FISHNET 2.0: A COMPREHENSIVE MULTIMODAL BENCHMARK FOR LONG-TAILED FINE-GRAINED FISH UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have demonstrated impressive cross-domain capabilities; however, their proficiency in specialized scientific domains such as marine biology remains underexplored. To systematically study this gap, we introduce FishNet2.0, a large-scale multimodal benchmark for long-tailed fine-grained fish understanding, comprising 17,393 species with over 700,000 dense key-part annotations and more than 400,000 curated question-answer pairs. Beyond traditional recognition, FishNet2.0 integrates structured morphological descriptions and dense part-level localizations, enabling comprehensive evaluation across classification, detection, and grounded reasoning tasks. Extensive benchmarking of state-of-the-art MLLMs reveals severe degradation under real-world long-tailed distributions, with top-performing open-source models achieving below 40% accuracy. To diagnose the underlying causes, we propose a controlled evaluation framework that disentangles domain-specific knowledge from pure visual perception. Our analysis shows that performance failures stem primarily from insufficient domain knowledge rather than perceptual limitations. Importantly, we demonstrate that leveraging structured expert annotations, particularly explicit morphological descriptions and keypoint guidance, substantially improves zero-shot recognition performance. This establishes FishNet2.0 not only as a rigorous benchmark but also as a testbed for developing domain-aware multimodal reasoning strategies. Data and code will be released upon acceptance.

1 INTRODUCTION

Healthy aquatic ecosystems are fundamental to global biodiversity, food security, and sustainable development Selig et al. (2019); Basurto et al. (2025); Barbier (2017). Effective conservation and management efforts rely on accurate documentation and monitoring of aquatic species, among which fish constitute a major component of marine biodiversity. However, fish recognition is inherently challenging: many species exhibit subtle morphological differences that require *fine-grained* discrimination, while real-world observations follow highly *long-tailed* distributions, with numerous rare species appearing infrequently. This combination of fine-grained visual ambiguity and severe class imbalance makes automated fish understanding a particularly demanding yet crucial problem for modern vision systems.

Recent advances in Multimodal Large Language Models (MLLMs) Zhu et al. (2023); Bai et al. (2025); Wang et al. (2025); AI (2024); OpenAI (2024b) have demonstrated remarkable generalist capabilities across a wide range of vision-language tasks, including recognition Peng et al. (2024); Gan et al. (2024), segmentation Huang et al. (2025); Wei et al. (2025), and reasoning Yang et al. (2025); Shao et al. (2024); Guo et al. (2025); Tong et al. (2025; 2026). However, it remains unclear whether such generalist proficiency translates to the expert-level performance required in specialized scientific domains, e.g., fish. This raises two critical questions: 1) *Do current MLLMs truly possess the specialized knowledge and visual discrimination capacity required for comprehensive fine-grained fish understanding?* 2) *If not, what factors primarily limit their performance?*

To systematically address these questions, we introduce **FishNet2.0**, a large-scale, comprehensive, and multimodal benchmark for long-tailed fine-grained fish understanding. Unlike prior fish recog-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

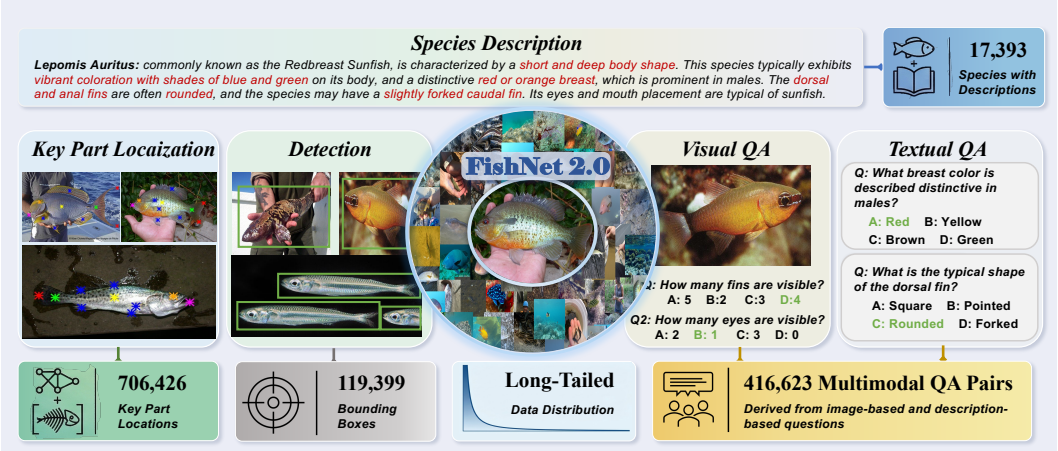

Figure 1: Overview of FishNet2.0, a comprehensive benchmark for long-tailed fine-grained fish understanding. It provides rich multimodal annotations, pairing each species with expert-level morphological descriptions and dense localizations, which enables rigorous diagnostic analysis of MLLMs.

nition datasets Chastine (2021); Conservancy (2017); Georgiou (2021); Nadarajan et al. (2009); Francescangeli & Jacopo (2023); Gilby et al. (2017); Boulais et al. (2021); iNaturalist (2021); Katija et al. (2022); Zhuang et al. (2020); Khan et al. (2023), FishNet2.0 provides structured, expert-level annotations that explicitly describe morphological traits, including species-specific shape, coloration, and distinguishing characteristics. In addition, we annotate fine-grained part locations, including the eye, mouth, pectoral/pelvic/anal fins, body center, and tail boundaries, enabling precise morphological reasoning. These multimodal annotations allow FishNet2.0 to serve not only as a benchmark for fine-grained recognition, but also as a diagnostic framework for understanding how visual and domain-specific knowledge contribute to model performance. In total, FishNet2.0 comprises 99,556 images spanning 17,393 fish species, enriched with 706,426 keypoint annotations, 119,399 bounding boxes, 17,393 curated textual descriptions, and 418,623 question-answer pairs. The benchmark supports a diverse set of tasks, including fine-grained classification, detection, part localization, and multimodal question answering.

Benchmarking several state-of-the-art MLLMs on FishNet2.0, we observe that despite their strong generalist performance on conventional vision-language tasks, they degrade substantially on real-world long-tailed fish recognition: the best open-source model achieves less than 40% overall accuracy, revealing limited generalization to specialized fine-grained domains. Incorporating expert-level textual descriptions and keypoint annotations during zero-shot inference significantly improves performance, demonstrating that structured domain knowledge and morphological guidance provide effective signals for fine-grained recognition.

To investigate the root causes of this degradation, we leverage the multimodal design of FishNet2.0 to construct controlled diagnostic evaluations that decouple domain knowledge from visual perception. We design two disentangled settings: a domain-knowledge setting of text-only tasks (textual question answering, category description, and name translation) that isolate conceptual and factual understanding of fish species, and a perception setting of vision-only tasks (visual question answering, detection, and keypoint localization) that require fine-grained visual discrimination without specialized priors. Our analysis reveals that degradation primarily stems from insufficient domain-specific knowledge rather than pure visual-perception limitations, highlighting a fundamental bottleneck in current MLLMs on expert-level biological recognition and underscoring the importance of domain-aware multimodal benchmarks.

Our contributions are summarized as follows:

- We introduce FishNet2.0, a comprehensive multimodal benchmark for long-tailed fine-grained fish understanding with structured expert-level annotations and dense key part annotations, supporting a broad set of tasks.

