A Hybrid Framework for Automated Coral Reef Risk Assessment by Zero-Shot Visual Perception and Fuzzy Cognitive Map Inference

Fabio Porcelli[1][0009-0008-5250-8662], Simone Cioffi[1][0009-0007-4054-7560], Emanuel Di Nardo[1][0000-0002-6589-9323], and Angelo Ciaramella[1][0000-0001-5592-7995]

Department of Science and Technologies, University of Naples Parthenope, Naples, Italy

Abstract. Coral reef ecosystems are among the most biodiverse and ecologically critical environments on Earth, yet they face unprecedented decline due to climate change, ocean warming, and anthropogenic pressure. Effective monitoring at scale remains a major challenge, as conventional survey methods rely heavily on expert labour and in-situ access. In this paper, we propose a hybrid framework for automated coral reef ecological risk assessment that combines zero-shot visual reasoning with causal inference. We employ Gemma 4-E4B, an open-weight multimodal large language model, as a perception layer that extracts twenty-five quantitative ecological features from raw underwater survey images drawn from the Coralscapes dataset, the first general-purpose semantic segmentation benchmark for coral reefs. The extracted feature vector is fed as clamped input activations into a Fuzzy Cognitive Map whose causal weight matrix is initialised from expert biological prior knowledge. Iterative sigmoid dynamics propagate causal influence through thirty-one concept nodes until convergence, yielding a six-dimensional risk profile that quantifies ecosystem health, pollution risk, thermal stress, habitat degradation, recovery potential, and heritage integrity. The entire pipeline operates in a zero-shot regime, requiring no task-specific fine-tuning and no manual annotation, and is deployable on commodity single-GPU hardware. Experiments on 50 test images demonstrate the feasibility of the approach and provide quantitative insight into the ecological risk landscape of reef systems.

Keywords: Fuzzy Cognitive Maps · Multimodal Large Language Models · Coral Reef Monitoring · Ecological Risk Assessment · Underwater Image Analysis · Zero-shot Visual Reasoning

1 Introduction

Coral reefs are among the most biodiverse and ecologically valuable ecosystems on Earth. Although they cover less than 1% of the ocean floor, they sustain approximately 25% of all known marine species [14]. Despite their disproportionate ecological importance, coral reefs are facing an unprecedented crisis driven by anthropogenic climate change, ocean warming, pollution, and destructive human activities. The fourth global mass coral bleaching event, which unfolded between 2023 and 2025, affected 84% of the world's reef systems across 82 countries and territories [11], the largest and most severe on record. As of 2024, the IUCN Red List has classified 44% of all reef-building coral species as threatened with extinction [13], a figure that has risen sharply from one third in 2008. Under a 2°C warming scenario, IPCC projections indicate that virtually all warm-water coral reefs could suffer catastrophic decline by the end of this century [12].

Effective conservation and management of these ecosystems requires continuous, high-resolution monitoring of reef health. However, conventional survey methods, including manual transect dives and point-intercept sampling, are inherently limited in spatial and temporal coverage, as they depend heavily on expert labour and in-situ access. The Red Sea, in particular, represents a critically understudied and ecologically unique reef system, characterised by high thermal tolerance but equally exposed to localised stressors such as coastal development, sedimentation, and tourism pressure. The Coralscapes dataset [23], collected from 35 dive sites across five Red Sea countries, underscores both the richness and the monitoring complexity of this region.

Recent advances in computer vision and deep learning have opened promising avenues for automating reef survey workflows. Nevertheless, supervised approaches require large volumes of expert-annotated images, which are expensive to produce and rarely available at scale in the marine domain. Multimodal Large Language Models (MLLMs) [22, 18, 8] have emerged as a compelling alternative, demonstrating strong zero-shot visual reasoning capabilities across diverse scientific domains. At the same time, Fuzzy Cognitive Maps (FCMs) [15, 7] provide a principled framework for encoding expert ecological knowledge as a causal graph and propagating it through iterative inference, a property that makes them well suited to risk assessment tasks where ground-truth labels are scarce and interpretability is critical.

In this paper, we propose a hybrid framework that combines these two complementary paradigms for automated coral reef risk assessment. We employ Gemma 4 [9, 6], an open-weight multimodal language model, as a zero-shot visual perception layer that extracts twenty-five quantitative ecological features from raw underwater survey images drawn from the Coralscapes dataset. These features are then fed as input activations into an FCM whose causal structure is initialised from expert biological prior knowledge. The FCM produces six ecosystem risk scores as output: *ecosystem health*, *pollution risk*, *thermal stress risk*, *habitat degradation risk*, *recovery potential*, and *heritage integrity risk*. The entire pipeline requires no task-specific fine-tuning and no manual annotation, making it immediately deployable on new reef systems.

The main contributions of this work are threefold:

- We introduce the first pipeline combining a zero-shot MLLM perception layer with a Fuzzy Cognitive Map for automated coral reef ecological risk assessment, requiring no task-specific fine-tuning and no manual annotation.

