UWBench: A Comprehensive Vision-Language Benchmark for Underwater Understanding

Da Zhang [ID], *Student Member, IEEE,* Chenggang Rong [ID], Bingyu Li [ID], Feiyu Wang [ID], Zhiyuan Zhao [ID], Junyu Gao [ID], *Member, IEEE,* and Xuelong Li [ID]†, *Fellow, IEEE*

Abstract—Large vision-language models (VLMs) have achieved remarkable success in natural scene understanding, yet their application to underwater environments remains largely unexplored. Underwater imagery presents unique challenges including severe light attenuation, color distortion, and suspended particle scattering, while requiring specialized knowledge of marine ecosystems and organism taxonomy. To bridge this gap, we introduce UWBench, a comprehensive benchmark specifically designed for underwater vision-language understanding. UWBench comprises 15,003 high-resolution underwater images captured across diverse aquatic environments, encompassing oceans, coral reefs, and deep-sea habitats. Each image is enriched with human-verified annotations including 15,281 object referring expressions that precisely describe marine organisms and underwater structures, and 124,983 question-answer pairs covering diverse reasoning capabilities from object recognition to ecological relationship understanding. The dataset captures rich variations in visibility, lighting conditions, and water turbidity, providing a realistic testbed for model evaluation. Based on UWBench, we establish three comprehensive benchmarks: detailed image captioning for generating ecologically informed scene descriptions, visual grounding for precise localization of marine organisms, and visual question answering for multimodal reasoning about underwater environments. Extensive experiments on state-of-the-art VLMs demonstrate that underwater understanding remains challenging, with substantial room for improvement. Our benchmark provides essential resources for advancing vision-language research in underwater contexts and supporting applications in marine science, ecological monitoring, and autonomous underwater exploration. Our code and benchmark are available at UWBench.

Index Terms—Underwater; Image Caption; Visual Grounding; Visual Question Answering; Multimodal Reasoning

Figure 1: Challenges for VLMs for understanding underwater scenes.

I. INTRODUCTION

RECENT years have witnessed remarkable advances in large vision-language models (VLMs) [1]–[3], which have demonstrated unprecedented capabilities in understanding and reasoning about visual content through natural language [4], [5]. State-of-the-art VLMs such as GPT-5, GLM-4.5, and InternVL have achieved impressive performance across a wide range of tasks, including detailed image captioning [6]–[8], complex visual question answering [9], [10], and precise visual grounding [11], [12]. These models have enabled significant progress by effectively bridging the gap between vision and language, and have been successfully deployed in various real-world applications, ranging from autonomous navigation [13], [14] to content moderation [15], [16] and assistive technologies [17], [18]. Much of this progress can be attributed to the availability of large-scale, high-quality datasets captured in terrestrial environments [19]–[21], which have provided VLMs with rich and diverse data for training and evaluation in natural scenes. However, a critical question remains largely unanswered: **can current VLMs effectively understand and interpret imagery from challenging underwater environments, where visual conditions differ fundamentally from terrestrial scenarios?**

This question carries substantial importance given the vital role of underwater observation in marine science, ecological conservation, and resource management [22]–[24]. As illustrated in Figure 1, applying VLMs to underwater imagery introduces three fundamental challenges that distinguish this domain from conventional vision-language tasks:

- **Degraded visual features**. Underwater scenes exhibit variable illumination with rapid light attenuation, wavelength-dependent color distortion, and fluctuating turbidity from

Da Zhang, Chenggang Rong, and Junyu Gao are with the Institute of Artificial Intelligence (TeleAI), China Telecom, China and also with the School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China. (E-mail: dazhang@mail.nwpu.edu.cn; rongcg5620@mail.nwpu.edu.cn; gjy3035@gmail.com).

Bingyu Li, Feiyu Wang, Zhiyuan Zhao, and Xuelong Li is with the Institute of Artificial Intelligence (TeleAI), China Telecom, China. (E-mail: libingyu0205@mail.ustc.edu.cn; wangfy25@m.fudan.edu.cn; tuzixini@gmail.com; xuelong_li@ieee.org).
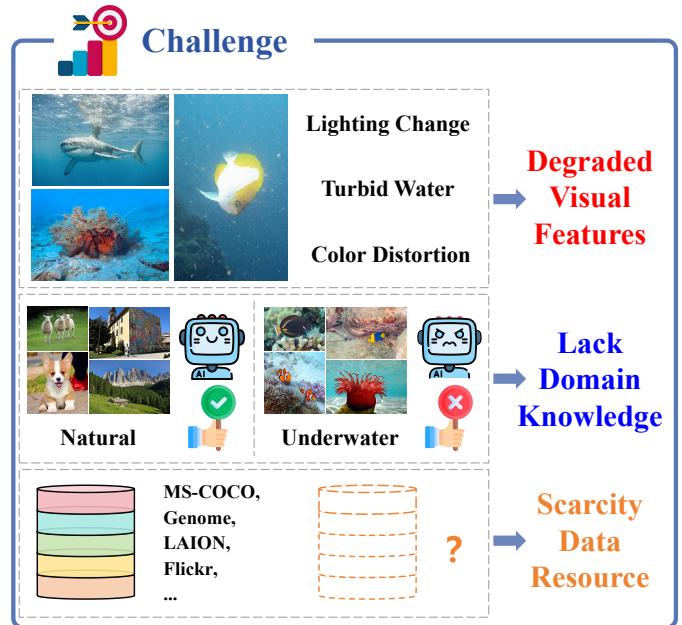
Object Referring
ID=0: The aircraft wreck lying on the sandy seabed, stretching from the left edge toward the center with a visible propeller at its nose.
ID=1: The diver near the center who is the top-most in the frame, positioned beside the wrecked aircraft's wing.
ID=2: The small diver in the bottom-right corner of the image.

