produce evidence sufficient to support an admissible ecological conclusion.

A.2.5 Structured Expert Rules

CoralAgent represents domain knowledge as structured expert rules that can be compiled into C-STAR and checked during execution, rather than using expert knowledge only as auxiliary prompt context. Each rule specifies its applicability conditions, the C-STAR fields it constrains, a normalized operational requirement, and provenance information. We organize these rules into six functional categories: measurement, diagnosis, statistics, validation, recovery, and conclusion. Depending on the task, the rules may constrain the measurement target and sampling procedure, specify taxonomic or diagnostic criteria, define statistical computations, impose evidence and consistency checks, prescribe admissible recovery actions, or restrict which conclusions may be reported. During task construction, only rules applicable to the current analysis are compiled into the corresponding C-STAR fields and analytical primitives, so that the same ecological requirements remain enforceable even when the underlying models, tools, or recovery paths change.

Table 8 summarizes the coverage of the structured expert-rule registry across the five Coral-Bench task groups. The registry contains 153 unique structured expert rules supporting all 23 benchmark tasks, including both hard constraints and advisory rules. Coral cover estimation contains the largest rule set with 62 structured expert rules, followed by coral community composition and structure analysis with 37, coral recognition with 23, visual knowledge QA with 22, and coral diversity calculation and evenness assessment with 9. In particular, tasks related to coral reef ecological analysis require explicit constraints on the survey region, ecological analysis definition, measurement denominator, evidence validity, and recovery behavior, whereas several tasks of cross-site and cross-year setting reuse a smaller set of shared statistical definitions and aggregation rules in single-image settings. Overall, the rule registry complements the analytical primitive library by specifying not only which operations may be performed, but also the scientific conditions under which their evidence, intermediate results, and final conclusions are considered valid.

A.3 Experiments

A.3.1 Implementation Details

We develop CoralAgent, a framework that leverages GPT-5.4 API as base backbone access for planning and visual reasoning. C-STAR representation and expert-knowledge rules are represented as structured records, which are translated into executable workflows through task-specific compilation routines. Each workflow step identifies an agent and the registered tools, together with its input dependencies, expected outputs, and execution parameters. The agent roles exchange intermediate results through a shared state containing visual evidence, measurements, knowledge, validation results, and artifact references. The executor resolves each step's inputs from this state and checks its declared outputs before advancing. For LLM-assisted planning, generated workflows undergo structural, dependency, and tool-compatibility checks; the planner permits two generation attempts.

Table 9: Comparison of GPT, coral-specific model augmented GPT-Agent, and CoralAgent on eight ecological analysis tasks.

Tasks	GPT-Agent	GPT	CoralAgent
Bleached coral cover estimation	14.06	14.06	**56.25**
Multi-site coral cover estimation	30.00	45.00	**65.00**
Coral composition analysis	25.93	27.16	**43.21**
Coral richness assessment	25.93	19.75	**55.56**
Dominant genus recognition	20.99	56.79	**65.43**
Shannon diversity index calculation	19.18	23.29	**52.05**
Gini–Simpson diversity index calculation	27.40	19.18	**61.64**
Pielou's evenness index calculation	34.25	49.32	**61.64**

Note: Scores are accuracy (%). All three configurations use GPT-5.4 as the backbone. GPT answers without external tools. GPT-Agent and CoralAgent have access to the same tool set. GPT-Agent autonomously selects and invokes tools without C-STAR guidance, whereas CoralAgent uses C-STAR-guided execution. Bold indicates the highest accuracy in each row.

The tool layer integrates specialist-models, API-based visual analysis, and local numerical calculation tools. Coral-cover workflows use precomputed segmentation artifacts indexed as CoralSCOP model and SAM2. For tasks that rely on coral genus level recognition, such as coral community composition and structure analysis, and coral diversity and evenness assessment, the CoralAgent additionally accesses CoralSRT (Zheng et al., 2025) annotations and semantic-mask evidence. Task-specific visual prompts elicit structured observations from full images, sampled-point overlays, or local image regions. Multi-image analyses retain image-to-site or image-to-year associations and use task-specific aggregation and comparison routines to form collection-level results.