– We demonstrate that structured ecological features can be reliably extracted from raw underwater images using prompt-engineered VLMs without any domain-specific fine-tuning.
– We release a fully reproducible pipeline built exclusively on open-weight models and public datasets, deployable on commodity single-GPU hardware, lowering the barrier for automated reef monitoring in resource-constrained research settings.

The remainder of this paper is organized as follows. Section 2 reviews related work on FCMs, their ecological applications, the Coralscapes dataset, and VLMs applied to marine imagery. Section 3 describes the proposed methodology in detail. Section 4 presents the experimental setup and results. Section 5 concludes the paper and describes directions for future work.

2 Related Work

The theoretical foundation of Fuzzy Cognitive Maps rests on the broader framework of fuzzy set theory, introduced by Zadeh [25] as a generalisation of classical Boolean logic to handle gradual, imprecise, or uncertain information. This framework has been extended by Mamdani and Assilian [16], who introduced the first rule-based fuzzy inference system for process control, and by Takagi and Sugeno [24], whose functional consequent model became a standard architecture for data-driven fuzzy modelling. More recently, fuzzy rule-based systems have attracted growing interest within the field of eXplainable Artificial Intelligence (XAI), owing to their ante-hoc interpretability: the knowledge encoded in a fuzzy rule base can be directly inspected by domain experts, provided that the rule base remains compact. Camastra et al. [4] address this compactness requirement by proposing a rule minimisation algorithm grounded in rough set theory, which reduces rule base size while preserving classification accuracy. The integration of fuzzy logic with deep learning architectures has been further explored by Di Nardo and Ciaramella [5], who introduce a Fuzzy Relational Neural Network for image classification that replaces standard convolutional operators with fuzzy counterparts, achieving improved gradient-based interpretability. Beyond classification tasks, fuzzy inference systems have been applied to environmental decision support: Camastra et al. [3] developed a Fuzzy Decision Support System for the risk assessment of the deliberate release of genetically modified organisms, demonstrating that fuzzy inference can operationalise expert ecological knowledge in domains characterised by incomplete and imprecise data — a setting closely analogous to the coral reef monitoring problem addressed in this work.

The transition from standalone fuzzy inference systems to relational causal structures is achieved by Fuzzy Cognitive Maps, which were introduced by Kosko [15] as weighted signed directed graphs for representing causal knowledge [7]. In an FCM, nodes (concepts) encode the activation level of domain variables in the continuous range $[0, 1]$, while edges carry signed weights in $[-1, 1]$ expressing the polarity and strength of causal influence between concepts. System dynamics are governed by an iterative update rule, typically involving a sigmoid activation function, that propagates concept activations until convergence to a fixed point. Beyond expert-elicited maps, recent work has explored data-driven FCM construction from unstructured sources: Ciaramella et al. [17] propose a method for automatically extracting FCMs from enriched Twitter conversations, enriching raw messages with domain-specific corpora before transforming them into weighted causal graphs, thereby reducing the manual annotation burden typically associated with FCM elicitation — a direction that parallels our use of a zero-shot vision-language model as an automated perceptual front-end.

FCMs have found extensive application in multiple fields for decision support, owing to their capacity to integrate heterogeneous expert knowledge into a unified causal model. Hobbs et al. [10] demonstrated the utility of FCMs for defining management objectives in complex freshwater ecosystems, showing that causal graph inference can guide policy decisions even in the absence of complete quantitative data. Özesmi and Özesmi [19] formalised a multi-step FCM methodology for ecological modelling based on stakeholder knowledge, applying it to wetland conservation and lake ecosystem management across multiple case studies. These works establish FCMs as a principled tool for ecological risk reasoning, particularly in domains where ground-truth labels are scarce and interpretability is a requirement. Our work extends this tradition to the coral reef domain, replacing manually elicited stakeholder inputs with automatically extracted visual features from a multimodal large language model, thus enabling scalable and annotation-free ecosystem monitoring.

Multimodal Large Language Models (MLLMs) have rapidly evolved from text-only systems to unified architectures capable of jointly reasoning over images and natural language [22, 18, 8]. This progression has opened new opportunities for zero-shot visual understanding in specialised scientific domains, where the cost of annotated training data is prohibitive. In this work, we adopt Gemma 4 [9, 6] as the visual perception backbone. Released by Google DeepMind in April 2026 under the Apache 2.0 license, Gemma 4 extends the multimodal capabilities of Gemma 3 with native audio input on the small variants and uniform image support across the entire family, available in four variants (E2B, E4B, 26B-A4B MoE, and 31B dense), with context windows up to 256K tokens. Two architectural features make it particularly suited to our pipeline. First, native system prompt support enables structured and controllable model behaviour, which we exploit to enforce domain-specific JSON output format without fine-tuning. Second, the Apache 2.0 open license ensures full scientific reproducibility and enables deployment on single-GPU environments such as the Kaggle T4 infrastructure used in this work.