Datasets	Properties			Tasks			
	Images	Species	Descriptions	Fine-Grained	Detection	Part-Location	QA
Fish-gres Chastine (2021)	3,248	8	0	✗	✗	✗	✗
NCFM Conservancy (2017)	16,915	8	0	✗	✗	✗	✗
Mediterranean Fish Species Georgiou (2021)	≈40,000	20	0	✗	✗	✗	✗
Fish4-Knowledge-A Nadarajan et al. (2009)	27,370	23	0	✗	✓	✗	✗
Image Dataset Francescangeli & Jacopo (2023)	33,805	30	0	✗	✗	✗	✗
Fish Abundance Gilby et al. (2017)	4,909	50	0	✗	✗	✗	✗
SEAMPD21 Boulais et al. (2021)	28,328	130	0	✓	✗	✗	✗
iNaturalist_Fish iNaturalist (2021)	54,006	369	0	✓	✗	✗	✗
FathomNet Katija et al. (2022)	84,454	2,244	0	✓	✓	✗	✗
WildFish++ Zhuang et al. (2020)	103,034	2,348	3,817	✓	✗	✗	✗
FishNet Khan et al. (2023)	94,532	17,357	0	✓	✓	✗	✗
FishNet2.0 (Ours)	99,556	17,393	17,393	✓	✓	✓	✓

Table 1: Comparison with existing datasets for marine species. FishNet2.0 provides textual descriptions for 17,393 species, while previous datasets only provide species labels. FishNet2.0 supports additional tasks for detection, key-part localization, and question answering(QA).

- We establish a controlled diagnostic evaluation framework that disentangles domain knowledge from visual perception, enabling principled analysis of multimodal reasoning under expert-level biological settings.
- We systematically benchmark state-of-the-art MLLMs on FishNet2.0, uncovering significant performance gaps and identifying domain-specific knowledge deficiency as the primary bottleneck.
- We demonstrate that leveraging our structured multimodal annotations, particularly expert-level textual descriptions and keypoint guidance, significantly improves zero-shot performance, highlighting the practical value of domain-aware multimodal design.

2 Related Work

Species Recognition. Species recognition is a major focus in ecology and biodiversity monitoring, which poses unique challenges (e.g., high intra-class variance, inter-class similarity, and class imbalance) Nadarajan et al. (2009); Boulais et al. (2021); Gilby et al. (2017); Francescangeli & Jacopo (2023); Conservancy (2017); Van Horn et al. (2018); Zhuang et al. (2020). For aquatic environments specifically, new datasets have been released. These include Fishnet Open Images Database Kay & Merrifield (2021), an open images dataset of $86,000$ fish from 34 species captured by vessel-borne cameras, which highlights domain conditions such as murky water, skewed species distribution, and occlusion. AutoFish Bengtson et al. (2025), another dataset with $1,500$ controlled setup collected images of 454 fish instances annotated with segmentation and IDs. Additionally, FathomNet Katija et al. (2022) provides a large-scale, underwater image database covering over 2,000 marine taxa, designed to support species-level recognition and detection. We compare FishNet2.0 dataset with further existing datasets in Table 1.

Fine-Grained Recognition. Fine-grained recognition (FGR) Wah et al. (2011); Berg et al. (2014); Van Horn et al. (2018); Khan et al. (2023) evaluates a model's ability to distinguish visually similar categories by identifying subtle, discriminative cues and reasoning over them. Recent benchmarks, such as FOCI Geigle et al. (2024), extend this evaluation to MLLMs by assessing performance using multiple-choice questions constructed via negative mining. To advance fine-grained recognition in the marine domain, we introduce FishNet2.0 and systematically evaluate MLLMs under different settings. Our results reveal that current models struggle substantially with fine-grained marine recognition, motivating a deeper analysis toward improving fine-grained discrimination in specialized domains.

Multimodal Large Language Models (MLLMs). MLLMs have advanced multimodal understanding and reasoning through large-scale pretraining, supervised fine-tuning, and often reinforcement learning from human feedback (RLHF) OpenAI (2024a;b); AI (2024); DeepSeek-AI et al.; Qwen et al. (2025); Yang et al. (2024); Ouyang et al. (2022). Despite this progress, they struggle to generalize to specialized or underrepresented domains that are sparsely represented during training Zhang et al. (2024). Thus, how MLLMs perform on marine species recognition, where many

visual concepts are likely underrepresented in their pretraining data, remains largely unclear. Motivated by this gap, we introduce our comprehensive benchmark, along with a diverse set of evaluation tasks, designed to serve as a principled testbed for systematically evaluating and analyzing MLLM capabilities for fine-grained understanding of marine species.

3 FISHNET2.0

While it is estimated that over 95% of the world's bird species have been described Barrowclough et al. (2016), the vast majority of marine life remains a mystery, with some estimates suggesting over 90% of species are yet to be discovered Mora et al. (2011). Despite this enormous knowledge gap, the computer vision community has predominantly focused on terrestrial animals Wah et al. (2011); Berg et al. (2014). To help bridge this disparity and advance aquatic science, we introduce FishNet2.0, a large-scale, long-tailed, multi-modal benchmark developed from the original FishNet dataset Khan et al. (2023). Our primary goal is to enable the development of models capable of large-scale, language-based species recognition, a foundational step towards the ultimate challenge of identifying unseen species.

FishNet2.0 is enriched with 17,393 textual species descriptions and annotations for detection, key-part localization, and question answering. Below we outline our data collection methodology, covering taxonomic correction, description generation, question-answer generation, and bounding-box and keypoint annotation. To ensure scientific validity, the entire process was conducted in close collaboration with marine biology experts.

Species Description. To generate descriptive text for each species, we drew on multiple reliable sources vetted by domain experts: FishBase Froese & Pauly (2025) as the primary source of morphological and ecological information, supplemented by data crawled from iNaturalist iNaturalist contributors (2025), WoRMS WoRMS Editorial Board (2025), and NOAA NOAA Fisheries (2025). We then used GPT-4o OpenAI (2024a) to produce a coherent, concise descriptive summary for each species. To validate reliability, we conducted a user study for each species. Three annotators were shown four images of the species alongside the description and rated its helpfulness for identification on a 1–5 scale (1 = "not helpful at all", 5 = "very helpful"). The descriptions received a mean of 3.89, a median of 3.94, and a mode of 4.00, indicating good recognition quality, a finding further supported by our experiments, which showed that descriptions improve performance.

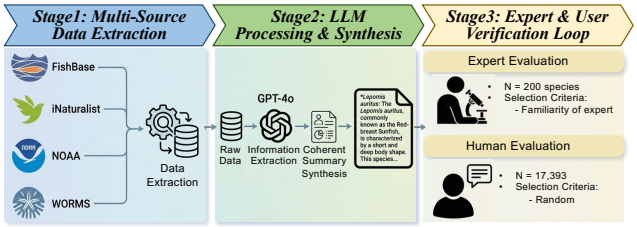

Figure 2: Data collection pipeline: multi-source extraction, LLM-based synthesis, and expert and human evaluation.

We also asked an expert marine taxonomist for 200 species they were familiar with and presented the corresponding generated descriptions; the expert verified that the information was accurate and that the distinguishing characteristics were visually discriminative, within the description guidelines (coherence and conciseness) with no wrong information. The overall pipeline is shown in Figure 2.

3.1 KEY-POINT SELECTION AND COLLECTION

We finalized six-part locations and one attribute to be collected for every image in FishNet2.0. The selection of these parts and attributes was done in consultation with experts to ensure the dataset's utility for both the machine learning and aquatic science communities. The parts are as follows: 1) Eye location, 2) Mouth location, 3) Pectoral, pelvic, and anal fin location, 4) Center of the main body, 5) Tail (caudal fin) start, and 6) Tail end. All parts were annotated by pixel location, and we additionally record whether the species is underwater or above water. Fin locations may involve multiple points, vary by species, and be affected by image angle and occlusion; similarly, the tail apex may have one or two points depending on its shape. The keypoints also yield the aspect ratio, which has been linked to species' behaviour, metabolism, ecological lifestyle, and response to