A.3.2 Ablation Study

Effect of C-STAR-guided execution. We compare GPT, GPT-Agent, and CoralAgent on eight ecological analysis tasks spanning coral cover estimation, community analysis, and diversity calculation (Table 9). All three configurations use GPT-5.4. GPT answers without external tools, while GPT-Agent and CoralAgent share the same tool set and use autonomous tool selection and C-STAR-guided execution, respectively. CoralAgent achieves the highest accuracy on all eight tasks, with improvements over GPT-Agent ranging from 17.28 to 44.44 percentage points. The largest gains occur in dominant genus recognition, where accuracy increases from 20.99% to 65.43%, and bleached coral cover estimation, where it increases from 14.06% to 56.25%. However, GPT-Agent does not consistently outperform the tool-free GPT baseline: it improves accuracy on coral richness assessment and Gini–Simpson diversity calculation, matches GPT on bleached coral cover estimation, and performs worse on the remaining five tasks. A plausible explanation is that successful invocation of a specialized model does not guarantee reliable or task-appropriate predictions. A model may execute successfully yet produce inaccurate outputs that provide an unreliable basis for subsequent ecological computations. If these outputs are accepted without adequate validation or recovery, errors can propagate through the workflow. Successful tool execution therefore does not necessarily translate into effective ecological assessment. C-STAR is designed to address this gap by governing evidence acquisition, ecological computation, validation, and recovery, so that intermediate results are checked against task-specific requirements and insufficient evidence can trigger corrective actions. With the same backbone and available tools, the observed gains support the value of this coordinated execution framework across the evaluated tasks.

Effect of validation, and recovery from C-STAR. We examine the contributions of validation and recovery on Gini–Simpson diversity calculation (Table 10). All configurations use GPT-5.4, and all agent variants have access to the same tools. Equipping GPT-Agent with tools increases accuracy from 19.18% to 27.40%, whereas additionally providing the expert rules used by CoralAgent yields no further improvement. A possible explanation is that these expert rules primarily guide VLM-based coral genus recognition, evidence acquisition, and evidence assessment. They do not directly modify specialized models or correct errors in predictions already returned by successful tool calls. Full

Table 10: Ablation study on Gini–Simpson diversity calculation.

Setting	Accuracy (%)
GPT (no tools)	19.18
GPT-Agent + same tools	27.40
GPT-Agent + expert rules	27.40
CoralAgent w/o recovery	38.36
CoralAgent w/o validation	49.32
CoralAgent (full)	**61.64**

Note: All settings use GPT-5.4; GPT-Agent and CoralAgent share same tools.

CoralAgent achieves 61.64% accuracy, outperforming both GPT-Agent variants by 34.24 percentage points. This comparison supports the value of incorporating domain knowledge into C-STAR-guided execution beyond making tools and rules available.

Removing validation reduces accuracy from 61.64% to 49.32%, a decrease of 12.32 percentage points. Gini–Simpson diversity is calculated from the relative frequencies of coral genera in the sampled community. Errors in genus recognition or composition estimation can therefore affect the final index, even when the calculation itself is correct. C-STAR checks whether the supporting evidence is complete, the calculations are consistent with that evidence, and the results meet task-specific requirements. The decline in accuracy after removing validation suggests that these checks help prevent unreliable intermediate outputs from propagating into the final diversity estimate.

Removing recovery reduces accuracy to 38.36%, a decrease of 23.28 percentage points from the full framework. Although validation can identify problems in intermediate evidence, these problems may remain unresolved without corrective action. C-STAR's recovery policy allows the system to collect additional evidence, correct or adjudicate existing observations, or use alternative analytical models or procedures while preserving the survey scope and measurement definition. The larger performance drop after removing recovery suggests that addressing evidence deficiencies is particularly important for this task. Together, the two ablations support the complementary roles of validation and recovery in C-STAR. Validation checks whether the evidence satisfies task-specific requirements, while recovery addresses identified problems so that the resulting evidence can be reassessed before reporting ecological conclusions. Overall, these results highlight the importance of integrating validation and recovery into C-STAR to improve the accuracy and reliability of quantitative ecological analysis.