The application of Vision-Language Models (VLMs) to marine and underwater imagery is a rapidly developing research frontier. Zheng et al. [26] introduced MarineGPT, an instruction-tuned multimodal model designed for ocean-domain queries, demonstrating that domain-specific adaptation of general-purpose LLMs enables richer biological descriptions of underwater scenes. More recently, Alawode et al. [1] proposed AquaticCLIP, a vision-language foundation model pre-

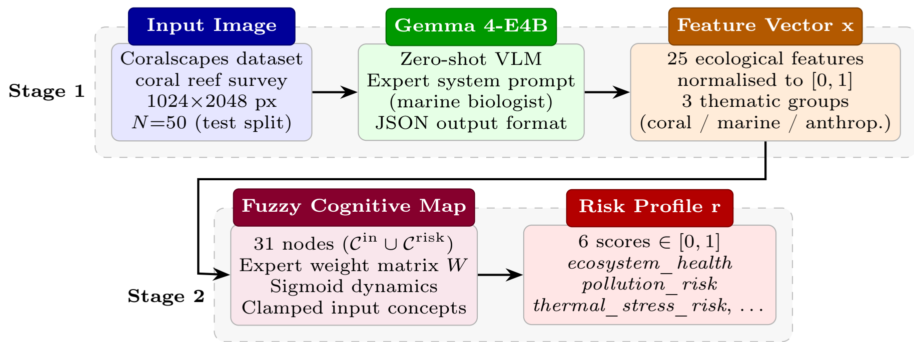

Fig. 1. Overview of the proposed pipeline. Stage 1: Gemma 4-E4B processes each image in a zero-shot regime and returns a 25-dimensional ecological feature vector. Stage 2: the feature vector is fed as input activations into a FCM with an expert-initialised weight matrix; iterative sigmoid dynamics converge to a six-dimensional risk profile $\mathbf{r}$.

trained on two million underwater image-text pairs, achieving strong zero-shot performance across detection, segmentation, and counting tasks in aquatic environments. Differently from these approaches, which require domain-specific pre-training or fine-tuning, our method leverages a general-purpose open-weight VLM (Gemma 4) in a pure zero-shot regime, relying entirely on prompt engineering to elicit structured ecological feature estimates. This design choice prioritises reproducibility and accessibility, while the FCM layer compensates for the absence of domain-specific supervision by encoding expert ecological knowledge directly into the causal graph structure.

3 Proposed Methodology

3.1 The Coralscapes Dataset

Coral reef monitoring has long relied on labour-intensive manual surveys, and the lack of large annotated datasets has historically constrained the development of automated vision tools for this domain. Sauder et al. [23] address this gap with the release of Coralscapes[1], the first general-purpose dense semantic segmentation dataset for coral reef environments. Coralscapes comprises 2,075 high-resolution images (1024×2048 px) collected across 35 dive sites in five countries in the Red Sea, annotated with 174k expert-labelled polygons spanning 39 benthic classes — ranging from live coral morphotypes and algal communities to anthropogenic debris and sediment. In our work, we use Coralscapes as the image source for our evaluation pipeline, exploiting its 39-class taxonomy to ground the ecological feature vocabulary extracted by Gemma 4.

The proposed framework consists of two sequential stages: (i) zero-shot ecological feature extraction via a multimodal large language model, and (ii) causal

[1] Dataset page: `https://huggingface.co/datasets/EPFL-ECEO/coralscapes`.

Table 1. Ecological features extracted by Gemma 4-E4B from each underwater survey image. Values are normalised to $[0, 1]$.

Group	Features
Coral-specific (12)	coral_cover_alive, coral_bleaching, coral_morphology_diversity, dead_coral_algae, juvenile_coral_recruitment, coralline_algae, substrate_rugosity, rubble_cover, sediment_stress, water_turbidity, algae_coverage, sand_cover
General marine (7)	fish_abundance, seagrass_coverage, macro_invertebrate_diversity, sponge_coverage, coral_disease_signs, crown_of_thorns_starfish, soft_coral_abundance
Anthropogenic (6)	anchoring_mechanical_damage, ghost_fishing_gear, blast_fishing_damage, image_assessability, sea_urchin_density, marine_debris

risk inference via a Fuzzy Cognitive Map with expert-initialised weights, as illustrated in Fig. 1.

3.2 Stage 1: Zero-Shot Feature Extraction with Gemma 4

Given an underwater survey image $\mathbf{I}$ drawn from the Coralscapes dataset [23], we query Gemma 4-E4B [9, 6] to produce a structured ecological assessment without any task-specific fine-tuning. The model is invoked with a system prompt that instructs it to behave as an expert marine biologist, and a user prompt that requests the estimation of $P = 25$ ecological features, each assigned a continuous value in $[0.0, 1.0]$, returned as a JSON object. The feature vocabulary $\mathcal{F} = \{f_1, \ldots, f_{25}\}$ is partitioned into three thematic groups, as summarised in Table 1.

The visual token budget is set to 1120 tokens to maximise image comprehension within GPU memory constraints. For the larger Gemma 4 variants (26B-A4B, 31B), the pipeline supports 4-bit NF4 quantisation via `bitsandbytes`, enabling inference on a single T4 GPU; the experiments reported here use the unquantised E4B variant. The model is conditioned via a structured system prompt that assigns it the role of an expert marine biologist specialised in reef ecosystems. The prompt instructs the model to analyse the visual content of the image strictly from an ecological standpoint, to estimate each feature independently based solely on observable visual evidence, and to return the output exclusively as a valid JSON object with no additional text. This design enforces output parsability without post-processing heuristics and grounds the model's reasoning in domain-specific ecological semantics, without requiring any parameter update or fine-tuning. The extracted feature vector $\mathbf{x} = (x_1, \ldots, x_{25}) \in [0, 1]^{25}$ is subsequently passed to the FCM stage.