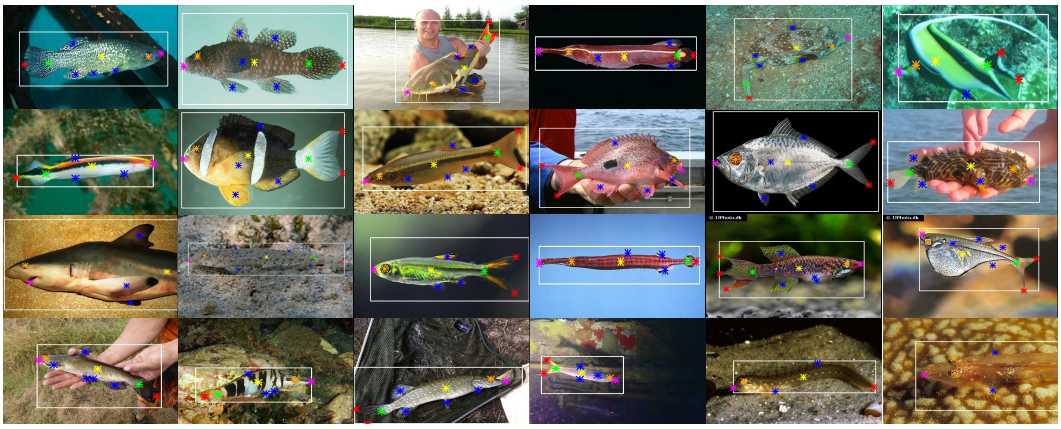

Figure 3: Example images from FishNet2.0 showcasing part-level annotations. Each keypoint is color-coded by semantic part: eye (orange), fins (blue), mouth (magenta), body center (yellow), tail start (green), and tail apex (red). The number and placement of fins vary across species, and some species exhibit a forked tail apex. For each image, we also display the annotated bounding box.

thermal stress Sambilay (1990); Campos et al. (2018); THOMSON & SIMANEK (2015), making it valuable for understanding species' ecology and conservation decision-making. Key-part locations can further serve as weak supervision for dense annotations such as segmentation (details in the appendix Section E).

To collect the part location annotations, we partnered with a data annotation company, with experts supervising and validating the process and regular manual checks for quality control. In total, FishNet2.0 provides $706,426$ key part locations: $86,589$ eye, $77,990$ mouth, $281,426$ fin, $80,653$ body, $73,785$ tail-start, and $105,983$ tail-end locations. Of these images, $38,326$ show fish above the water surface.

3.2 Dataset Refinement and Expansion

In addition to the newly annotated species descriptions and keypoint annotations, we further refine FishNet through taxonomic standardization and dataset expansion.

Taxonomic Corrections. The taxonomy of species worldwide is continuously evolving Bouchet et al. (2023), making it essential to ensure that datasets reflect the most up-to-date and accurate species names. We compared all labels in FishNet Khan et al. (2023) against the taxonomic standards (as per Froese & Pauly (2025)) and found that 266 species names in the FishNet dataset no longer aligned with current taxonomy. To address this, we manually remapped these outdated names to their correct, updated counterparts. Following this, we also associated each species in our dataset with its corresponding FishBase species code, which remains constant even as taxonomic names change. This provides a straightforward mechanism for keeping our dataset aligned with current taxonomic nomenclature. In the end, we identified 36 of our newly added images that did not correspond to any known species in the entire taxonomy. These may represent entirely new species to science.

Additional Images. The original Fishnet Khan et al. (2023) is highly long-tailed, with only 495 species with ten or more images. The bias in the number of images is often associated with species that are not commercially exploited at large scales (either for fisheries, ecotourism, or the aquarium trade), those found in less well-researched parts of the world, or those found in less accessible depth ranges. For FishNet2.0, we sought to increase the number of species with reasonable image representation. We collected additional images for species from various underrepresented regions worldwide, including Egypt, Indonesia, Oman, Seychelles, Papua New Guinea, and Saudi Arabia, sourced through a wider network of collaborators who provided access to their extensive collections. In total, we gathered an additional 5,024 images, increasing the number of species with at least ten images from 495 to 804, significantly enhancing their diversity and representation.

3.3 BENCHMARK TASKS

FishNet2.0 supports a diverse set of evaluation tasks designed to assess multimodal reasoning under long-tailed fine-grained settings, covering recognition, localization, and knowledge-driven reasoning.

Fine-grained Classification. We evaluate species-level recognition across highly similar categories under realistic long-tailed distributions, measuring model robustness to subtle morphological distinctions and rare classes. Together with our newly proposed text description and key-point annotations, we enable the fine-grained classification under various multimodal inputs.

Object Detection. Bounding-box annotations enable localization of fish instances within images, supporting instance-level evaluation under diverse environmental conditions.

Keypoint Localization. The structured keypoint annotations from FishNet2.0 for morphological landmarks (e.g., eye, fin, and tail boundaries) facilitate fine-grained localization of structural keypoints.

Text-based Question Answering. To evaluate domain knowledge in isolation, we construct a text-based QA benchmark using curated species descriptions. For each species, GPT-4o generates five multiple-choice questions focusing on discriminative morphological traits such as eye color, body shape, and fin characteristics. The questions are designed to be answerable solely from species-level knowledge without requiring visual input. Each question includes one correct answer and three plausible distractors.

Visual Question Answering. For the visual QA benchmark, we create perception-focused questions directly tied to observable image attributes. For each image, we formulate four questions targeting measurable or spatial properties, including the number of visible fins, the number of eyes, the tail endpoints, and the environmental context (e.g., underwater, on land, or suspended). Numerical distractors are deliberately chosen to be close to the ground-truth values (e.g., ±1 or ±2) to increase difficulty, while habitat labels are manually verified to ensure accuracy. Together, these two QA benchmarks enable controlled evaluation of domain knowledge and visual perception in isolation.

4 EXPERIMENTS

Data Splits. The dataset is categorized into head, middle, and tail classes, comprising 829, 4,966, and 11,598 categories, respectively. The head consists of 62,456 instances, the middle consists of 21,900 instances, and the tail consists of 15,200 instances. Species with at least 9 images are put in the head classes, and species with 3-8 images are included in the middle classes. Species with fewer than 3 images (i.e., only 1 or 2 images) are included in the tail classes.

We adopt a 75-5-20 train-validation-test split for species belonging to the head and middle classes of FishNet2.0. For species with 4-9 images, we assign two images to the test set and use the remaining images for training. For species with exactly 3 images, one image is reserved for testing. The rest follow the 75-5-20 split. This splitting strategy is inspired by FishNet Khan et al. (2023), which drops species with very few samples (1 or 2) for the classification experiments. However, in contrast, we retain these underrepresented species in the tail classes to thoroughly evaluate the recognition capabilities of vision-language models.

4.1 MAIN RESULTS ON FINE-GRAINED RECOGNITION

Task Setup. We first evaluate various MLLMs on fine-grained species recognition. Following prior work He et al. (2025); Geigle et al. (2024); Khan et al. (2025), the MLLM is given a query image and four candidate species. However, instead of CLIP Radford et al. (2021), we construct distractors with BioCLIP2 Gu et al. (2025), which is better suited for species-level recognition and yields substantially harder options: we select the non-ground-truth species whose name embedding is most cosine-similar to the query image embedding, whereas CLIP produces an easier setting (see appendix Section B). We then analyze model performance and strategies for improving fine-grained recognition.

	Image Only			Image + Descriptions			Image + Desc. + KP		
	Head	Middle	Tail	Head	Middle	Tail	Head	Middle	Tail
LLavaNeXt Li et al. (2024b)	23.1	20.1	19.4	24.4	24.4	22.9	23.7	22.2	19.6
Phi-4 Abdin et al. (2024)	24.7	19.5	17.5	27.1	23.6	21.3	27.1	23.1	21.2
Molmo-2 Clark et al. (2026)	22.9	21.2	20.3	37.6	37.5	33.2	39.5	37.4	32.7
LLavaOneVision Li et al. (2024a)	27.5	25.0	22.7	30.7	28.7	24.6	31.1	28.8	25.5
InternVL-3.5 Wang et al. (2025)	29.7	25.0	20.4	39.5	35.3	31.2	37.9	34.3	29.8
Idefics-3 Laurençon et al. (2024)	34.7	30.6	28.4	37.1	32.2	29.7	33.9	29.9	27.5
Gemma-3n Team et al. (2025)	33.4	30.7	26.7	40.1	38.3	34.2	41.7	40.1	35.1
MiniCPM-V-4.5 Yao et al. (2024)	32.8	28.3	25.5	43.5	41.0	37.1	45.9	42.1	38.6
Qwen3-VL Bai et al. (2025)	**41.6**	**39.1**	**37.0**	**52.6**	**48.9**	**45.3**	**55.6**	**49.1**	**45.9**

Table 2: Fine-grained recognition accuracy across head, middle, and tail categories under different conditioning strategies: Image-Only (image + species names), Image + Descriptions, and Image + Descriptions + Key Points (KP).

