

Evaluating Vision-Language Models Underwater: A Benchmark Study Toward Domain-Specific VLM Design

Aditya Tejawani

Department of Electrical and Computer Engineering

University of Maryland

College Park, MD, USA

atejwani@umd.edu

Abstract—Vision-language models (VLMs) such as CLIP [1] perform well on land-based imagery but struggle in underwater environments due to color distortion, turbidity, and noise [2]. A benchmark is performed to assess CLIP ViT-B/32, CLIP ViT-L/14, and DINOv2 performance on the AQUA20 dataset [3] of 6,559 testing images across 20 marine species. CLIP achieves 42.99% and 54.08% accuracy, while DINOv2 [4] reaches 92.45% using only visual features. Two main limitations are identified: semantic ambiguity between labels and degradation of visual features under poor imaging conditions. These results suggest that language-supervised models are poorly suited for underwater recognition without domain-specific training.

Index Terms—vision-language models, underwater imaging, marine species classification, CLIP, DINOv2

I. INTRODUCTION

Autonomous Underwater Vehicles (AUVs) are being increasingly utilized for biodiversity assessment, aquaculture management, and sea-life monitoring [5]. However obtaining useful data from marine images presents challenges that do not exist on land. Light attenuation, color distortion, turbidity, and low resolution degrade image quality in ways that standard computer vision models are not trained to handle [2].

Vision language models (VLMs) such as CLIP [1] have demonstrated strong performance in zero-shot classification when tested on land-based benchmarks. However, CLIP performance declines significantly under domain shift, particularly in remote sensing applications [6]. Underwater imagery represents an even more extreme domain shift where the pre-training data for these models is comprised almost entirely of land-based images with little marine representation.

Despite the growing use of VLMs in robotic perception, their underwater performance remains poorly understood, limiting their viability in safety-critical marine missions. This benchmark shows that CLIP often confuses classes such as

“fish” and “rayfish” due to language bias, and investigates how these failures can inform the design of more effective domain-specific VLMs.

A benchmark test is performed comparing CLIP ViT-B/32, CLIP ViT-L/14, and DINOv2 [4] on the AQUA20 dataset [3], comprising of 6,559 underwater training images and 20 different marine species. This comparison produced three key insights:

- A benchmark of both CLIP models and DINOv2 on marine species classification revealed a 38.37% accuracy gap between language-supervised models and unlabeled visual models.
- The two main sources of error are from semantic ambiguity between visually similar species, and feature degradation under underwater imaging conditions.
- There is evidence that CLIP failures are driven by language bias rather than image degradation, supported by DINOv2 achieving 93.4% accuracy on fish images where CLIP achieves only 2.59%

II. RELATED WORK

A. Vision-Language Models

Vision-Language Models (VLMs) are a class of machine learning systems designed to align visual and textual information in the same representation space [7]. These models use large datasets of image-text pairs to learn the correspondence between natural language descriptions and their visual representation. Since both modalities are embedded in the same space, VLMs enable a large array of machine learning tasks like image classification, visual question answering, and captioning, often without domain-specific tuning.

A commonly used VLM is CLIP (Contrastive Language Image Pretraining) [1], which utilizes an architecture consisting of an image encoder and a text encoder. During training,

CLIP tries to maximize the cosine similarity between the text embedding and the matched image while minimizing the similarity between mismatched pairs. At inference time, CLIP enables zero-shot classification by comparing an image embedding with the embeddings of the textual prompts representing a candidate’s class. The benefit of CLIP is its ability to learn semantic representations without supervision for specific tasks. There is no longer a need to label the datasets in many applications.

