ORCA: Object Recognition and Comprehension for Archiving Marine Species

Supplementary Material

This supplementary material contains additional details regarding the dataset statistic (Section A), image captioning fine-tuning (Section B), and negative caption analysis (Section C).

A. Dataset Statistics

This section provides a comprehensive overview of the statistical properties of the ORCA dataset. Figure 1 summarizes the number of unique taxonomic labels across hierarchical levels, while Figures 3 and 2 visualize the structural composition and sample distribution within the dataset.

As shown in Figure 1, the ORCA dataset captures a wide spectrum of biological diversity, encompassing two kingdoms and 478 distinct species. The hierarchical distribution in Figure 3 indicates that most specimens belong to the phylum *Chordata*, with substantial representation from classes such as *Aves* (birds) and *Mammalia* (mammals). Other phyla, including *Mollusca*, *Arthropoda*, and *Cnidaria*, are present to a lesser extent, reflecting the natural sampling bias toward more frequently imaged taxa in ecological and wildlife datasets. This taxonomic heterogeneity underscores both the ecological breadth and inherent imbalance characteristic of large-scale biological image collections.

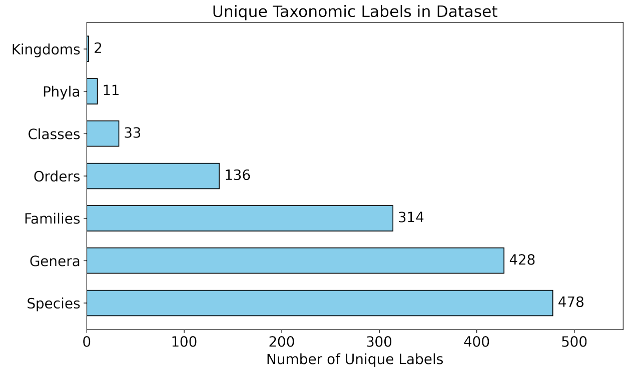

Figure 1. Number of unique labels at each taxonomic level.

Figure 2 illustrates the distribution of samples across species, revealing a pronounced long-tailed pattern. On average, each species label contains 63.01 images. The dataset exhibits substantial variation in sample counts, from single-instance records of rare species to a maximum of 8,004 samples labeled as "fish." The aggregated "fish" category reflects practical annotation conventions where specimens of uncertain or undefined identity are grouped under a general label. This natural imbalance poses meaningful challenges for modeling and provides a realistic testbed for evaluating algorithms under imbalanced data scenarios.

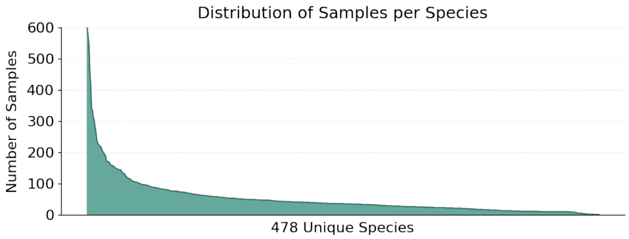

Figure 2. Distribution of sample counts per species in ORCA. The maximum count is clipped to 600 for visualization.

B. Image Captioning Fine-Tuning

We fine-tune MiniGPT-4 on ORCA to assess the influence of our dataset on image captioning performance. As shown in Table 1, fine-tuning on ORCA yields consistent improvements across nearly all evaluation metrics, demonstrating the dataset's effectiveness in enhancing the model's alignment between visual content and textual descriptions. The fine-tuned model achieves higher scores on both semantic and n-gram–based metrics, suggesting that exposure to our dataset enables the generation of captions that are more contextually relevant and descriptively rich.

It is worth noting that the METEOR score remains unchanged. This can be attributed to our fine-tuning strategy, where only the linear projection layer component of MiniGPT-4 was optimized while the language model remained frozen. Consequently, the semantic vocabulary of the model did not expand beyond the limitations of the original MiniGPT-4 language model, resulting in a stable METEOR score. Nevertheless, fine-tuning with the ORCA dataset allows the model to produce captions that are more accurate and semantically appropriate, aligning more closely with human-authored descriptions. This improvement is reflected in the higher BLEU-4, ROUGE, and CIDEr scores, indicating enhanced ability to convey fine-grained and distinctive visual details.

C. Negative Caption Analysis

We employ GPT-4 to generate initial image captions, which are subsequently reviewed and refined by marine biologists. Captions identified by the experts as *negative* (incorrect or