Visual Question Answer
Question1: What is the large structure on the seabed? _Answer_: Wreckedaircraft.   *Question2*: How many divers are visible? _Answer_: 2.
Question3: Are small fish visible in the scene? _Answer_: Yes.   *Question4*: Where is the top-most diver? _Answer_: Center.
Question5: What is the orientation of the aircraft's wing across the frame? _Answer_: Horizontal.
Question6: What is the dominant substrate on the bottom? _Answer_: Sand.   *Question7*: Is there a wreckedaircraft present? _Answer_: Yes.

Detailed Captioning
Clear blue water reveals a sandy seabed with scattered rocky patches and an intact aircraft wreck resting on the bottom. The wreckedaircraft has an exposed nose and propeller, with a long wing extending horizontally across the foreground. Two scuba divers are present: one near the center above the wing and another smaller diver in the bottom-right. Trails of bubbles rise into the water column, and small fish are scattered around the scene. The view highlights the scale of the wreck relative to the divers across the open seabed.
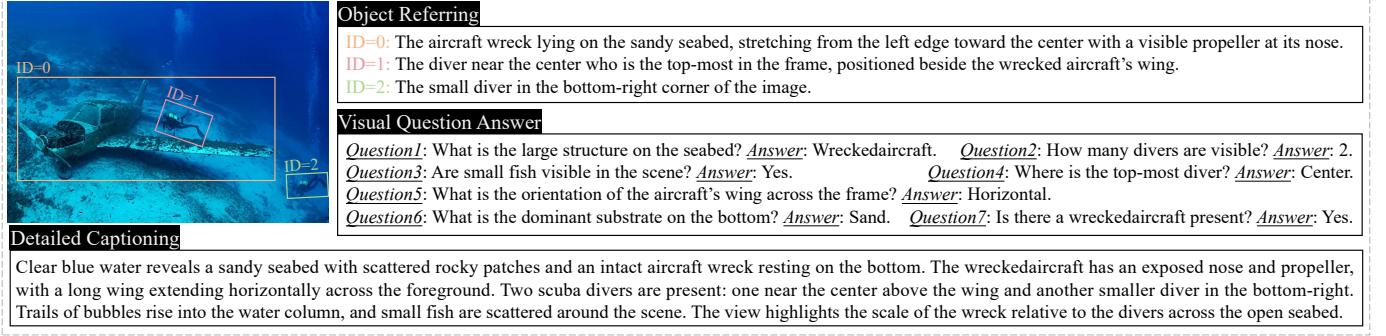
Figure 2: Examples of an image and corresponding annotations in UWBench dataset. Our annotations include object referring, visual question answering, and detailed captions.

suspended particles, resulting in reduced contrast and limited visibility that confound terrestrial-trained approaches [25], [26].

- **Lack of scientific domain knowledge**. Accurate underwater interpretation requires specialized knowledge of marine taxonomy, organism morphology, behavioral patterns, and ecological relationships, representing a substantial gap for systems trained on common terrestrial scenarios [27].
- **Scarcity of data resources**. Existing vision-language datasets focus predominantly on terrestrial scenarios, while underwater-specific datasets lack the multimodal annotations necessary for comprehensive understanding [28].

To address these challenges and advance vision-language understanding in underwater contexts, we introduce UWBench, the first large-scale benchmark dataset specifically designed for comprehensive underwater image understanding. UWBench comprises 15,003 high-resolution images spanning diverse underwater environments including shallow reefs, mid-depth zones, and deep-sea habitats, with rich variations in water clarity, illumination conditions, and marine biodiversity that reflect realistic observation scenarios. Each image is enriched with meticulously crafted annotations verified by marine biology experts: 15,003 detailed captions providing ecologically informed scene descriptions, 15,281 object referring expressions for precise organism localization, and 124,983 question-answer pairs assessing diverse reasoning capabilities from basic recognition to complex ecological understanding (Figure 2). Through a collaborative approach combining automated GPT-5 generation with rigorous human verification, we ensure scientific accuracy in taxonomic classifications, behavioral descriptions, and environmental context. Based on these comprehensive annotations, UWBench establishes three interconnected evaluation benchmarks for detailed image captioning, visual grounding, and visual question answering, enabling systematic assessment of vision-language models across diverse underwater understanding capabilities and supporting development of specialized systems for marine research and ecological monitoring applications.