Despite their impressive capabilities, VLMs exhibit many limitations. One key limitation is a sensitivity to distributional shift and image transformations, such as occlusion, compositional rearrangement, and repeated object occurrences [8]. Additionally, VLMs often fail to correctly associate attributes with multiple objects in a scene, a phenomenon known as the binding problem [9]. For example, the model would struggle to distinguish between “a blue fish swimming above coral” and “coral with a blue fish swimming underneath”. This indicates that VLMs have a lack of fine-grained relational reasoning. Additionally, biases in training data can propagate into model predictions, which leads to reliability concerns in specialized domains. Addressing these limitations is an active area of machine learning research.

B. Self-Supervised Visual Models

Unlike CLIP, DINOv2 [4] is a self-supervised model that does not rely on paired text-image data. DINOv2 is built using self-distillation training, where a student network is trained to match the output of its teacher network through a series of augmentations on the same image [4]. These augmentations include random cropping, color modifications, resolution changes, which teach the model to recognize the object regardless of how it appears. This removes the need for manual annotation and makes the approach practical at scale.

DINOv2 is built on Vision Transformers [4], which are well-suited for capturing fine-grained spatial relationships in images. Additionally, DINOv2 implements several training refinements, including feature normalization, sharpening of the teacher outputs, and exponential-moving average updates to stabilize the teacher-student training process and prevent representation collapse. Data curation is applied to ensure that the model is exposed to diverse and high-quality images during pretraining.

The resulting visual features transfer well across various computer vision tasks without fine-tuning. Although DINOv2 has no language component, it can serve as an encoder in VLM pipelines. When implemented in a VLM, it is used as a language-free baseline to determine whether CLIP failures

in underwater environments stem from poor visual encoding or from the language-vision alignment mechanism itself.

C. Out-of-Distribution Performance

While CLIP-style VLMs demonstrate remarkable zero-shot capabilities on natural images, their performance suffers when applied to out-of-distribution (OOD) data. This trend has been well-documented in specialized fields such as medical imaging, where CLIP struggles with fine-grained medical features like tumors [10], and remote sensing, where the aerial perspective of satellite imagery differs significantly from the images the model is trained on [11]. In fact, VLM frameworks such as MMLGNet demonstrates that CLIP’s visual data processing structure must often be augmented with specialized encoders and multimodal alignment strategies to achieve the necessary semantic performance in domain specific applications [11]. The underwater environment similarly presents a set of domain-specific challenges, including light attenuation, scattering, and turbidity, which pose difficulties for VLMs trained on terrestrial data. By discussing underwater perception in the larger context of OOD performance, we show that the problem is a general limitation of current foundation models, not just a problem with one dataset.

D. Underwater Computer Vision

Computer vision underwater is much harder than it is in terrestrial environments because the water distorts with what the camera captures [2]. Light absorption and scattering cause images to appear blurry and low contrast, meaning that marine species in the images will rarely look the way they would on the surface. Waves introduce additional noise that further degrades image quality. Additionally, marine life moves unpredictably and frequently blends into natural features like coral, making it tough for computer vision models to accurately extract useful information [5].

Approaches to marine species classification have changed considerably over time. Earlier methods focused on manually designing features, like color patterns, textures, or shapes, and then applying algorithms like SVMs or k-NN to classify them. More recently, Convolutional Neural Networks, have become the most commonly applied approach [3]. These neural networks learn features automatically from data instead of relying on user designed features. Additionally, augmentation has been used during preprocessing steps to further improve classification models by training the model on a more diverse dataset.

III. DATASET AND MODELS

A. AQUA20 Dataset

All experiments use the AQUA20 dataset [3], a dataset of underwater images designed for marine species classification. The full dataset contains 8,171 images spread across 20 classes, each corresponding to a specific marine species. For this study, all benchmark tests are performed on the testing partition, which consists of 6,559 images. The dataset is heavily imbalanced, with some species significantly overrepresented compared to others. For example, the “fish” class has approximately 2,200 images, while rare classes such as “octopus” contain only about 20 images. This imbalance makes the benchmark more realistic, since models must perform well even when certain species have very little data to draw from. Furthermore, the dataset reflects underwater conditions, including variations in lighting, visibility, and color distortion making AQUA20 a suitable dataset for assessing underwater computer vision models.