3.3 Stage 2: Fuzzy Cognitive Map Risk Inference

FCM Formulation. Following Kosko [15], a Fuzzy Cognitive Map is a weighted signed directed graph $\mathcal{G} = (\mathcal{C}, W)$ where $\mathcal{C}$ is a set of N concepts and $W \in [-1, 1]^{N\times N}$ is the causal weight matrix. A positive weight $w_{ij} > 0$ indicates that concept c_i causally increases concept c_j; a negative weight $w_{ij} < 0$ indicates an inhibitory relationship; $w_{ij} = 0$ denotes absence of direct causal influence. In our framework, the concept set is partitioned into:

$$\mathcal{C} = \mathcal{C}^{\text{in}} \cup \mathcal{C}^{\text{risk}} \tag{1}$$

where $\mathcal{C}^{\text{in}} = \{f_1, \dots, f_{25}\}$ are the *input concepts* clamped to the values extracted by Gemma 4, and $\mathcal{C}^{\text{risk}}$ comprises six *risk concepts* whose activation is inferred by the FCM dynamics: *ecosystem_ health*, *pollution_ risk*, *thermal_ stress_ risk*, *habitat_ degradation_ risk*, *recovery_ potential*, and *heritage_ integrity_ risk*.

Dynamics and Convergence. The activation of risk concepts at iteration $t+1$ is governed by:

$$c_j^{(t+1)} = \sigma\left(\sum_{i\in\mathcal{C}^{\text{in}}} w_{ij}\, x_i \;+\; \sum_{k\in\mathcal{C}^{\text{risk}}} w_{kj}\, c_k^{(t)}\right), \quad j \in \mathcal{C}^{\text{risk}}, \tag{2}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid activation function and x_i denotes the clamped value of input concept i as provided by Gemma 4. Input concept activations remain fixed throughout the simulation, preserving the perceptual information extracted by the vision-language model. Risk concepts are initialised to zero and iterate according to Eq. (2) until convergence, defined as $\|\mathbf{c}^{(t+1)} - \mathbf{c}^{(t)}\|_\infty < 10^{-5}$, or a maximum of 50 iterations is reached.

Expert-Initialised Weight Matrix. The causal weight matrix W is initialised from a biological prior encoding established ecological relationships [15, 10, 2]. To make the construction reproducible, we follow a three-step elicitation protocol. First, the 25 perceptual concepts are selected from the Coralscapes taxonomy and from visually assessable reef health indicators commonly used in benthic surveys, excluding variables that cannot be inferred from a single image, such as water temperature time series or chemical measurements. Second, each candidate causal relation is assigned a polarity and an ordinal strength according to marine-ecological prior knowledge: weak links are mapped to ±0.20, moderate links to ±0.40–±0.60, and strong links to ±0.70–±0.90. Third, relations are checked for ecological consistency by scenario simulation, verifying that known reef states produce the expected ordering of health and risk indicators. Excitatory links (e.g. $w_{coral\_\,cover\_\,alive,\,ecosystem\_\,health} = +0.80$) and inhibitory links (e.g. $w_{coral\_\,bleaching,\,recovery\_\,potential} = -0.70$) reflect this expert-defined prior. Zero entries indicate absence of direct causal connection. The weight matrix is kept fixed throughout inference, ensuring full reproducibility and interpretability of the risk scores.

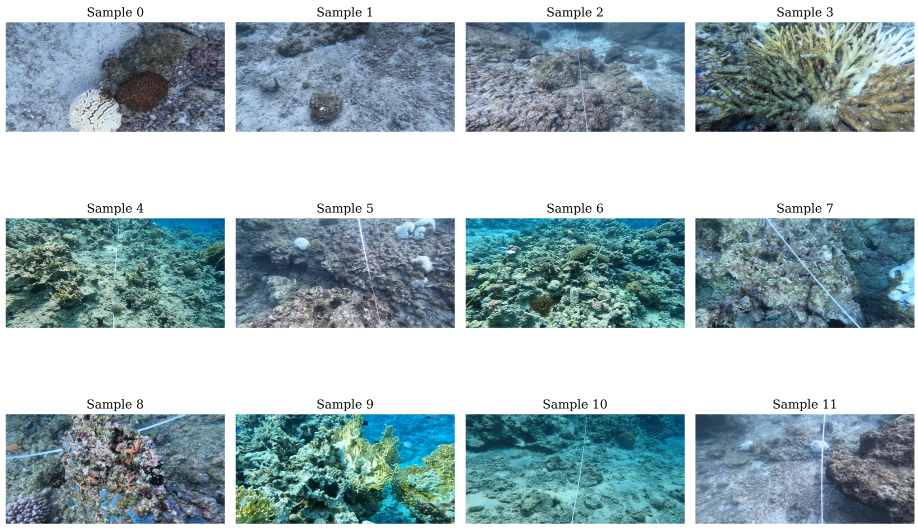

Fig. 2. Representative sample of twelve underwater survey images from the Coralscapes test split used in this study. Images span a wide range of reef conditions, from structured coral assemblages to heavily degraded substrates dominated by rubble and sand.