We conduct comprehensive experiments on UWBench to evaluate the performance of existing VLMs under three evaluation settings. Specifically, we assess multiple state-of-the-art models including closed-source systems such as GPT-4o [29], GPT-5, and Gemini, as well as open-source models including InternVL3.5 series [30], Qwen2.5-VL series [31], Qwen3 series [32], and GLM-4 series [33]. Experimental results demonstrate that underwater image understanding remains highly challenging even for state-of-the-art models. Compared to their performance on terrestrial benchmarks, all models exhibit substantial performance degradation on UWBench, highlighting the domain gap between terrestrial and underwater vision-language understanding. Through comprehensive analysis including task-specific evaluation, attribute-based assessment, and qualitative error analysis, we identify key factors contributing to performance decline: difficulty in recognizing degraded visual features under poor visibility conditions, insufficient domain knowledge for accurate taxonomic identification and ecological reasoning, and challenges in fine-grained spatial reasoning required for precise object localization in complex underwater scenes. These findings underscore the necessity of specialized benchmarks like UWBench and illuminate potential avenues for future improvement in underwater vision-language understanding. The key contributions of our work are summarized as follows:

- We construct UWBench, the first large-scale benchmark dataset specifically designed for underwater vision-language understanding, comprising 15,003 high-resolution images with comprehensive human-verified annotations including detailed captions, object referring expressions, and question-answer pairs.
- Three comprehensive evaluation benchmarks are established for detailed image captioning, visual grounding, and visual question answering, providing standardized protocols for systematic assessment of vision-language models in underwater contexts.
- We provide detailed insights into factors limiting current model performance and identify promising directions for future research, contributing to the development of specialized vision-language systems capable of supporting marine research and ecological monitoring applications.

II. RELATED WORKS

A. Vision-Language Models and Benchmarks

The rapid advancement of large vision-language models has revolutionized multimodal understanding across diverse applications [34], [35]. Foundational works such as CLIP

[36] demonstrate the effectiveness of contrastive learning for aligning visual and textual representations at scale [37], [38]. Building upon these foundations, recent generative VLMs including LLaVA [39], InternVL [40], Qwen-VL [41], and GLM [33] have achieved remarkable performance by integrating powerful vision encoders with large language models, enabling sophisticated reasoning about visual content through natural language. These models excel at tasks requiring joint understanding of vision and language, including image captioning [42], visual question answering [9], and visual grounding [43].

The development of comprehensive benchmarks has been instrumental in advancing VLM capabilities [1], [44], [45]. Datasets such as COCO [46] provide large-scale annotations for object detection and image captioning, while VQAv2 [47] and OK-VQA [48] introduce challenging visual question answering scenarios requiring reasoning and commonsense knowledge. RefCOCO [49] and its variants establish benchmarks for referring expression comprehension and visual grounding tasks. More recently, specialized benchmarks like MMBench [50] and SEED-Bench [51] provide holistic evaluation frameworks assessing multiple dimensions of multimodal understanding. However, these benchmarks predominantly focus on everyday terrestrial scenarios, leaving significant gaps in specialized domains. UWBench addresses this limitation by providing the first comprehensive vision-language benchmark specifically designed for underwater environments, enabling systematic evaluation of VLM capabilities in challenging aquatic contexts.

B. Underwater Image Datasets and Benchmarks

Underwater vision research has progressed through the development of specialized datasets targeting various perception tasks [52]–[54]. Early efforts focused on low-level image enhancement [55] and restoration [56]. The UIEBD dataset [52] provides paired underwater and reference images for enhancement algorithm evaluation, while EUVP [53] offers a large-scale collection for underwater image restoration and color correction. More recent datasets like LSUI [57] and RUIE [58] extend this work with diverse underwater scenes captured under varying environmental conditions.

For object-level understanding, several datasets have been proposed to advance underwater object detection and segmentation. The URPC dataset [59] introduces annotations for marine organisms in challenging underwater conditions, while Brackish [60] includes diverse aquatic species for detection and tracking. SUIM [61] provides semantic segmentation annotations for underwater imagery, and more recently, UIIS [62] offers instance-level segmentation masks for marine objects. The USOD10K dataset [63] marks significant progress by providing 10,255 images with pixel-wise annotations for underwater salient object detection, covering 70 object categories across diverse underwater environments. Building upon this foundation, USIS16K [64] further expands the scale to 16,151 images with 158 categories, offering comprehensive instance segmentation annotations.