Fig. 1 illustrates the class imbalance across all 20 species. The dataset is dominated by common species such as fish (2,200 images) and coral (1,562 images), while rare classes such as octopus contain as few as 20 images. A logarithmic scale is used to visualize the smallest classes alongside the largest.

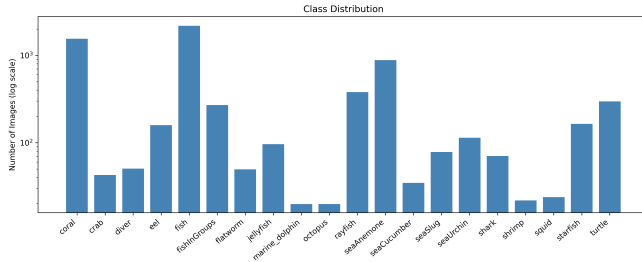


Fig. 1: Class distribution of the AQUA20 dataset across 20 marine species on a logarithmic scale.

B. Models

The benchmark tests 3 models: CLIP ViT-B/32, CLIP ViT-L/14 [1], and DINOv2 [4]. The two CLIP models classify images by comparing image embeddings against text embeddings for each class. Text prompts of the form “a photo of a [species]” are created for all 20 marine species. The model predicts the class whose text embedding is most similar to the image embedding. The primary difference between the two CLIP variants is model size. ViT-L/14 is significantly larger than ViT-B/32, which allows us to investigate whether increasing model size improves performance in underwater conditions.

DINOv2 [4] is used in a different way, serving as a visual feature extractor rather than a zero-shot classifier. Each image is passed through the model to obtain a token embedding, and a k-Nearest Neighbors classifier with $k = 5$ is used to assign labels based on the closest embeddings in the feature space. An 80/20 train-test split is applied for this evaluation. This benchmark test enables us to assess how well self-supervised visual features perform compared to language-supervised approaches like CLIP, and more importantly, to determine whether observed failures are due to limitations in visual understanding or in the alignment between vision and caption.

It should be noted that the comparison between models is not identical in terms of experimental conditions. CLIP is evaluated fully zero-shot with no exposure to the AQUA20 dataset, while DINOv2 is provided labeled training examples via the 80/20 data split. This is intentional, since the goal is not to compare models under identical conditions, but to evaluate the best capability each paradigm can offer. CLIP represents the strongest language-supervised approach, while DINOv2 represents the strongest visual feature approach with minimal supervision.

IV. EXPERIMENTS

A. Evaluation Metrics

Two metrics are used to evaluate model performance. The first is overall classification accuracy, which measures the percentage of images where the model correctly identifies the species. Accuracy is the most straightforward way to compare how well each model performs, but it can produce misleading results. Since AQUA20 is heavily imbalanced, a model could show high accuracy overall while having low accuracy on a rare species. For this reason, the per-class accuracy of all 20 species was also collected.

ROC curves and AUC are also reported for both CLIP models, capturing how well each model ranks the correct class regardless of threshold. AUC is omitted for DINOv2 since k-NN does not produce calibrated probability scores. This is particularly informative in cases where a model assigns reasonably high probability to the correct class but is consistently outranked by a competing class. For example, with the fish class, CLIP achieves an AUC of 0.87 despite only 2.59% classification accuracy, indicating that the failure is not due to the correct class being ranked poorly overall, but rather due to a specific class consistently scoring higher.