In Table 2, we report zero-shot fine-grained recognition performance under different multimodal input contexts using FishNet2.0. We evaluate models across all the splits (Head, Middle, and Tail) to better understand where fine-grained discrimination succeeds or fails.

Image Only. We begin with a strict zero-shot setting in which the query image is presented alongside four candidate species names, with no additional semantic guidance. As shown in the left section of Table 2, accuracy remains limited across all models: even the strongest achieves only 39.2% on average, demonstrating that even state-of-the-art MLLMs struggle with fine-grained discrimination in the specialized marine domain when only label names are provided.

Image + Description. Under this setting, we augment the species names with our curated text descriptions, providing the models with external knowledge to facilitate visual reasoning. As seen in the middle section of Table 2, the inclusion of high-quality descriptive information consistently improves performance for all models across every level of the distribution (Head, Middle, and Tail). Qwen3-VL, the best-performing model, shows an average increase of 9.7% across the head, middle, and tail classes. Across all the MLLMs, performance increases by an average of 7.2%. This improvement indicates that MLLMs can effectively leverage linguistic cues to resolve visual ambiguities in species identification and also highlights the high-quality text descriptions in FishNet2.0.

Image + Description + Key Parts. Finally, we further try to enhance the descriptions with explicit key-point guidance, providing localized cues corresponding to discriminative anatomical regions. We do this by highlighting various key points using the collected keypoint annotations. As shown in the right section of Table 2, this additional structural guidance yields further improvements, especially for models (Qwen3-VL, MiniCPM-V-4.5, Gemma-3n), which show strong gains from additional textual context. Moreover, as shown in Table 3, a consistent trend emerges: models with stronger key-point localization tend to benefit from explicit key-point guidance, except for InternVL-3.5, while models with poor key-point localization skills, like Idefics-3, and LlavaNeXt, degrade. This observation suggests that improvements in fine-grained recognition are closely linked to a model's capacity to align textual attributes with precise visual regions, reinforcing the importance of spatially grounded reasoning.

Overall, the progressive performance gains from Image Only to Descriptions to Descriptions + Key Points reveal two important insights: (1) semantic enrichment substantially improves the zero-shot performance, and (2) explicit visual guidance can further enhance fine-grained discrimination for models with strong visual capabilities. These findings indicate that while both domain knowledge and region-level perception help improve recognition, language appears to be the major bottleneck, as evidenced by the performance boost.

4.2 Disentangling Domain Knowledge from Visual Perception.

To investigate the primary source of the observed performance gap, we disentangle domain knowledge from visual perception through controlled experimental designs. Specifically, we construct

	Vision				Text			
	VQA	Detection	Key-Point	Avg.	QA	Description	Name	Avg.
LLavaNeXt Li et al. (2024b)	42.7	64.5	1.3	36.2	33.4	26.4	5.5	21.8
Phi-4 Abdin et al. (2024)	60.6	37.8	1.5	33.3	37.0	14.9	3.5	18.5
Molmo-2 Clark et al. (2026)	63.2	26.2	7.6	32.3	45.8	16.3	4.8	22.3
LLavaOneVision Li et al. (2024a)	57.0	44.9	7.4	36.4	41.9	13.1	2.2	19.1
InternVL-3.5 Wang et al. (2025)	44.3	68.8	21.5	44.9	37.4	20.5	2.0	20.0
Idefics-3 Laurençon et al. (2024)	46.3	24.9	1.2	24.1	44.2	16.2	8.8	23.1
Gemma-3n Team et al. (2025)	48.6	53.2	10.9	37.6	38.5	22.4	4.6	21.8
MiniCPM-V-4.5 Yao et al. (2024)	59.8	68.8	12.6	47.1	41.2	32.3	6.0	26.5
Qwen3-VL Bai et al. (2025)	**65.5**	**72.1**	**33.1**	**56.9**	**46.6**	**34.0**	**11.7**	**30.8**

Table 3: Analysis results from both visual and textual-only VQA

two complementary evaluation settings: one that relies solely on textual reasoning to assess domain knowledge, and another that relies solely on visual reasoning to assess perceptual capabilities.

a) Domain Knowledge. Below we evaluate whether MLLMs possess the fundamental marine domain knowledge required for our task, independent of their visual perception capabilities.

Text-Based QA. This setup forms the basis of our text-only Question Answering (QA) analysis. We prompt each MLLM with species-specific, attribute-focused questions with four options to assess their domain knowledge of aquatic species, without providing any visual input. As reported in the "Text-QA" column of Table 3, the best-performing model (Qwen3-VL) achieves an accuracy of only 46.6%, with other models performing even worse. Given a random chance of success is 25%, these results reveal a substantial deficiency in the marine domain knowledge of current MLLMs.

Description Consistency. We ask each model to describe every species in FishNet2.0 and use GPT-4o to assess consistency with expert ground truth. A description is accepted if it contains no conflicting information, even when details are omitted. As shown in the "Description" column of Table 3, the best model achieves only 34.0%, revealing substantial hallucinations and knowledge gaps in marine biology.

Name Translation. In this test, We evaluate bidirectional translation between common and scientific names for all 17,393 species using FishBase Froese & Pauly (2025). The "Name" column of Table 3 reports accuracy averaged across both directions. The best model, Qwen3-VL, achieves only 11.7%, compared with 94.3% on CUB-200-2011 birds Wah et al. (2011). This contrast indicates a specific lack of marine-biological knowledge rather than a general MLLM limitation.

Overall, the consistently poor performance across all three text-only evaluations, spanning attribute-based QA, description consistency, and name translation, provides strong evidence that current MLLMs lack fundamental domain-specific knowledge in marine biology, a major bottleneck for fine-grained recognition of aquatic species.

4.2.1 B) VISUAL PERCEPTION.

Below, we evaluate whether MLLMs possess the visual perception capabilities necessary for fine-grained marine recognition, independent of domain-specific textual knowledge.

Visual Question Answering (VQA). We prompt each MLLM with image-specific perception questions to assess visual reasoning. As shown in the "Vision-VQA" column of Table 3, all models substantially outperform their text-only QA counterparts. Qwen3-VL, for example, reaches 65.5% accuracy, nearly 20 points higher than its text-only QA score (46.6%), indicating that MLLMs can reliably extract visual attributes despite limited domain knowledge.

Coarse-Grained Vision (Detection). We evaluate object localization using the ground-truth bounding boxes in FishNet2.0 and report mean IoU over thresholds from 0.50 to 0.90 (details in the appendix Section D). Detection performance is strong: Qwen3-VL achieves 72.1 IoU, close to a supervised YOLO-12 Tian et al. (2025) model trained on FishNet2.0 (76.4), suggesting localization is not the main bottleneck.

Fine-Grained Vision (Keypoint Localization). To evaluate fine-grained visual understanding, we conduct a keypoint localization task requiring models to predict six anatomical landmarks. We report the average PCK@0.1 Novotny et al. (2018); Yang & Ramanan (2013) over all parts, which measures the proportion of predicted keypoints within a normalized distance threshold relative to object size. Results for each part separately are presented in the appendix Section D.

In the "Key-Point" column of Table 3, Qwen3-VL achieves 33.1 PCK@0.1, approaching the performance of a YOLO-12 model trained on FishNet2.0 (38.1). Although absolute performance remains modest, the fact that a general-purpose MLLM is competitive with a task-specific supervised model demonstrates that strong MLLMs possess meaningful fine-grained spatial perception capabilities.

Taken together, the VQA, detection, and keypoint results indicate that modern MLLMs possess strong coarse- and fine-grained visual perception. In contrast to their much weaker text-only performance, this suggests that the primary bottleneck in fine-grained marine recognition lies in insufficient domain knowledge encoded in the language component rather than visual perception.