For temporal understanding, tracking datasets such as UTB180 [65] and VMAT [66] provide annotated video sequences for single object tracking in underwater scenarios.

More recently, WebUOT-1M [67] introduces a million-scale benchmark with 1,500 video sequences spanning 408 categories, significantly advancing underwater object tracking research. These datasets have substantially contributed to advancing underwater computer vision. However, they primarily focus on traditional vision tasks without incorporating vision-language annotations. In contrast, UWBench bridges this critical gap by providing comprehensive multimodal annotations including detailed captions, referring expressions, and question-answer pairs, thereby enabling the development and evaluation of vision-language models specifically tailored for underwater understanding. Table I provides a comprehensive comparison of representative underwater datasets across different task categories.

C. Underwater Visual Analysis Tasks

Research in underwater visual analysis encompasses three primary directions: image quality enhancement [73], object recognition and localization [74], and environmental modeling [75]. For image enhancement, numerous methods address the degradation caused by light absorption and scattering in aquatic environments [76]. Physics-based approaches such as Dark Channel Prior and its underwater adaptations model the image formation process to restore color and contrast [77]. Learning-based methods including WaterGAN [78], UGAN [79], and FUnIE-GAN [53] leverage adversarial training for end-to-end enhancement. More recent works employ diffusion models [80] and transformer architectures [57], demonstrating improved restoration quality on challenging underwater imagery.

Object recognition in underwater environments has seen substantial progress through specialized detection and classification frameworks [81]. Methods tailored for marine species recognition address challenges including small object scales, camouflage, and inter-species similarity. The DeepFish dataset [68] supports instance segmentation and classification of fish species, while FishNet [69] provides a large-scale benchmark for fish recognition and functional trait prediction covering over 17,000 species. For coral reef analysis, specialized approaches including CoralSCOP [82] and semantic segmentation methods enable automated monitoring of reef health and biodiversity. Recent vision-language models like MarineInst [83] and MarineGPT [84] begin exploring multimodal understanding for marine imagery, incorporating both visual recognition and language description capabilities. While these advances demonstrate significant progress in individual tasks, UWBench uniquely enables integrated evaluation of vision-language capabilities across multiple complementary tasks, providing a holistic framework for assessing comprehensive underwater scene understanding.

D. Multimodal Understanding in Specialized Domains

The application of vision-language models to specialized domains beyond everyday scenarios has gained increasing attention. In medical imaging, datasets such as MIMIC-CXR [85] and PadChest [86] provide radiograph-report pairs enabling the development of models for medical visual question answering and report generation. Models like Medunifier

Table I: Comprehensive comparison of underwater datasets and benchmarks. UWBench is the first image-based benchmark providing integrated vision-language annotations including detailed captions, referring expressions, and visual question answering, alongside traditional detection and segmentation tasks.

Dataset	Year	Type	Images	Videos	Categories	Annotation Quality	Enhancement	Detection	Segmentation	Tracking	Caption	Referring	VQA
UIEBD [52]	2019	Image	950	-	-	Manual	✓	-	-	-	-	-	-
EUVP [53]	2020	Image	12K	-	-	Manual	✓	-	-	-	-	-	-
LSUI [57]	2023	Image	5K	-	10	Manual	✓	-	-	-	-	-	-
RUIE [58]	2021	Image	4K	-	-	Semi-auto	✓	-	-	-	-	-	-
URPC [59]	2020	Image	6K	-	4	Manual	-	✓	-	-	-	-	-
Brackish [60]	2020	Image	14K	89	8	Manual	-	✓	-	✓	-	-	-
SUIM [61]	2020	Image	1.6K	-	8	Manual	-	-	✓	-	-	-	-
UIIS [62]	2023	Image	4.6K	-	7	Manual	-	-	✓	-	-	-	-
USOD10K [63]	2023	Image	10.3K	-	70	Manual	-	-	✓	-	-	-	-
USIS16K [64]	2025	Image	16.2K	-	158	Manual+Expert	-	✓	✓	-	-	-	-
DeepFish [68]	2022	Image	4.5K	-	20	Manual	-	✓	✓	-	-	-	-
FishNet [69]	2023	Image	17K	-	17K	Manual	-	✓	-	-	-	-	-
UTB180 [65]	2018	Video	-	180	-	Manual	-	-	-	✓	-	-	-
VMAT [66]	2023	Video	57K	33	17	Manual	-	-	-	✓	-	-	-
WebUOT-1M [67]	2024	Video	1M	1.5K	408	Semi-auto	-	-	-	✓	-	-	-
DRUVA [70]	2023	Video	6K	20	20	Manual	✓	-	-	-	-	-	-
MarineInst	2024	Image	2.4M	-	-	Auto-generated	-	-	✓	-	✓	-	-
UVLM [71]	2025	Video	860K	2.1K	419	GPT+Human	-	-	-	-	-	-	✓
CoralVQA [72]	2025	Image	12K	-	-	GPT+Human	-	-	-	-	-	-	✓
UWBench (Ours)	**2025**	**Image**	**15K**	**-**	**158**	**GPT+Expert**	-	✓	✓	-	✓	✓	✓

[87] and BiomedCLIP [88] demonstrate that domain-specific pretraining significantly improves performance on medical tasks compared to generic VLMs. Similarly, in remote sensing, the VRSBench dataset [89] provides comprehensive annotations including detailed captions, object referring, and visual question answering for aerial imagery.