In addition to accuracy and AUC, precision, recall, and F1 score are reported for each model. Given the class imbalance in AQUA20, overall accuracy alone is an insufficient metric. Since fish and coral together account for the majority of

images in the dataset, a model could achieve a high overall accuracy while completely failing on rare classes such as octopus and shrimp. This failure would be largely unrepresented in the accuracy number. The F1 metric addresses this by computing F1 separately for each class and averaging them equally, ensuring that a failure on octopus penalizes the summary metric just as much as a failure on fish. For a given class, precision measures the fraction of positive predictions that are correct, and recall measures the fraction of true positives that are detected. F1 is the harmonic mean, defined by the equation:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

Table I reports precision, recall, and F1. DINOv2 achieves balanced scores of 91.25% and 91.93%, while CLIP ViT-L/14 shows a recall of 73.49% against a precision of 61.90%, indicating excess false positives. The F1 scores confirm the accuracy gap is not an artifact of class imbalance.

TABLE I: Overall Performance on AQUA20

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
CLIP ViT-B/32	42.99	50.83	53.84	42.75
CLIP ViT-L/14	54.08	61.90	73.49	59.80
DINOv2 ViT-B/14	92.45	91.25	91.93	91.50

Fig. 2 shows the per-class ROC curves for both CLIP variants across all 20 marine species.

ROC Curves – CLIP Small vs CLIP Large

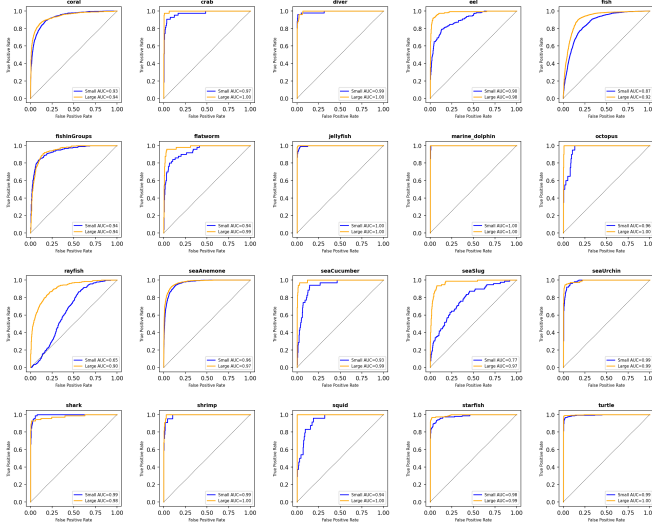


Fig. 2: Per-class ROC curves for CLIP ViT-B/32 (blue) and CLIP ViT-L/14 (orange) across all 20 marine species on the AQUA20 dataset.

B. Results

Table I summarizes overall accuracy. DINOv2 achieves 92.45%, compared to 54.08% for CLIP ViT-L/14 and 42.99% for CLIP ViT-B/32, displaying a 38.37% gap showing that zero-shot language-supervised methods struggle in underwater environments. Table II shows per-class results. DINOv2 achieves 100% on seven classes including jellyfish, shark, turtle, and squid, and exceeds 90% on fourteen of twenty classes. CLIP ViT-L/14 performs well on visually distinctive species such as diver (100%) and jellyfish (96.9%), but struggles on fish (10.6%) and fishInGroups (11.0%). The fish and fishInGroups classes are the most problematic for both CLIP variants, despite fish having the largest image count, both models fail to classify it reliably, suggesting language bias rather than data scarcity as the root cause. Fig. 3a shows the per-class accuracy comparison visually.

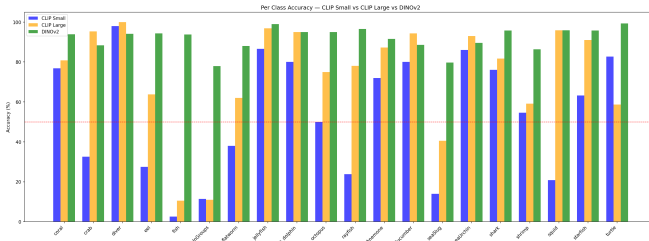
TABLE II: Per-Class Accuracy (%) on AQUA20