Risk Score Output. Upon convergence of Eq. (2), the final activations of the six risk concepts constitute the output risk profile $\mathbf{r} \in [0, 1]^6$ for the input image. Values close to 1 indicate high risk or high health depending on the concept polarity, while values close to 0 indicate the opposite. The full pipeline processes $N = 50$ images from the test split of Coralscapes, writing per-image feature vectors and risk scores to a `.jsonl` output file for subsequent analysis.

4 Experiments and Results

4.1 Experimental Setup

All experiments are conducted on the test split of the Coralscapes dataset [23], processing $N = 50$ underwater survey images representative of reef conditions. A sample of twelve input frames is shown in Fig. 2, illustrating the diversity of reef states present in the dataset — from relatively structured coral assemblages to heavily degraded substrates dominated by rubble, sand, and bleached colonies. Feature extraction is performed using Gemma 4-E4B [9, 6] with a visual token budget of 1120 and zero-shot prompting, running on a single NVIDIA T4 GPU via the Kaggle platform. The FCM inference stage uses the expert-initialised weight matrix described in Section 3, with sigmoid dynamics and input concept clamping.

4.2 Comparison with Existing Approaches

Direct quantitative comparison with supervised coral segmentation systems is not straightforward, because the proposed framework does not predict pixel-wise semantic masks and does not use segmentation ground truth during inference. Existing supervised approaches for underwater scene understanding primarily optimise class-level or pixel-level recognition accuracy, whereas our output is a six-dimensional ecological risk profile derived from image-level visual evidence and causal reasoning. We therefore position the method as complementary to segmentation pipelines: a supervised model can estimate benthic classes with high precision when annotated training data are available, while the present zero-shot pipeline trades pixel-level accuracy for annotation-free deployment, interpretable causal aggregation, and explicit risk indicators.

Qualitatively, the framework addresses two limitations of purely vision-based baselines. First, raw VLM feature estimates remain local and descriptive; by themselves they do not encode how bleaching, rubble, algae, disease, structural complexity, and anthropogenic damage jointly affect reef status. The FCM layer converts these heterogeneous observations into causally linked risk dimensions. Second, black-box classifiers provide limited transparency about why a reef scene is considered degraded. In contrast, each FCM output can be traced to signed weighted links, enabling inspection of the ecological factors that increase or decrease each risk score.

4.3 Component Ablation

Because no image-level ground-truth risk labels are available for the Coralscapes test images, we report an ablation-oriented analysis of component roles rather than supervised accuracy scores. Removing the Gemma 4 perception layer leaves the FCM without image-specific evidence; the model can still simulate synthetic scenarios, but it no longer performs automated assessment of real survey images. Removing the FCM layer yields only 25 independent visual feature estimates, which are useful descriptors but do not provide the six target risk indicators or account for indirect causal interactions among reef stressors. Finally, replacing the expert matrix with an identity or zero-interaction mapping suppresses causal propagation, eliminating the moderation effects visible in the full model, such as the inhibitory influence of ecosystem health and recovery potential on habitat degradation risk. This analysis indicates that the two components play distinct roles: Gemma 4 provides the perceptual interface to raw imagery, while the FCM provides transparent ecological aggregation and inference.

4.4 FCM Structure and Causal Graph

The causal structure of the FCM is visualised in Fig. 3. The graph comprises 31 nodes (25 input concept nodes fed by Gemma 4 feature estimates and 6 risk concept nodes computed by FCM dynamics) connected by 94 directed weighted edges. Green solid edges represent excitatory causal links ($w > 0$) and red dashed edges represent inhibitory links ($w < 0$). The most highly connected risk concept is *habitat_degradation_risk*, which receives both direct excitatory input from stressor features (e.g. blast fishing damage +0.90, rubble cover +0.75) and indirect inhibitory modulation through *recovery_potential* (−0.55) and *ecosystem_health* (−0.60).