These specialized domain applications reveal consistent patterns: domain-specific visual characteristics, specialized vocabulary and knowledge requirements, and limited availability of annotated multimodal data present significant challenges for general-purpose VLMs [90]–[92]. Successful approaches typically involve constructing large-scale domain-specific datasets, incorporating expert knowledge into annotation processes, and adapting model architectures or training procedures to address domain-specific challenges [93]. The underwater domain shares these characteristics, exhibiting unique visual degradation patterns, requiring extensive marine biological knowledge, and lacking comprehensive vision-language resources. UWBench follows best practices established in other specialized domains while addressing the unique challenges of underwater environments, providing ecologically informed annotations verified by marine experts and supporting multiple interconnected tasks essential for comprehensive underwater scene understanding.

III. UWBENCH CONSTRUCTION

A. Overview

The construction of UWBench addresses three fundamental challenges: capturing underwater visual degradation, incorporating marine biological knowledge, and ensuring annotation quality through expert verification. Our pipeline consists of five stages: data collection, attribute extraction, prompt design, GPT-5 assisted generation, and expert verification. This approach produces 15,003 high-resolution images with 15,281 referring expressions and 124,983 QA pairs, all verified by marine biology experts. Figure 3 illustrates the complete annotation pipeline, demonstrating the integration of automated generation and human expertise throughout the construction process.

B. Data Source and Image Collection

We adopt a multi-source collection strategy drawing from three complementary sources:

- Web-based collection from Google, Bing, and Flickr targeting marine organisms and underwater habitats, yielding diverse imagery across global locations;
- Samples from existing underwater detection and segmentation datasets providing high-quality instance-level annotations with precise boundaries and taxonomic classifications [61], [62], [64], [94];
- In-situ images from underwater robotic operations capturing authentic challenges including occlusion and backscatter.

This yields over 35,000 candidate images.

We implement rigorous quality control through systematic filtering by volunteers with marine science backgrounds. The process eliminates severe quality issues, duplicates, watermarked images, and aquarium scenes. Selection criteria emphasize diversity across habitat types (coral reefs, open ocean, kelp forests), water conditions (clear, turbid, low-light), object scales, and scene complexity. Through careful curation, we select 15,003 high-quality images with instance segmentation masks covering 158 object categories.

C. Attribute Extraction

Attribute extraction transforms segmentation annotations into structured information for language model generation. We extract image-level attributes including resolution, habitat type, water quality indicators, and illumination conditions. Object-level extraction derives taxonomic information (species name, category), spatial information (bounding box, position, relationships), and morphological attributes (size, aspect ratio, visibility status, uniqueness within category). This rich representation provides the foundation for generating diverse vision-language annotations.