4.3 Performance Analysis: Vision vs. Language

To further validate our hypothesis that the language component is the primary bottleneck in MLLM recognition performance, we conduct supervised fine-tuning (SFT) using LoRA Hu et al. (2022) on three components of Qwen3-VL: (1) the vision encoder, (2) the language decoder, and (3) the full model. We choose Qwen3-VL because it achieves the strongest zero-shot performance on FishNet2.0. Implementation details are provided in the appendix Section A.

Table 4: Recognition Accuracy of Qwen3-VL when fine-tuned.

Model	LoRA	Head	Middle	Tail
Qwen3-VL	Zero-Shot	52.6	48.9	45.3
Qwen3-VL	Vision	62.1	54.6	50.6
Qwen3-VL	Language	86.3	80.3	77.4
Qwen3-VL	Full	**86.8**	**81.0**	**78.7**

As shown in Table 4, the gains depend strongly on the component being fine-tuned.

Vision-Only Fine-Tuning. Fine-tuning only the vision encoder yields modest improvements, increasing Head accuracy from 52.6% to 62.1%. While better visual representations help, the model still struggles to map them to the correct taxonomic labels.

Language-Only Fine-Tuning. In contrast, fine-tuning only the language decoder produces a dramatic improvement, boosting Head accuracy to 86.3%, with similar gains on Middle (80.3%) and Tail (77.4%) classes. This suggests that the model already possesses strong visual representations but lacks domain-specific linguistic knowledge.

Full Model Fine-Tuning. Fine-tuning the entire model yields the highest overall results (86.8% on Head classes), though the marginal gain over language-only fine-tuning is minimal (0.5%).

These results indicate that the main limitation in zero-shot recognition is taxonomic mapping, not visual perception: the vision encoder already extracts discriminative features, while the language decoder lacks domain-specific associations to link them to species categories. Language-based fine-tuning is especially effective for rare "Tail" classes, boosting accuracy from 45.3% to 77.4%. Aligning visual cues with correct taxonomic and semantic representations is thus key, making domain-aware language adaptation a more efficient path than enhancing visual representations.

5 Conclusion

We introduced FishNet2.0, a multimodal benchmark for long-tailed fine-grained marine understanding, combining broad species coverage with expert annotations such as descriptions, keypoints, bounding boxes, and multimodal question-answer pairs. Evaluations of state-of-the-art MLLMs reveal substantial performance gaps, with controlled analyses identifying domain-specific knowledge, not visual perception, as the primary bottleneck. Structured species descriptions and keypoint guidance also improve zero-shot performance, demonstrating the value of combining domain knowledge with strong visual representations. We hope FishNet2.0 supports future research on robust, long-tailed multimodal learning in specialized scientific domains.

6 Ethics Statement

Our benchmark is intended to support scientific research on fine-grained marine-species understanding and does not study human identity or sensitive personal attributes. As with other biodiversity datasets, FishNet2.0 may reflect geographic, taxonomic, and imaging biases, particularly for rare species. Moreover, the evaluated models can hallucinate or misidentify species. We encourage users to consider these limitations and the potential ecological sensitivity of marine-species data when developing downstream applications.

7 AI Use Statement

Generative AI tools were used as part of our experiments to generate species descriptions and to evaluate model outputs, as detailed in the paper and supplementary material. We document the models, prompts, and evaluation protocols used in these processes.

8 Reproducibility Statement

The main paper describes the construction of FishNet2.0, the evaluated tasks, models, and metrics. The appendix and supplementary material provide details on data collection, annotation procedures, dataset splits, prompts, model and evaluator versions, implementation settings, and evaluation protocols. We will release the complete dataset under a Creative Commons license, providing unrestricted download access for reproducibility and future research. We will also release the preprocessing and evaluation code, configuration files, and relevant model outputs.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Google AI. Gemini: Google's multimodal ai model. *Google AI Research*, 2024. https://fireflies.ai/blog/gemini-vs-gpt-4.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Edward B. Barbier. Marine ecosystem services. *Current Biology*, 27(11):R507–R510, 2017. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2017.03.020. URL https://www.sciencedirect.com/science/article/pii/S0960982217302890.

George F Barrowclough, Joel Cracraft, John Klicka, and Robert M Zink. How many kinds of birds are there and why does it matter? *PLoS one*, 11(11):e0166307, 2016.

Xavier Basurto, Nicolas L. Gutierrez, Nicole Franz, et al. Illuminating the multidimensional contributions of small-scale fisheries. *Nature*, 637(8047):875–884, 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08448-z. URL https://doi.org/10.1038/s41586-024-08448-z.

Stefan Hein Bengtson, Daniel Lehotský, Vasiliki Ismiroglou, Niels Madsen, Thomas B. Moeslund, and Malte Pedersen. Autofish: Dataset and benchmark for fine-grained analysis of fish, 2025. URL https://arxiv.org/abs/2501.03767.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L. Alexander, David W. Jacobs, and Peter N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Philippe Bouchet, Wim Decock, Britt Lonneville, Bart Vanhoorne, and Leen Vandepitte. Marine biodiversity discovery: the metrics of new species descriptions. *Frontiers in Marine Science*, Volume 10 - 2023, 2023. ISSN 2296-7745. doi: 10.3389/fmars.2023.929989. URL https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2023.929989.

Océane Boulais, Simegnew Yihunie Alaba, John E Ball, Matthew Campbell, Ahmed Tashfin Iftekhar, Robert Moorehead, James Primrose, Jack Prior, Farron Wallace, Henry Yu, et al. Seamapd21: A large-scale reef fish dataset for fine-grained categorization. In *Proceedings of the FGVC8: The Eight Workshop on Fine-Grained Visual Categorization, Online*, volume 25, 2021.

D.F. Campos, A.L. Val, and V.M.F. Almeida-Val. The influence of lifestyle and swimming behavior on metabolic rate and thermal tolerance of twelve amazon forest stream fish species. *Journal of Thermal Biology*, 72:148–154, 2018. ISSN 0306-4565. doi: https://doi.org/10.1016/j.jtherbio.2018.02.002. URL https://www.sciencedirect.com/science/article/pii/S030645651730236X.

Prasetyo Eko; Suciati Nanik; Fatichah Chastine. Fish-gres dataset for fish species classification. *Mendeley Data*, 2021. doi: 10.17632/76cr3wfhff.1.

Christopher Clark, Jieyu Zhang, Zixian Ma, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

The Nature Conservancy. The nature conservancy fisheries monitoring. *Kaggle Data*, 2017. URL https://www.kaggle.com/competitions/the-nature-conservancy-fisheries-monitoring/overview.

DeepSeek-AI, Daya Guo, Dejian Yang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. URL https://arxiv.org/abs/2501.12948.

Marco; Marini Simone; Martínez Enoc;Del Río Joaquín;Toma Daniel M.;Nogueras Marc; Francescangeli and Aguzzi Jacopo. Image dataset for benchmarking automated fish detection and classification algorithms. *Scientific Data*, 117(24):13596–13602, 2023. doi: https://doi.org/10.1038/s41597-022-01906-1. URL https://www.pnas.org/doi/abs/10.1073/pnas.1922686117.

Rainer Froese and Daniel Pauly. Fishbase. https://www.fishbase.org, 2025. World Wide Web electronic publication, accessed February 2025.

Yaozong Gan, Guang Li, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Cross-domain few-shot in-context learning for enhancing traffic sign recognition. In *ICIP*, 2024.

Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv*, abs/2406.14496, 2024. URL https://arxiv.org/abs/2406.14496.

Giannis Georgiou. Fish species. *Kaggle Data*, 2021. URL https://www.kaggle.com/datasets/giannisgeorgiou/fish-species.

Ben L. Gilby, Andrew D. Olds, Rod M. Connolly, Nicholas A. Yabsley, Paul S. Maxwell, Ian R. Tibbetts, David S. Schoeman, and Thomas A. Schlacher. Umbrellas can work under water: Using threatened species as indicator and management surrogates can improve coastal conservation. *Estuarine, Coastal and Shelf Science*, 199:132–140, 2017. ISSN 0272-7714. doi: https://doi.org/10.1016/j.ecss.2017.10.003. URL https://www.sciencedirect.com/science/article/pii/S0272771417309216.

Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, James Balhoff, Wasla Dahdul, Daniel Rubenstein, Hilmar Lapp, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP 2: Emergent properties from scaling hierarchical contrastive learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=yPC9zmkQgG.

Daya Guo, Dejian Yang, Haowei Zhang, and other. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p3NKpom1VL.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. MLLM-for3d: Adapting multimodal large language model for 3d reasoning segmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=IiPSP4OUYx.

iNaturalist. inaturalist, 2021. URL https://www.inaturalist.org. Accessed: May 6, 2024.

iNaturalist contributors. inaturalist. https://www.inaturalist.org, 2025. Accessed March 2025.

Kakani Katija, Eric Orenstein, Brian Schlining, Lonny Lundsten, Kevin Barnard, Giovanna Sainz, Oceane Boulais, Megan Cromwell, Erin Butler, Benjamin Woodward, and Katherine L. C. Bell. Fathomnet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1):15914, 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-19939-2. URL https://doi.org/10.1038/s41598-022-19939-2.

Justin Kay and Matt Merrifield. The fishnet open images database: A dataset for fish detection and fine-grained categorization in fisheries, 2021. URL https://arxiv.org/abs/2106.09178.

Faizan Farooq Khan, Xiang Li, Andrew J. Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20496–20506, October 2023.

Faizan Farooq Khan, Jun Chen, Youssef Mohamed, Chun-Mei Feng, and Mohamed Elhoseiny. Neural catalog: Scaling species recognition with catalog of life-augmented generation, 2025. URL https://arxiv.org/abs/2505.05635.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. URL https://arxiv.org/abs/2408.12637.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024b. URL https://arxiv.org/abs/2407.07895.

Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011.

Gayathri Nadarajan, Yun-Heh Chen-Burger, and Robert B Fisher. A knowledge-based planner for processing unconstrained underwater videos. In *IJCAI'09 Workshop on Learning Structural Knowledge From Observations (STRUCK'09)*, 2009.

NOAA Fisheries. Noaa fisheries – national oceanic and atmospheric administration. https://www.fisheries.noaa.gov, 2025. Accessed March 2025.

David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection, 2018.

OpenAI. Gpt-4o: Enhanced multimodal language model. *OpenAI Research*, 2024a. https://openai.com/index/hello-gpt-4o/.

OpenAI. Gpt-4v: Multimodal language model with vision capabilities. *OpenAI Research*, 2024b. https://openai.com/index/gpt-4/.

Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Kunyu Peng, Jia Fu, Kailun Yang, Di Wen, Yufan Chen, Ruiping Liu, Junwei Zheng, Jiaming Zhang, M Saquib Sarfraz, Rainer Stiefelhagen, et al. Referring atomic video action recognition. In *ECCV*, 2024.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Victor C. Sambilay, Jr. Interrelationships between swimming speed, caudal fin aspect ratio and body length of fishes. *Fishbyte*, 8(3):16–20, 1990.

Elizabeth R. Selig, David G. Hole, Edward H. Allison, Katie K. Arkema, Madeleine C. McKinnon, Jingjie Chu, Alex de Sherbinin, Brendan Fisher, Louise Glew, Margaret B. Holland, Jane Carter Ingram, Nalini S. Rao, Roly B. Russell, Tanja Srebotnjak, Lydia C.L. Teh, Sebastian Troëng, Will R. Turner, and Alexander Zvoleff. Mapping global human dependence on marine ecosystems. *Conservation Letters*, 12(2):e12617, 2019. doi: https://doi.org/10.1111/conl.12617. URL https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/conl.12617.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

KEITH STEWART THOMSON and DAN E. SIMANEK. Body form and locomotion in sharks. *American Zoologist*, 17(2):343–354, 08 2015. ISSN 0003-1569. doi: 10.1093/icb/17.2.343. URL https://doi.org/10.1093/icb/17.2.343.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*, 2025.

Jintao Tong, Shilin Yan, Hongwei Xue, Xiaojun Tang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. Swimbird: Eliciting switchable reasoning mode in hybrid autoregressive mllms. *arXiv preprint arXiv:2602.06040*, 2026.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Jul 2011.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20193–20203, October 2025.

WoRMS Editorial Board. World register of marine species. https://www.marinespecies.org, 2025. Accessed March 2025.

An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, and other. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013. doi: 10.1109/TPAMI.2012.261.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models, 2024. URL https://arxiv.org/abs/2402.06599.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Peiqin Zhuang, Yali Wang, and Yu Qiao. Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Transactions on Multimedia*, 23:3603–3617, 2020.

APPENDIX: FISHNET 2.0: A COMPREHENSIVE MULTIMODAL BENCHMARK FOR LONG-TAILED FINE-GRAINED FISH UNDERSTANDING

The appendix material provides:

- Section A provides detailed experimental settings for both training and inference.
- Section B presents a comparison between CLIP and BioCLIP2 for distractor mining.
- Section C provides details on the species description generation process.
- Section D reports detailed results for detection and keypoint localization.
- Section E describes the process of obtaining segmentation masks from keypoints.
- Section F provides details on the data collection process.
- Section G discusses the limitations of FishNet2.0.

A EXPERIMENTAL DETAILS.

All inference experiments were conducted on a single NVIDIA A100 GPU. Across all Multimodal Large Language Models (MLLMs), the target species name was generated via multiple-choice prompting.

For the baseline inference using only species names as candidate options, the following prompt was employed:

> You are an expert marine biologist specializing in fish species recognition.
> Task:
> You are given:
> 1. An image of a fish.
> 2. Four candidate species names.
> You must pick exactly one species from the four options.
> Respond ONLY with the chosen species name, with no explanation.
> Options:
> Option A, Option B, Option C, Option D

When incorporating textual morphological descriptions into the inference process, the prompt was expanded as follows:

> You are an expert marine biologist specializing in fish species recognition.
> Task:
> You are given:
> 1. An image of a fish.
> 2. Four candidate species names with their descriptions.
> You must pick exactly one species from the four options.
> Respond ONLY with the chosen species name, with no explanation.
> Options:
> Option A with description, Option B with description, Option C with description, Option D with description

For inference utilizing spatial keypoint annotations, the prompt explicitly defined the color-coded anatomical landmarks to guide the model:

You are an expert marine biologist specializing in fish species recognition.
Task:
You are given:
1. An image of a fish. Keypoints are marked with colored squares indicating anatomical landmarks:
Pink: mouth start
Orange: eye
Yellow: body center
Dark blue: center of fins
Green: start of tail
Red: end of tail
2. Four candidate species names with their descriptions.
You must pick exactly one species from the four options.
Respond ONLY with the chosen species name, with no explanation.
Options:
Option A with description, Option B with description, Option C with description, Option D with description

Training Details LORA-based supervised fine-tuning of Qwen3-VL was performed on one A100 GPU with 80GB of memory for 4000 steps, with an effective batch size of 16 and rank 8. Optimization was conducted using AdamW with an initial learning rate of 0.00001, a cosine learning rate schedule, and a warmup ratio of 0.05.

YOLO-based model training for detection and keypoint localization was performed using the Ultralytics YOLO framework for 30 epochs with a mini-batch size of 16 images. All input images were uniformly resized to 640×640 pixels. Optimization was conducted using Stochastic Gradient Descent on an NVIDIA V100 GPU with 32 GB of memory, employing an initial learning rate of 0.01, a momentum factor of 0.937, and a weight decay of 0.0005. A 3-epoch warmup phase was employed, linearly increasing the momentum from 0.8 and the bias learning rate from 0.1. We used the corresponding YOLO model as the base model, with pre-trained weights to speed up convergence and enhance performance.