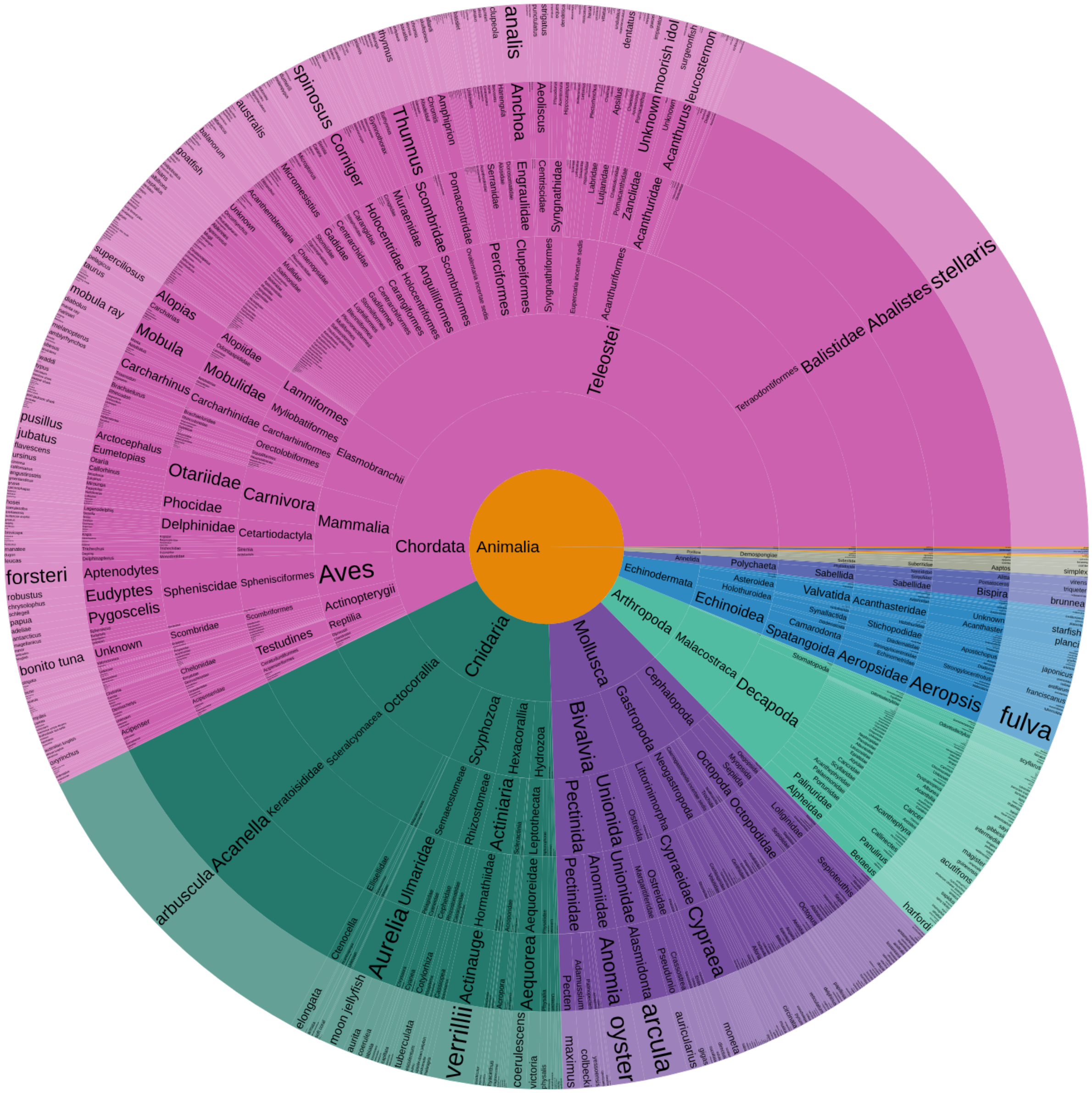

Figure 3. Hierarchical distribution of taxonomic labels in ORCA. Each segment represents a unique label at a given taxonomic rank.

Method	CLIPScore↑	RefCLIPScore↑	CIDEr↑	BLUE-4↑	METEOR↑	ROUGE↑
Vanilla	74.48	73.43	5.72	7.18	16.90	28.03
Fine-tuned	77.96+3.48	77.51+4.08	17.36+11.64	14.79+7.61	16.90+0	33.71+5.68

Table 1. Results of fine-tuning MiniGPT-4 on ORCA.

misleading) are retained to facilitate a detailed examination of captioning errors. A systematic analysis is conducted on these erroneous outputs, categorizing them into 11 distinct types based on their semantic and compositional characteristics. Representative examples and corresponding statistical distributions for each error type are provided in Table 2.

Object-related error. Our analysis indicates that a substantial proportion of captioning errors stems from object-related issues, particularly those involving fine-grained object classification. This pattern suggests that the model exhibits limited capacity to identify marine species. Additional errors involving nonexistent objects or inaccurate background contexts

Properties		Example	Number
Object	Classification	This is a yellow fish. *vs.* This is a yellow coral.	6,875
	Background	The turtle is in the ocean. *vs.* The turtle is in the sky.	1,343
	Unexisting	The shark has a long tail. (There is no tail in the image.)	3,264
Relation	Spatial	This fish is under the coral. *vs.* This fish is on the coral.	816
	Action	The penguin is walking. *vs.* The penguin is sitting.	938
Attribute	Size	The shark is large. *vs.* The shark is small.	271
	Color	This is a yellow fish. *vs.* This is a blue fish.	2,031
	Shape	This is a oval seashell. *vs.* This is a triangle seashell.	312
	Texture	The seashell is smooth. *vs.* The seashell is rough.	321
	Material	The fish is probably made of plastic.	316
	Counting	There are three penguins *vs.* There are four penguins.	831

Table 2. The detailed explanations of the constructed 11 error categories and corresponding data statistics.

point to tendencies toward hallucination and contextual misinterpretation, reflecting an overreliance on dataset priors rather than on visual evidence.

Relation-related error. Although relation-based errors, which concern spatial or action-level inconsistencies, occur less frequently, they expose deficiencies in the model's reasoning over inter-object relationships.

Attribute-related error. Within the attribute-related error categories, color misclassification emerges as the most prevalent, underscoring the model's sensitivity to variations in illumination, shading, and surface textures. Collectively, these findings highlight the need for improved visual grounding, compositional reasoning, and quantitative perceptual mechanisms to enhance both the semantic precision and contextual reliability of image captioning systems.