D. Prompt Engineering

We design comprehensive instructions guiding GPT-5 to produce scientifically accurate annotations across three tasks.
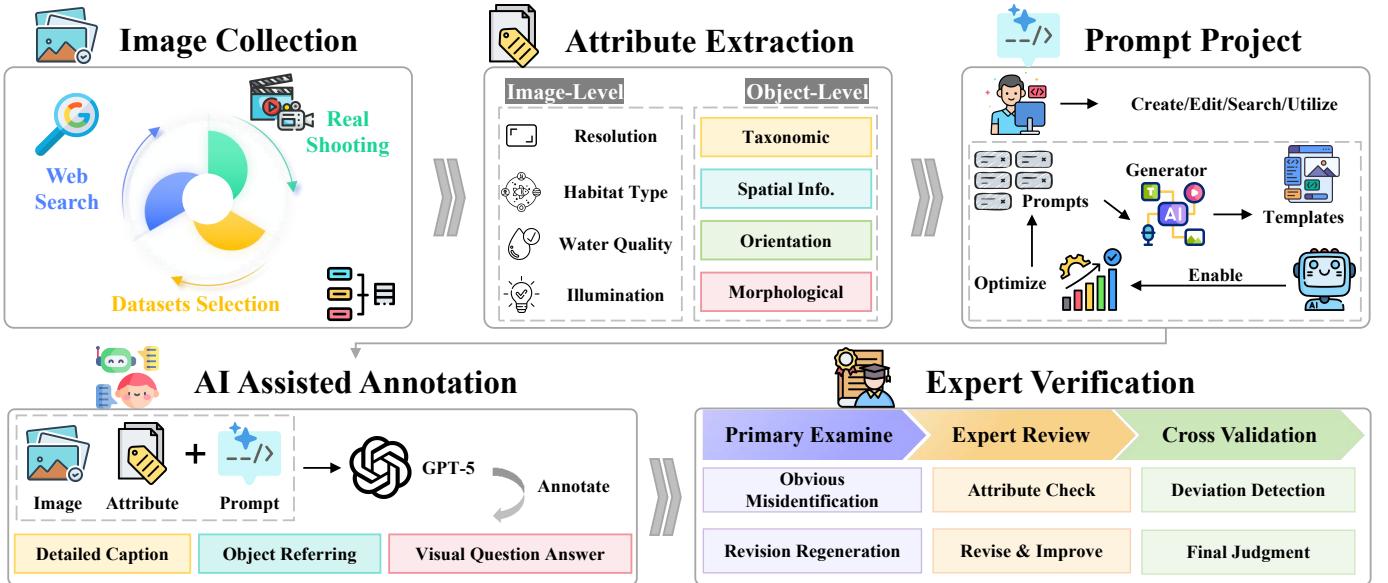
Figure 3: Overview of UWBench Construction. The pipeline starts with multi-source underwater image acquisition—including web search, public dataset selection, and real-world in-situ shooting—followed by refined attribute extraction that structurally organizes image and object-level features such as resolution, habitat type, water quality, illumination, taxonomy, spatial, and morphological information. Comprehensive prompt engineering then guides GPT-5 to automatically generate detailed captions, distinctive referring expressions, and diverse visual QA pairs. Finally, annotation quality is verified and enhanced through a three-stage process—primary review, expert assessment, and cross-validation—resulting in a scientifically rigorous, ecologically informative, and broadly representative underwater vision-language dataset.

Our prompts explicitly incorporate marine biological knowledge requirements: proper taxonomic nomenclature, morphological characteristics, behavioral patterns, and ecological context.

For image captioning, prompts instruct GPT-5 to begin with habitat overview including substrate, vegetation, water clarity, and structures, then characterize marine organisms with species identification, morphological features, and positioning. Descriptions use only visually verifiable attributes, avoiding speculation about depth or temperature.

For referring expressions, prompts guide generation of unambiguous descriptions using distinctive visual attributes: morphological features, texture patterns, relative size, coloration, and spatial relationships to visible structures. Each expression must independently identify its target without ordinal references.

For visual QA, prompts elicit diverse question types spanning object identification, existence verification, quantity, shape, size, position, orientation, and scene classification. We generate 3-10 pairs per image with concise answers, prohibiting questions about invisible attributes and requiring exact category names from the taxonomy.

E. AI Assisted Annotation Generation

We employ GPT-5 to generate initial annotations by constructing multimodal prompts combining images, extracted attributes, and task instructions. GPT-5 outputs standardized JSON containing captions, object-referring pairs, and QA pairs.

- **Iterative refinement**: Recursively invoking GPT-5 up to five times to eliminate uncertain language and ambiguous descriptions.

- **Content filtering**: Removing self-answering questions, tautological reasoning, hallucinated information, and questions about invisible properties.
- **Format validation**: Ensuring proper JSON structure, normalized coordinates, and standalone referring expressions.

These mechanisms produce high-quality initial annotations for expert verification.

F. Expert Verification

We implement a rigorous three-stage verification process with marine biology experts:

- **Stage 1: General quality assessment** by trained annotators identifies taxonomic errors, morphological inconsistencies, logical errors, linguistic issues, and guideline violations. Flagged annotations undergo revision or regeneration.
- **Stage 2: Domain expert review** by senior marine biologists validates taxonomic accuracy, ecological plausibility, attribute precision, scientific terminology, and question appropriateness. Experts can modify annotations, add context, or request regeneration.
- **Stage 3: Cross-validation** involves 2,000 images reviewed by multiple experts to assess agreement and identify systematic biases. This reveals error patterns, informs prompt refinement, and establishes quality benchmarks. Senior biologists adjudicate disagreements.

The verification process requires approximately 150 seconds per image, totaling over 600 hours. This investment ensures scientifically rigorous annotations with expert-verified taxo-

nomic identifications, ecologically informed descriptions, and diverse question-answer pairs maintaining factual accuracy.

IV. UWBENCH STATISTICS

A. Overview

UWBench contains 15,003 underwater images with comprehensive human-verified annotations comprising 15,003 detailed captions, 15,281 object referring expressions, and 124,983 visual question-answer pairs. The dataset covers 158 underwater object categories spanning diverse taxonomic groups including marine fishes, shellfish, marine animals, underwater facilities, and debris. All annotations integrate automated GPT-5 generation with rigorous expert verification to ensure scientific accuracy and ecological validity. The detailed statistical results are shown in Table II.