Species	CLIP-S	CLIP-L	DINOv2
coral	76.9	80.9	93.6
crab	32.6	95.4	66.7
diver	98.0	100.0	90.0
eel	27.5	63.8	93.8
fish	2.6	10.6	93.4
fishInGroups	11.4	11.0	74.1
flatworm	38.0	62.0	70.0
jellyfish	86.6	96.9	100.0
marine_dolphin	80.0	95.0	100.0
octopus	50.0	75.0	100.0
rayfish	23.8	78.0	93.5
seaAnemone	72.0	87.3	93.3
seaCucumber	80.0	94.3	100.0
seaSlug	13.9	40.5	62.5
seaUrchin	86.1	93.0	91.3
shark	76.1	81.7	100.0
shrimp	54.6	59.1	75.0
squid	20.8	95.8	100.0
starfish	63.3	91.6	97.0
turtle	82.7	99.3	100.0
Overall	43.0	54.1	92.5

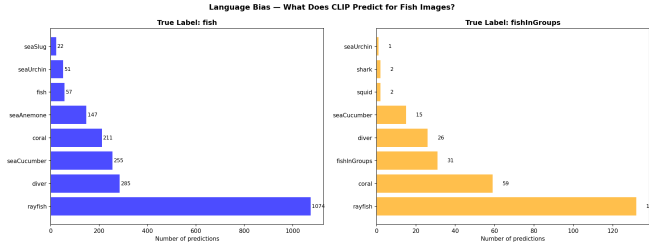
V. ANALYSIS

A. Limitation 1: Language-Semantic Ambiguity

One major source of error with the CLIP models is language-semantic ambiguity, where species that appear identical are confused due to overlapping or imprecise textual descriptions. A clear example of this occurs with the “fish” and “rayfish” classes. Despite substantial visual differences, both classes share the word “fish” in their text prompts, which biases the language-supervised model toward systematic misclassification. Analysis of all 2,200 fish images shows that CLIP ViT-B/32 predicts rayfish in 1,074 cases, accounting



(a) Per-class accuracy comparison across models.



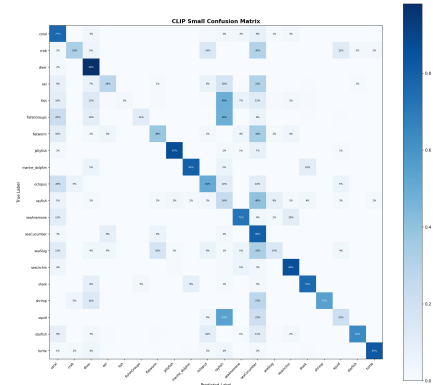
(b) CLIP prediction bias for fish classes.

Fig. 3: Model performance and failure case analysis on AQUA20.

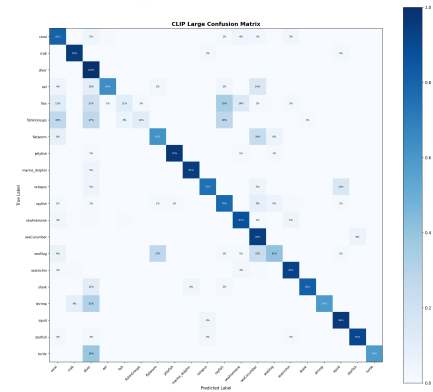
for 48.8% of all fish misclassifications. This pattern is consistent across both CLIP variants, as shown in Fig. 3b. This suggests that the failure is likely not caused by poor visual encoding, but rather stems from the way CLIP aligns textual embeddings with visual features [6]. CLIP uses Byte-Pair Encoding (BPE) for subword segmentation. In CLIP, rare or domain-specific words like ‘rayfish’ are likely not present as single tokens in the models dictionary. Instead, the tokenizer decomposes them into more frequent subwords, such as ‘ray’ and ‘fish’. Consequently, the model’s text embedding for ‘rayfish’ becomes heavily biased toward the generic ‘fish’ component, leading to the high misclassification rate observed in the results [1].