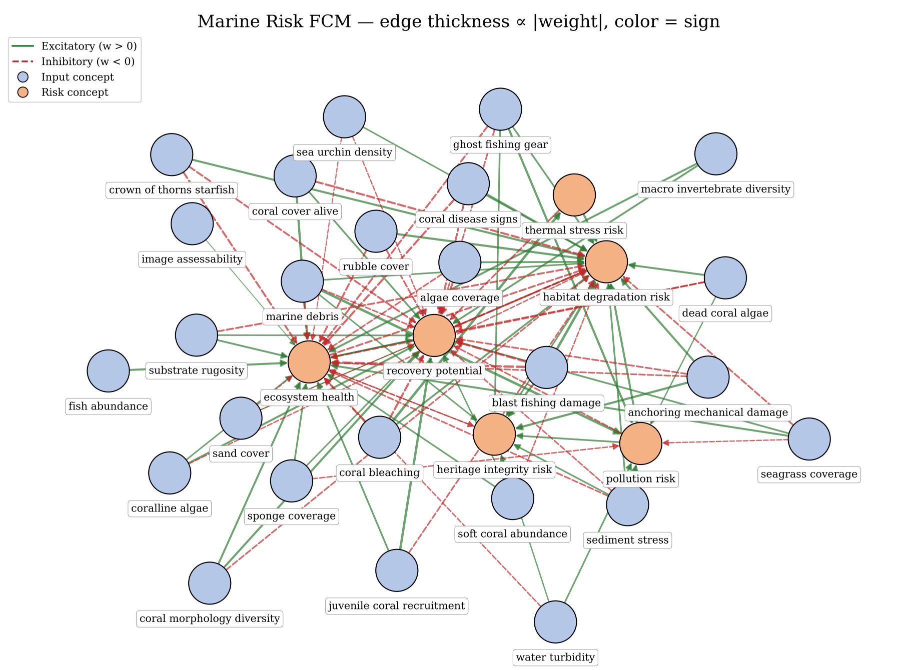

Fig. 3. Causal graph of the expert-initialised Fuzzy Cognitive Map. Green solid edges denote excitatory causal links ($w_{ij} > 0$); red dashed edges denote inhibitory links ($w_{ij} < 0$). The graph comprises 31 concept nodes and 94 directed weighted edges.

4.5 Convergence Analysis

Fig. 4 shows the activation trajectory of the six risk concepts over FCM iterations for image `Sample 0`. All risk concepts converge within approximately 4–5 iterations from zero initialisation, demonstrating the rapid and stable inference behaviour of the clamped-input FCM formulation. The final converged risk profile for `test_0` is: *ecosystem_health* = 0.69, *pollution_risk* = 0.59, *thermal_stress_risk* = 0.52, *habitat_degradation_risk* = 0.48, *recovery_potential* = 0.51, and *heritage_integrity_risk* = 0.50. This profile is consistent with a moderately degraded reef exhibiting residual structural complexity (substrate rugosity = 0.70, coral morphology diversity = 0.60) alongside early-stage stressors.

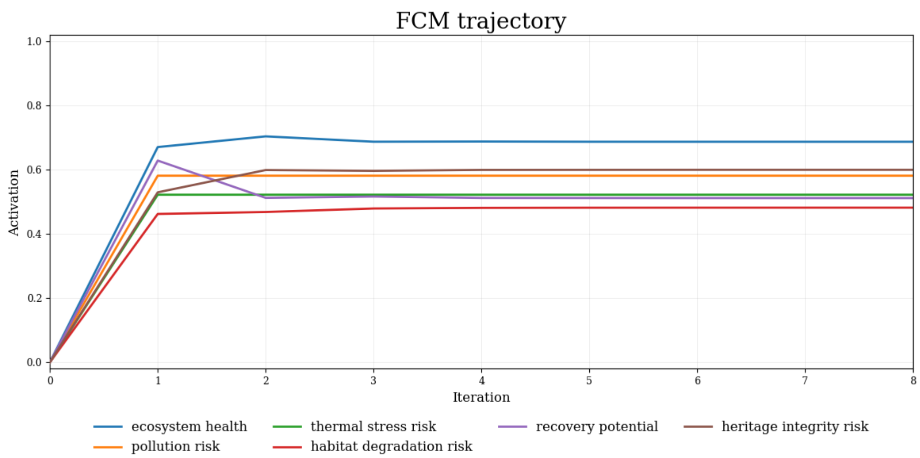

Fig. 4. Activation trajectory of the six risk concepts over FCM iterations for image `test_0`. All concepts converge within 4–5 iterations. *Ecosystem_health* reaches the highest final activation (0.69), while *habitat_degradation_risk* stabilises at 0.48.

The post-convergence state of the FCM is further illustrated in Fig. 5, where node size and colour encode the final activation level. Input concept nodes show a wide range of activations reflecting the ecological heterogeneity of the scene, with *image_assessability* (0.90) and *substrate_rugosity* (0.70) dominating, while stress-related features such as *coral_bleaching* (0.10) and *coral_disease_signs* (0.00) remain low.

4.6 Scenario Analysis

To validate the ecological consistency of the expert weight matrix, we evaluate the FCM on four synthetic scenarios constructed from domain knowledge, spanning the full range of reef health conditions observable in the Coralscapes dataset. Results are reported in Table 2.