Table II: Statistics of UWBench.

Caption	Total Captions	Object Referring	Total Objects	VQA	Total Pairs
	15,003		15281		124,983
	Avg. Words		Categories		Question Types
	68.60		158		301
	Avg. Sentences		Avg. Objects		Avg. Q Num
	4.35		1.02		8.33
	Unique Words		Max Object		Avg. Q Length
	3560		3		6.91 words
	1st Common Word		Unique (%)		Avg. A Length
	water (14,905)		99.20%		1.13 words

B. Detailed Caption

UWBench captions provide comprehensive descriptions integrating underwater environmental characteristics with detailed object-specific information. Each caption follows a structured format beginning with scene-level context including substrate type such as sandy seabed, rocky bottom, or coral reef, water quality indicators encompassing clarity, turbidity, and color cast, visible vegetation or algae presence, and artificial structures when present. Following this environmental overview, captions describe prominent marine organisms with attention to species identification, morphological characteristics, relative positioning, and quantity where appropriate. Descriptions emphasize visually verifiable attributes including texture patterns such as spotted or striped markings, shape characteristics, relative size comparisons, and spatial relationships expressed using image-relative terms. Captions explicitly avoid speculation about invisible attributes including exact depth, water temperature, temporal information, or behavioral patterns unless clearly observable. Statistical analysis reveals captions average 68.60 words across 4.35 sentences, ranging from 30 to 120 words. The dataset vocabulary encompasses 3,560 unique words, with frequent underwater-specific terms including coral, reef, and fish. This reflects the dual focus on environmental context and biological content characteristic of underwater understanding. A summary of these caption statistics is detailed in Figure 4.

C. Object Referring

UWBench provides referring expressions for 15,281 objects across 158 categories, carefully crafted to enable unambiguous identification without spatial deixis or ordinal references. Each referring sentence independently identifies its target object using distinctive visual attributes rather than relying on positional descriptors or sequential ordering. The expressions emphasize species-specific morphological features, texture patterns including spotted, striped, or encrusted characteristics, relative size compared to other visible organisms, coloration
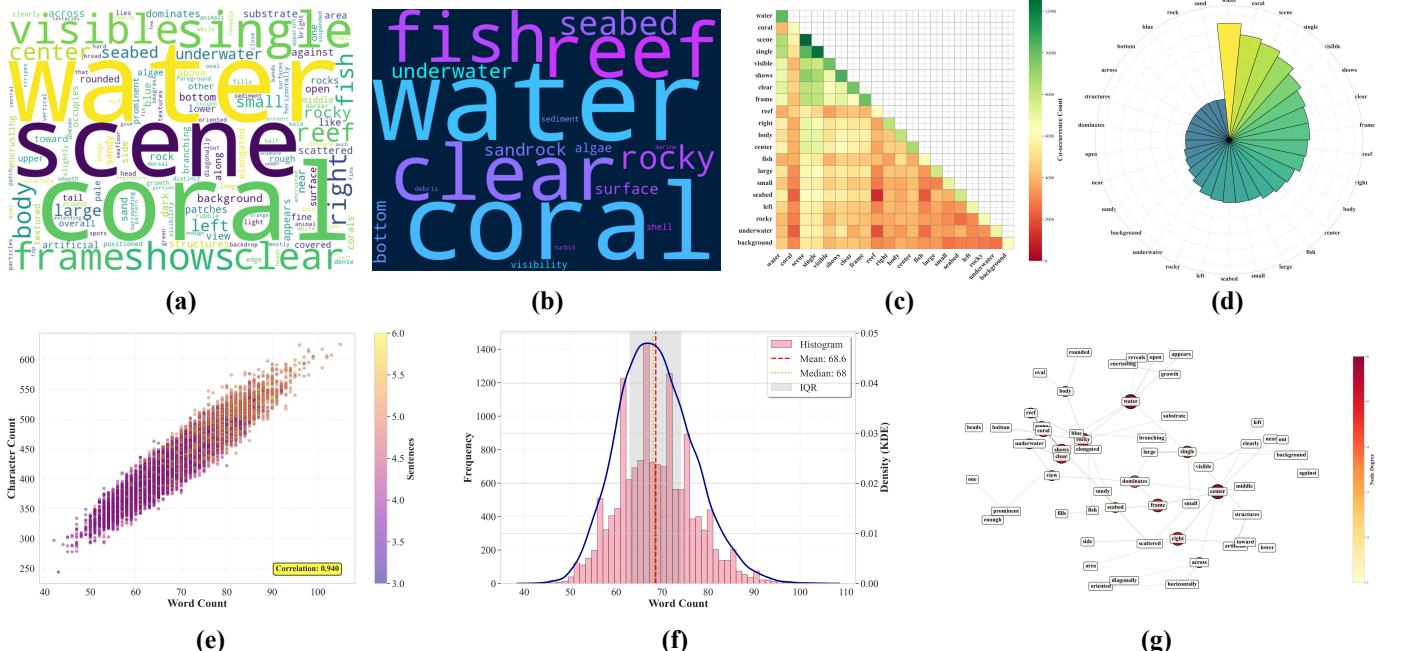

