This study explores generative AI visual language models for object detection in underwater ROV environments. Testing Gemini, ChatGPT, and other models across aquaculture, exploration, and monitoring scenarios revealed strong zero-shot performance without specialized underwater training. Google Gemini achieved the highest mAP (0.85), followed by ChatGPT (0.81), with others nearing 0.80. These results demonstrate the significant potential of visual language models in complex underwater environments. Future improvements through domain-specific fine-tuning and edge computing enhancement could further increase detection accuracy, offering promising directions for underwater robotics and intelligent ocean exploration.

Image Analysis of VLM Models

Gemini

A school of fish swims around a coral reef with marine debris including plastic bottles and and line.

ChATGPT

"large group of small, reflective possibly sardines, amchiaten, navigates lonstructure. Pieces of clear bottles and tangled mess the vibrael, tine, visilel, annbrant corals, Th aline aioon fiiltering rays, creale CHATPP's is t most detaiied.

Claude

Fish near colorful corals with plastic pollution h coral, and human-made trash undre ocean.

Underwater End

ROV Onboard Camera + Jetson Nano

Image Capture & Pre-processing

Cloud/Oshore End

RTX 3070 Server

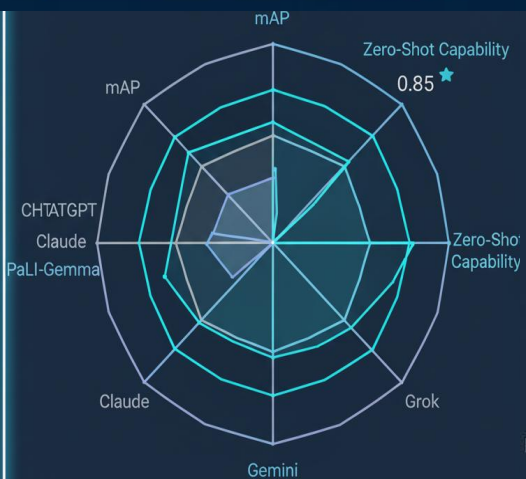
Multi-Model VLM Inference & Analysis

CHATGPT (GPT-4V), Anthropic Claude

DeepSeek-VL

Google PaLI-Gemma

xAI Grok



Generative AI Opportunities

Fish Species: Blue Tang

Coral Type: Brain Coral

Status: Healthy

Decision Feedback

Control Station: Decision Support & Feedback