B. Limitation 2: Visual Degradation

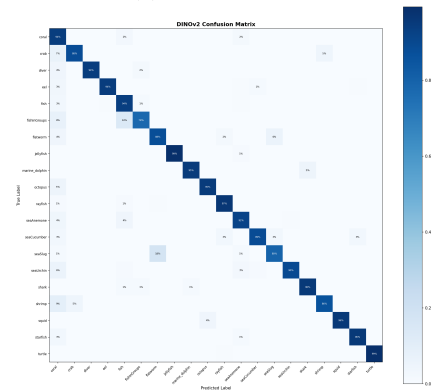
Another source of error stems from visual degradation under underwater conditions [2]. CLIP models frequently misclassify coral as diver, particularly in images with strong color distortion, low contrast, or dark lighting. These degradations obscure the distinguishing features in an image, making it difficult for the model to differentiate between objects with similar shapes or patterns. Fig. 5 shows examples where CLIP fails while DINOv2 is successful, illustrating how environmental degradations remove the visual cues that CLIP models rely on. In contrast, DINOv2, which relies solely on visual feature representations, is less sensitive to these distortions.



(a) CLIP ViT-B/32



(b) CLIP ViT-L/14



(c) DINOv2

Fig. 4: Confusion matrices for all three models on the AQUA20 dataset.

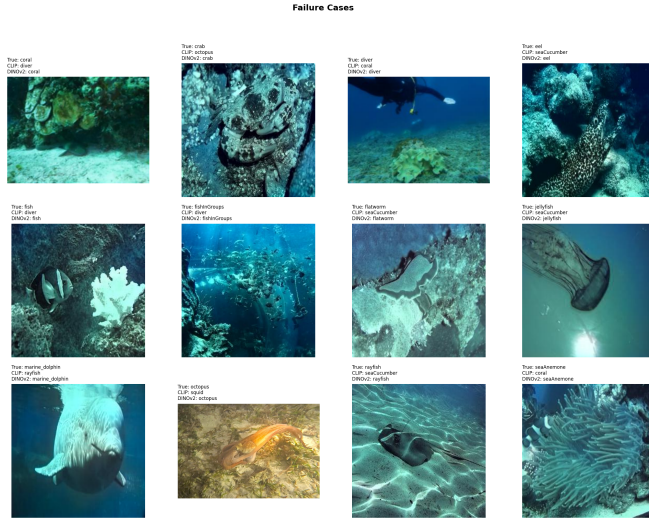


Fig. 5: Examples of images where CLIP fails but DINOv2 succeeds. Underwater colour distortion and low contrast remove visual cues that CLIP relies upon for classification.

C. Effect of Model Size

Scaling from CLIP ViT-B/32 to ViT-L/14 improves overall accuracy by 11.09%, with notable accuracy increases on species such as crab (+62.8%) and squid (+75.0%). However, increasing the model size from CLIP ViT-B/32 to ViT-L/14 does not eliminate failures due to language-semantic ambiguity. While the larger model is better at extracting general visual information, increasing model size predominantly helps with feature recognition rather than resolving biases introduced by overlapping textual prompts. These results suggest that language-driven failures in underwater classification cannot be overcome by using a larger model, and that architectural changes or domain-specific training are needed to address the root cause of error.

VI. CONCLUSION

A benchmark was conducted comparing CLIP ViT-B/32, CLIP ViT-L/14, and DINOv2 on the AQUA20 underwater species classification dataset. DINOv2 outperformed both CLIP models by a wide margin, achieving 92.45% accuracy compared to 54.08% for CLIP ViT-L/14 and 42.99% for CLIP ViT-B/32. Two areas of failure were identified in CLIP: language-semantic ambiguity between overlapping class labels, driven by BPE tokenization separating domain-specific terms into generic subwords, and visual degradation under underwater imaging conditions. Scaling to a larger model improved general recognition ability but did not resolve language-driven failures, confirming that the root cause of

error is related to model architecture as opposed to model capacity.