	BioClip2			**Clip**		
	Head	Middle	Tail	Head	Middle	Tail
LLavaNeXt Li et al. (2024b)	23.1	20.1	19.4	24.5	21.2	19.8
Phi-4 Abdin et al. (2024)	24.7	19.5	17.5	24.9	20.0	18.2
Molmo-2 Clark et al. (2026)	22.9	21.2	20.3	24.1	22.1	21.0
LLavaOneVision Li et al. (2024a)	27.5	25.0	22.7	32.8	30.1	27.5
InternVL-3.5 Wang et al. (2025)	29.7	25.0	20.4	33.1	26.7	23.3
Idefics-3 Laurençon et al. (2024)	<u>34.7</u>	30.6	<u>28.4</u>	38.7	37.3	35.6
Gemma-3n Team et al. (2025)	33.4	<u>30.7</u>	26.7	42.9	38.5	36.4
MiniCPM-V-4.5 Yao et al. (2024)	32.8	28.3	25.5	<u>46.1</u>	<u>40.2</u>	<u>38.1</u>
Qwen3-VL Bai et al. (2025)	**41.6**	**39.1**	**37.0**	**62.4**	**57.4**	**56.8**

Table 5: Fine-grained recognition accuracy across head, middle, and tail categories under different models used for obtaining distractors.

B CLIP-BASED RESULTS.

In this section, we compare the zero-shot results when using CLIP Radford et al. (2021) to mine the hard negatives against BioCLIP2 Stevens et al. (2024). From the results reported in Table 5, we can clearly see a significant performance disparity that is heavily dependent on the distractor generation strategy. As anticipated, employing BioCLIP2 produces substantially harder distractors than standard CLIP, resulting in a noticeably more rigorous evaluation setting and universally lower accuracy across all MLLMs.

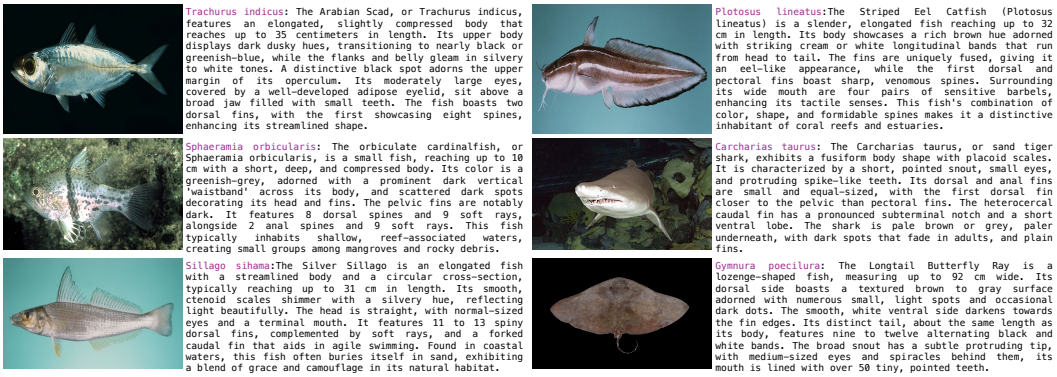

Figure 4: Examples of species description from FishNet2.0 obtained by the process explained in Section C.

C Species Description Generation

To construct the textual component of our knowledge base, we first collect species information from multiple authoritative sources. For fish species, we use **FishBase** Froese & Pauly (2025) as the primary source of morphological and ecological information, supplemented with information retrieved from **iNaturalist** iNaturalist contributors (2025), **WoRMS** WoRMS Editorial Board (2025), and **NOAA** NOAA Fisheries (2025). Aggregating information from multiple sources provides complementary descriptions of morphology, coloration, body shape, fins, markings, and other diagnostic characteristics.

The collected information is then provided to GPT-4o OpenAI (2024a), which generates a single species-level description. The prompt instructs the model to produce a concise description that emphasizes visually observable and discriminative characteristics useful for species identification while removing redundant and non-visual information. Specifically, the model is instructed to focus on coloration, body shape, fins, markings, size, and other distinguishing morphological traits, while excluding taxonomy, habitat, and other information that is not directly useful for visual recognition. The complete prompt is shown below.

Figure 4 presents an example species together with its generated description. Unlike an image caption, the generated description summarizes the species' characteristic appearance rather than describing a particular image, enabling it to capture visual attributes that may not be visible in any single photograph.

Prompt for Species Description Generation

You are given information about a fish species collected from multiple reliable and authoritative sources.
Your task is to write a concise, objective, and visually descriptive species-level summary that can be used for fine-grained species recognition.
Requirements:

- Focus on the physical appearance of the species.
- Include body shape, coloration, fins, markings, size, and other diagnostic anatomical features.
- Preserve all important identifying characteristics.
- Integrate information from all sources.
- Exclude habitat, behavior, taxonomy, and other non-visual information.
- Write the description as if explaining the appearance to someone who cannot see it.
- Produce a single paragraph of at most 100 words.

Species Information:
{`Retrieved species information`}

Method	Average IoU(%)				
	IoU50	IoU60	IoU70	IoU80	IoU90
YOLO-12 Tian et al. (2025)	**94.0**	**92.6**	**86.4**	**71.0**	**37.9**
LLavaNeXt Li et al. (2024b)	92.9	86.8	74.4	51.5	17.1
Phi-4 Abdin et al. (2024)	57.5	48.9	39.3	29.9	13.4
Molmo-2 Clark et al. (2026)	39.2	34.8	24.0	21.0	12.0
InternVL-3.5 Wang et al. (2025)	92.1	87.1	76.8	58.2	29.6
LLavaOneVision Li et al. (2024a)	84.9	71.7	48.1	18.0	2.0
MiniCPM-V-4.5 Yao et al. (2024)	92.8	86.7	74.0	50.0	15.7
Gemma-3n Team et al. (2025)	65.6	45.1	22.5	8.2	2.0
Qwen3-VL Bai et al. (2025)	<u>93.7</u>	<u>89.5</u>	<u>80.5</u>	<u>62.9</u>	<u>34.1</u>

Table 6: Performance of various MLLMs on the detection task. The performance is reported using IoU thresholds (50–90).

Method	Frequent Species (%)					
	Tail End	Fin	Tail Start	Body	Mouth	Eye
YOLO-12 Tian et al. (2025)	**30.3**	**15.9**	**46.4**	**46.2**	**45.7**	**44.0**
LLavaNeXt Li et al. (2024b)	1.0	1.0	1.1	1.7	1.3	1.5
Phi-4 Abdin et al. (2024)	1.3	1.3	1.3	2.1	1.4	1.6
Molmo-2 Clark et al. (2026)	1.3	5.2	14.3	15.5	7.6	1.7
InternVL-3.5 Wang et al. (2025)	15.8	12.7	18.1	27.2	33.4	22.0
LLavaOneVision Li et al. (2024a)	1.4	5.4	19.3	10.5	5.6	1.9
MiniCPM-V-4.5 Yao et al. (2024)	4.0	4.0	18.6	26.1	12.2	10.4
Gemma-3n Team et al. (2025)	3.2	11.5	12.3	21.5	11.4	5.2
Qwen3-VL Bai et al. (2025)	<u>29.7</u>	<u>13.4</u>	<u>27.7</u>	<u>40.0</u>	<u>45.3</u>	<u>42.6</u>

Table 7: Performance of various MLLMs on the key-point localization task. The performance is reported using PCK @0.1 Novotny et al. (2018); Yang & Ramanan (2013).

D Detailed detection and key-point results

In Table 6, we report detailed detection results for IoU@50-IoU@90 across all evaluated MLLMs, along with a trained YOLO-based baseline. The results show that although MLLMs generally achieve lower IoU scores than YOLO, the best-performing model (Qwen3-VL) reaches comparable performance, demonstrating its strong capability to localize fish within images.

Similarly, in Table 7, we report key-point detection results across all six body parts. We observe a similar trend: while the overall performance of MLLMs remains relatively modest, Qwen3-VL achieves the best performance and is comparable to a trained YOLO-based baseline, highlighting its ability to capture fine-grained visual details.

E Segmentation

Obtaining segmentation masks is often time-consuming and labor-intensive. To address this, we employed a semi-automated pipeline to generate segmentation masks for FishNet2.0. Specifically, we used keypoints collected for each image as prompts to guide the Segment Anything Model Kirillov et al. (2023), enabling it to better infer the approximate structure of the target object. This keypoint-guided approach proved highly effective. Some examples are shown in Figure 5.

