Visual Question Answering for Underwater Scene Understanding

Master/Eng internship

General information

- **Duration :** 6 months (standard stipend). To start between February and March 2027.
- **Institute :** Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE), team Systèmes Intelligents de Perception
- **Location :** 45 rue des Saints-Pères, 75006, Paris
- **Supervision :** Ayoub Karine [1] , Camille Kurtz [2], Laurent Wendling [3]
- **Keywords :** computer vision, deep learning, underwater images, vision-language models
- **Application :** send an email with CV to ayoub.karine@u-pariscite.fr, camille.kurtz@u-pariscite.fr and laurent.wendling@u-pariscite.fr with subject "[VQA-UI]" (the application through message in Linkedin will not be considered)

Proposed topic

Context

Marine biodiversity monitoring is crucial for understanding and preserving marine ecosystems, including species identification, abundance estimation, taxonomic diversity and ecological interactions. Recent advances in Vision-Language Models (VLMs), particularly Visual Question Answering (VQA), offer promising opportunities to facilitate these tasks through natural language-based queries (Figure 1). Several underwater VQA datasets have recently emerged, such as NautData [1], UWBench [2], MarineEval [3] and CoralVQA [4]. However, they primarily address general underwater scene understanding or focus on specific ecosystems, rather than comprehensive marine biodiversity monitoring. Conversely, VLM4Bio [5] introduced a biodiversity-oriented VQA benchmark covering multiple biological groups, including fish, birds and butterflies, but it is not specifically designed for underwater imagery or marine environments. To the best of our knowledge, no existing benchmark combines comprehensive marine biodiversity understanding with underwater imagery. We therefore propose to develop a dedicated marine biodiversity VQA dataset by leveraging existing underwater vision and vision-language datasets.

Moreover, underwater images are affected by specific degradations, such as light attenuation, color distortion and light scattering by suspended particles, which challenge VLMs trained mostly on terrestrial images. We therefore aim to investigate the capabilities of general-purpose VLMs on marine biodiversity tasks and develop a specialized marine biodiversity VLM. In particular, we will explore the integration of domain-specific knowledge, including underwater imaging physics and marine biological knowledge, to improve model performance and robustness.

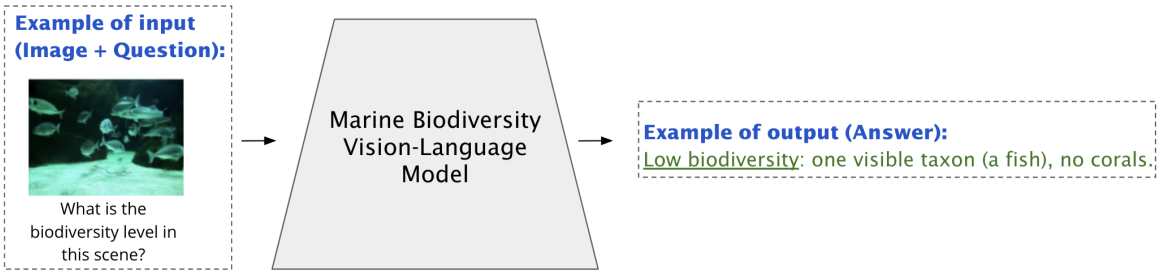


FIGURE 1 – Example of marine biodiversity VQA.

1. https://www.ayoub-karine.com/
2. https://www.camille-kurtz.com/
3. https://helios2.mi.parisdescartes.fr/ lwendlin/

Work to be done

▷ **Step 1 – Literature review :** Review recent vision-language datasets and models dedicated to visual question answering in underwater environments, as well as biodiversity-oriented VQA benchmarks.

▷ **Step 2 – Build a marine biodiversity VQA dataset :** Build the dataset from images and annotations of existing underwater datasets, enriched with marine biodiversity information (e.g., taxonomy). The construction will be inspired by VLM4Bio [5] and CoralVQA [4].

▷ **Step 3 – Test generic VLM on constructed dataset :** Evaluate generic VLMs (e.g., LLaVA, Qwen-VL, BLIP-2) on the constructed dataset in a zero-shot setting, then after parameter-efficient fine-tuning (e.g., LoRA) and full fine-tuning.

▷ **Step 4 – Integration of underwater knowledge in VLM :** Our research team has recently proposed two methods to integrate underwater knowledge, namely imaging physics [6] and visual quality [7], into vision models. Building on these developments, we aim to develop a specialized marine biodiversity VLM integrating underwater imaging physics and marine biological knowledge, and to evaluate whether it improves performance and robustness compared with generic VLMs and with existing underwater-specific VLMs (e.g., Nautilus [1], AquaticCLIP [8]).

Desired background for the candidate

We are looking for a Master 2 or final-year MSc student, or an engineering school student, in computer science. The ideal candidate should have knowledge in deep learning, computer vision and Python programming, and an interest in underwater computer vision.

Bibliography

[1] Wei Xu, Cheng Wang, Dingkang Liang, Zongchuang Zhao, Xingyu Jiang, Peng Zhang et Xiang Bai. "NAUTILUS : a large multimodal model for underwater scene understanding". In : *Advances in Neural Information Processing Systems* 38 (2026), p. 25079-25105.

[2] Da Zhang, Chenggang Rong, Bingyu Li, Feiyu Wang, Zhiyuan Zhao, Junyu Gao et Xuelong Li. "Uwbench : A comprehensive vision-language benchmark for underwater understanding". In : *arXiv preprint arXiv :2510.18262* (2025).

[3] Yuk Kwan Wong, Tuan-An To, Jipeng Zhang, Ziqiang Zheng et Sai-Kit Yeung. "Marineeval : assessing the marine intelligence of vision-language models". In : *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2026, p. 1459-1470.

[4] Hongyong Han, Wei Wang, Gaowei Zhang, Mingjie Li et Yi Wang. "CoralVQA : A Large-Scale Visual Question Answering Dataset for Coral Reef Image Understanding". In : *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2026.

[5] M Maruf, Arka Daw, Kazi S Mehrab, Harish B Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P Balhoff, Yasin Bakiş, Bahadir Altintas et al. "VLM4Bio : A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images". In : *The Thirty-Eighth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track* 37 (2024), p. 131035-131071.

[6] Alex Delaveau, Ayoub Karine, Camille Kurtz et Laurent Wendling. "PGDD-UW : Physics-Guided Dataset Distillation for Efficient Underwater Classification". In : *Proceedings of the 22nd International Conference on Computer Vision Theory and Applications, VISAPP 2027, Valetta, Malta* (2027).

[7] Amine Aissaoui, Alex Delaveau, Ayoub Karine, Camille Kurtz et Laurent Wendling. "EQUIS : Efficient Underwater Instance Segmentation via Quality-Guided Foundation Model Distillation". In : *Proceedings of the 22nd International Conference on Computer Vision Theory and Applications, VISAPP 2027, Valetta, Malta* (2027).

[8] Basit Alawode, Iyyakutti Iyappan Ganapathi, Sajid Javed, Mohammed Bennamoun et Arif Mahmood. "AquaticCLIP : A Vision-Language Foundation Model and Dataset for Underwater Scene Analysis". In : *IEEE Transactions on Neural Networks and Learning Systems* (2026).