The results demonstrate that the FCM weight matrix correctly encodes the expected ecological ordering across all scenarios. The healthy reef scenario yields the highest ecosystem health (0.986) and recovery potential (0.959) with near-zero habitat degradation risk (0.068). The COTS and blast fishing scenario produces the most severe risk profile, with ecosystem health collapsing to 0.081 and habitat degradation risk reaching 0.962, consistent with the catastrophic structural damage associated with destructive fishing practices and crown-of-thorns outbreaks. The post-bleaching scenario correctly reflects intermediate degradation with elevated thermal stress risk (0.556) and strongly suppressed

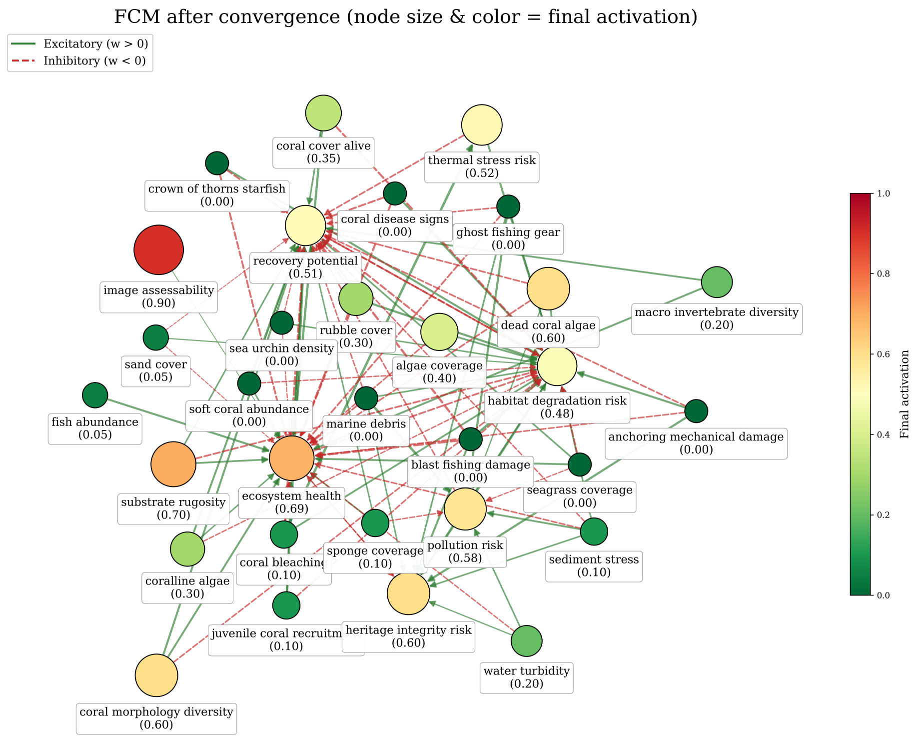

Fig. 5. FCM state after convergence for image `test_0`. Node size and colour (green = low activation, red = high activation) encode the final activation value. Risk concepts are identifiable by their intermediate-to-high activations clustered in the centre of the graph.

recovery potential (0.144). The urchin barren scenario yields an intermediate profile (EH = 0.674, HDR = 0.540), reflecting the ambiguous ecological role of high sea urchin density — simultaneously controlling algae but inducing bioero-

Table 2. FCM risk profiles for four synthetic ecological scenarios. EH = ecosystem health, PR = pollution risk, TSR = thermal stress risk, HDR = habitat degradation risk, RP = recovery potential, HIR = heritage integrity risk. Higher EH and RP indicate healthier conditions; higher PR, TSR, HDR, and HIR indicate greater risk.

Scenario	**EH**	**PR**	**TSR**	**HDR**	**RP**	**HIR**
Healthy reef	0.986	0.451	0.511	0.068	0.959	0.476
Post-bleaching	0.339	0.616	0.556	0.819	0.144	0.697
COTS + blast fishing	0.081	0.692	0.500	0.962	0.034	0.855
Urchin barren	0.674	0.500	0.500	0.540	0.459	0.568

sion. These results confirm that the expert-defined causal structure captures the principal ecological risk gradients of reef systems.

5 Conclusion

In this paper, we presented a hybrid framework for automated coral reef ecological risk assessment that combines zero-shot visual reasoning with expert-guided causal inference. By employing Gemma 4-E4B as a zero-shot perception layer, this paper demonstrated that twenty-five quantitative ecological features can be reliably extracted from raw underwater survey images without any task-specific fine-tuning or manual annotation. These features are fed as clamped input activations into a Fuzzy Cognitive Map whose causal weight matrix encodes established marine ecological knowledge, yielding a six-dimensional risk profile that characterises ecosystem health, pollution risk, thermal stress, habitat degradation, recovery potential, and heritage integrity. Experiments on 50 images from the Coralscapes test split confirm rapid FCM convergence within 4–8 iterations across all tested inputs. Scenario analysis over four synthetic ecological conditions — ranging from a pristine reef to a COTS and blast fishing degradation event — validates the ecological consistency of the expert weight matrix, demonstrating that the causal structure correctly ranks risk severity across the full spectrum of reef health conditions. The entire pipeline runs on commodity single-GPU hardware with open-weight models and public datasets, making it immediately reproducible and deployable in resource-constrained monitoring contexts.

Several directions remain open for future investigation. First, the expert-defined weight matrix constitutes a fixed biological prior that does not adapt to the statistical distribution of the processed images. Incorporating online Hebbian learning strategies [21, 20] would allow the causal structure to self-refine as more observations become available, potentially improving risk discrimination in specific reef ecosystems. Second, the current evaluation is qualitative; future work should establish ground-truth risk labels, derived from expert-annotated surveys or existing reef health indices, to enable quantitative validation of the predicted risk profiles. Third, scaling the pipeline to the full Coralscapes dataset and extending it to temporal sequences of survey images would enable longitudinal reef health monitoring, a key requirement for conservation practice. Finally, the prompt engineering approach could be further refined through systematic ablation of feature vocabulary and output format constraints, exploring the trade-off between feature granularity and extraction reliability across different Gemma 4 model variants.