To evaluate the effectiveness of this approach, we manually assessed the generated segmentation mask on the test set comprising the head, middle, and tail classes. In this test, the annotators were asked whether the provided segmentation mask fully covered the marine species, including all parts of its body. From all 31,885 images of the test set, 24,278(76%) were considered to be perfect by users, and the remaining ones captured most of the body but often missed parts like tails and fins, as shown in Figure 6. This clearly shows that our approach is highly effective for obtaining automated segmentation masks.

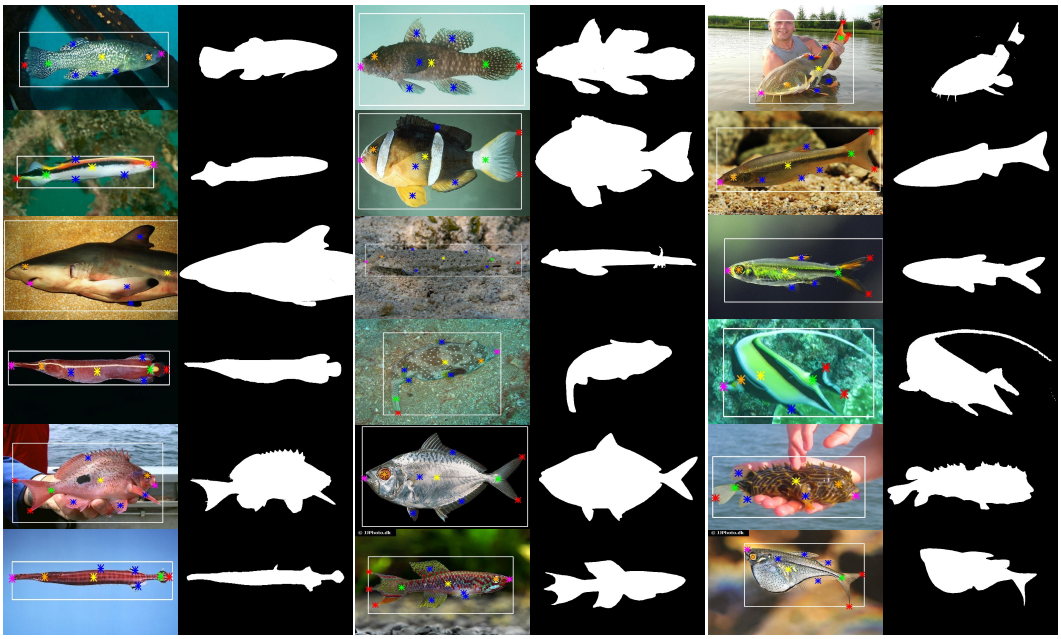

Figure 5: We show the same images from Figure 3 with segmentation masks obtained from our automated pipeline using key-points as supervision.

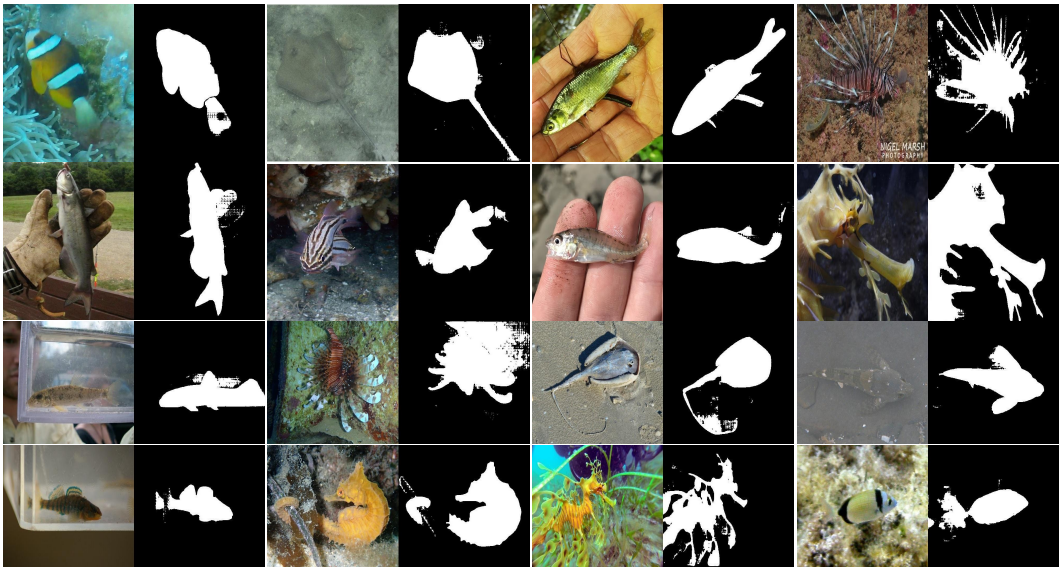

Figure 6: We show the samples with erroneous segmentation masks obtained from our automated pipeline using key-points as supervision.

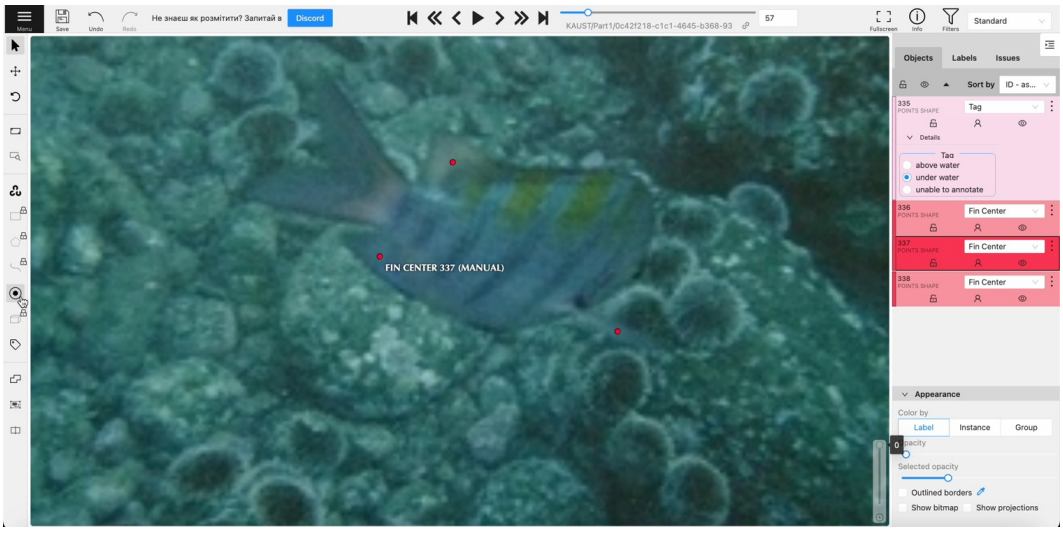

Figure 7: This figure illustrates the interface used for part location annotation. Here, the user has zoomed in on the image to accurately label the fin's center. Two fins have already been labeled, as indicated by the red-colored dot at the fin centers. The entire labeling process is efficient and user-friendly.

F Data Collection Details.

To enable efficient and accurate collection of data, we worked with an annotation service provider.[1]. The custom-designed interface was developed to facilitate the collection and verification of part location and segmentation masks. We show the interface in Figure 7.

G Limitations.

Despite the extensive coverage and high-quality annotations provided by FishNet2.0, several limitations remain:

- While FishNet2.0 includes a large number of species and diverse annotations, the dataset remains constrained by the available imagery. Certain ecological regions and rare species remain underrepresented, limiting the generalizability of models trained on this data to truly global scenarios.
- Prompt-based evaluation for MLLMs can be highly sensitive to the structure and content of the prompt, which may introduce bias in comparisons.
- Underwater imagery presents extreme domain shifts (lighting, turbidity, occlusion) that remain difficult for both MLLMs and task-specific models. Performance in these conditions, while informative, may not fully reflect real-time field performance.

[1] https://labelyourdata.com/