This study has two main limitations. The comparison is intentionally asymmetric, since DINOv2 uses labeled training examples while CLIP is fully zero-shot, so the gap reflects the best of each paradigm rather than an identical comparison. Results are also limited to a single dataset. Future work should explore underwater preprocessing and marine-specific fine-tuning for CLIP, and evaluate DINOv2 as a visual backbone in a full VLM pipeline for reliable marine perception. This work serves as a diagnostic benchmark for understanding the limitations of current vision-language models under underwater domain shift and provides a foundation for the development of domain-specific vision-language systems for marine perception.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [2] H. Lu, Y. Li, X. Xu, J. Li, Z. Liu, X. Li, J. Yang, and S. Serikawa, "Underwater image enhancement method using weighted guided trigonometric filtering and artificial light correction," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 504–516, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1047320316300335>
- [3] T. R. Fuad, S. Ahmed, and S. Ivan, "Aqua20: A benchmark dataset for underwater species classification under challenging conditions," *Arabian Journal for Science and Engineering*, Mar. 2026. [Online]. Available: <http://dx.doi.org/10.1007/s13369-026-11231-z>
- [4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024. [Online]. Available: <https://arxiv.org/abs/2304.07193>
- [5] A. Monterroso Muñoz, M.-J. Moron-Fernández, D. Cascado-Caballero, F. Diaz-del Rio, and P. Real, "Autonomous underwater vehicles: Identifying critical issues and future perspectives in image acquisition," *Sensors*, vol. 23, no. 10, 2023. [Online]. Available: <https://www.mdpi.com/1424-8220/23/10/4986>
- [6] F. Sammani and N. Deligiannis, "Interpreting and analysing clip's zero-shot image classification via mutual knowledge," 2024. [Online]. Available: <https://arxiv.org/abs/2410.13016>
- [7] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, M. Ibrahim, M. Hall, Y. Xiong, J. Lebensold, C. Ross, S. Jayakumar, C. Guo, D. Bouchacourt, H. Al-Tahan, K. Padthe, V. Sharma, H. Xu, X. E. Tan, M. Richards, S. Lavoie, P. Astolfi, R. A. Hemmat, J. Chen, K. Tirumala, R. Assouel, M. Moayeri, A. Talattof, K. Chaudhuri, Z. Liu, X. Chen, Q. Garrido, K. Ullrich, A. Agrawal, K. Saenko, A. Celikyilmaz, and V. Chandra, "An introduction to vision-language modeling," 2024. [Online]. Available: <https://arxiv.org/abs/2405.17247>
- [8] A. M. Anis, H. Ali, and S. Sarfraz, "On the limitations of vision-language models in understanding image transforms," 2025. [Online]. Available: <https://arxiv.org/abs/2503.09837>

- [9] D. Campbell, S. Rane, T. Giallanza, N. De Sabbata, K. Ghods, A. Joshi, A. Ku, S. M. Frankland, T. L. Griffiths, J. D. Cohen, and T. Webb, "Understanding the limits of vision language models through the lens of the binding problem," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 113 436–113 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/cdcc6d47c1627350014a3076112ab824-Paper-Conference.pdf
- [10] Z. Zhao, Y. Liu, H. Wu, M. Wang, Y. Li, S. Wang, L. Teng, D. Liu, Z. Cui, Q. Wang, and D. Shen, "Clip in medical imaging: A survey," 2025. [Online]. Available: <https://arxiv.org/abs/2312.07353>
- [11] A. Chaudhary, S. Barman, M. Singha, A. Jha, G. Mishra, and B. Banerjee, "Mmlgnet: Cross-modal alignment of remote sensing data using clip," 2026. [Online]. Available: <https://arxiv.org/abs/2601.08420>