References

1. Alawode, B., et al.: AquaticCLIP: A Vision-Language Foundation Model for Underwater Scene Analysis. arXiv preprint arXiv:2502.01785 (2025)

2. Apostolopoulos, I.D., Groumpos, P.P.: Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence. Applied Sciences **13**(6), 3412 (2023)
3. Camastra, F., Ciaramella, A., Giovannelli, V., Lener, M., Rastelli, V., Staiano, A., Staiano, G., Starace, A.: A fuzzy decision support system for the environmental risk assessment of genetically modified organisms. In: Recent Advances of Neural Network Models and Applications: Proceedings of the 23rd Workshop of the Italian Neural Networks Society (SIREN), May 23-25, Vietri sul Mare, Salerno, Italy. pp. 241–249. Springer (2014)
4. Camastra, F., Ciaramella, A., Salvi, G., Sposato, S., Staiano, A.: On the interpretability of fuzzy knowledge base systems. PeerJ Computer Science **10**, e2558 (2024)
5. Di Nardo, E., Ciaramella, A., et al.: Advanced fuzzy relational neural network. In: WILF (2021)
6. Farabet, C., Lacombe, O.: Gemma 4: Byte for Byte, the Most Capable Open Models. `https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/` (2026), google DeepMind Blog, Accessed: April 2026
7. Felix, G., Nápoles, G., Falcon, R., Froelich, W., Vanhoof, K., Bello, R.: A Review on Methods and Software for Fuzzy Cognitive Maps. Artificial Intelligence Review **52**(3), 1707–1737 (2019)
8. Gemma Team: Gemma: Open Models Based on Gemini Research and Technology. arXiv preprint arXiv:2403.08295 (2024)
9. Google DeepMind Team: Gemma 4 Model Card. `https://ai.google.dev/gemma/docs/core/model_card_4` (2026), accessed: April 2026
10. Hobbs, B.F., et al.: Fuzzy Cognitive Mapping as a Tool to Define Management Objectives for Complex Ecosystems. Ecological Applications **12**(5), 1548–1565 (2002)
11. ICRI: 84% of the World's Coral Reefs Impacted in the Most Intense Global Coral Bleaching Event Ever. `https://icriforum.org/4gbe-2025/` (2025), international Coral Reef Initiative
12. IPCC: Special Report on Global Warming of 1.5°C. `https://www.ipcc.ch/sr15/` (2018), intergovernmental Panel on Climate Change
13. IUCN: Over 40% of Coral Species Face Extinction — IUCN Red List. `https://iucn.org/press-release/202411/over-40-coral-species-face-extinction-iucn-red-list` (2024), announced at COP29, November 2024
14. Knowlton, N., Brainard, R.E., Fisher, R., Moews, M., Plaisance, L., Caley, M.J.: Coral Reef Biodiversity. In: Life in the World's Oceans: Diversity, Distribution, and Abundance, pp. 65–78. Wiley-Blackwell (2010)
15. Kosko, B.: Fuzzy Cognitive Maps. International Journal of Man-Machine Studies **24**(1), 65–75 (1986). https://doi.org/10.1016/S0020-7373(86)80040-2
16. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. International journal of man-machine studies **7**(1), 1–13 (1975)
17. Maratea, A., Ciaramella, A., Santillo, M.: Fuzzy cognitive maps extraction from enriched tweets. In: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–8. IEEE (2022)
18. OpenAI: GPT-4 Technical Report. Tech. rep., OpenAI (2023), arXiv:2303.08774
19. Özesmi, S.L., Özesmi, U.: Ecological Models Based on People's Knowledge: A Multi-Step Fuzzy Cognitive Mapping Approach. Ecological Modelling **176**(1–2), 43–64 (2004)
20. Papageorgiou, E.I., Groumpos, P.P.: A Weight Adaptation Method for Fuzzy Cognitive Map Learning. Soft Computing **9**(11), 846–857 (2005)

21. Papageorgiou, E.I., Stylios, C.D., Groumpos, P.P.: Active Hebbian Learning Algorithm to Train Fuzzy Cognitive Maps. International Journal of Approximate Reasoning **37**(3), 219–249 (2004)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning (ICML) **139**, 8748–8763 (2021)
23. Sauder, J., et al.: The Coralscapes Dataset: Semantic Scene Understanding in Coral Reefs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2025), arXiv:2503.20000
24. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE transactions on systems, man, and cybernetics **15**(1), 116–132 (1985)
25. Zadeh, L.A.: Fuzzy sets. Information and control **8**(3), 338–353 (1965)
26. Zheng, Z., et al.: MarineGPT: Unlocking Secrets of Ocean to the Public. arXiv preprint arXiv:2310.13596 (2023)