Figure 4: Statistics of the UWBench caption dataset. (a) Overall Vocabulary Word Cloud. (b) Underwater Domain-Specific Terms. (c) Top 20 Words Co-occurrence Heatmap. (d) Top 30 Words Radial View. (e) Word vs Character Count Scatter. (f) Word Count Distribution. (g) Word Association Network.
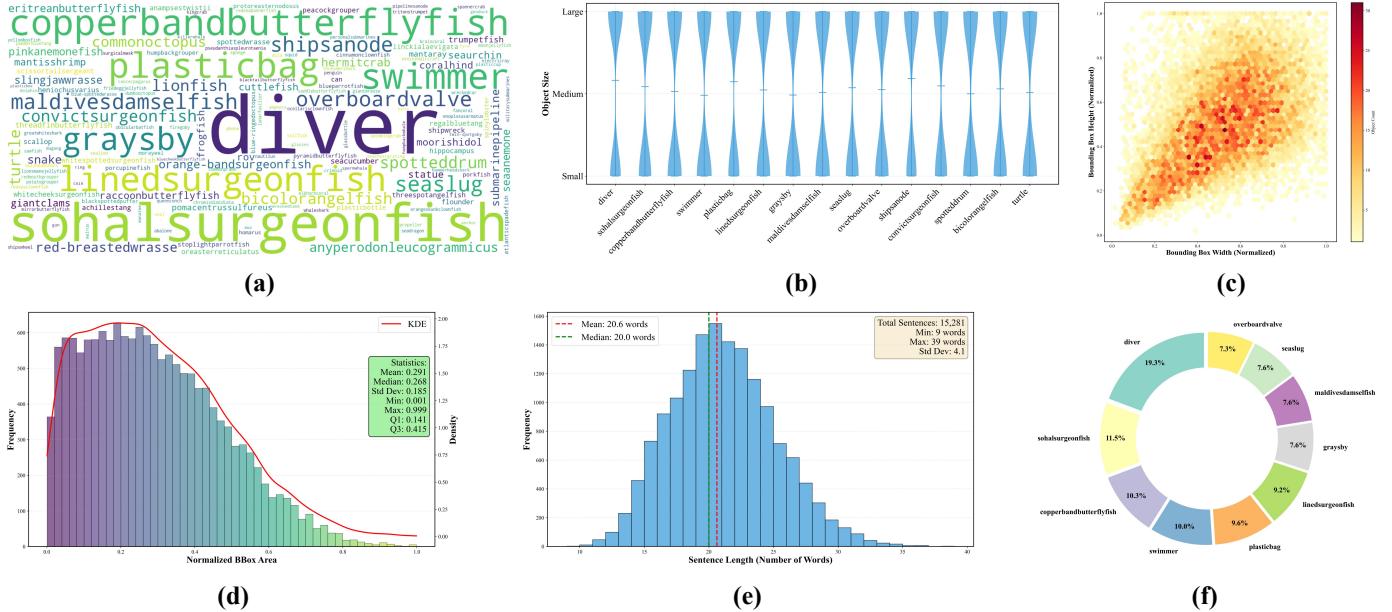
Figure 5: Statistics of object referring of UWBench. (a) Word Cloud of Categories. (b) Object Size by Category (Top 15). (c) Density Heatmap for BBox Dimensions. (d) BBox Area Distribution. (e) Referring Sentence Length Distribution. (f) Top 10 Categories Pie Chart (Relative Proportion).

when reliably discernible under underwater lighting conditions, and spatial relationships expressed through adjacency to clearly visible structures such as rocks, coral formations, or artificial objects. Referring expressions average 20.6 words, ranging from 8 to 40 words. The dataset prioritizes unique objects comprising approximately 60% of expressions, while remaining expressions address challenging multi-instance scenarios. Category distribution reflects underwater ecosystem composition, which ensures diverse representation across taxonomic groups and object types encountered in underwater observation scenarios. Figure 5 provides a summary of UWBench.

D. Visual Question Answering

UWBench provides 124,983 question-answer pairs covering a variety of reasoning types essential for underwater understanding. The questions span object category identification, existence detection, quantity counting, attribute recognition (such as color and shape), spatial relationships, and scene classification (e.g., coral reefs, sandy seabeds). Additional questions address substrate and material identification, orientation assessment, and complex reasoning that integrates visual and ecological knowledge, providing a comprehensive evaluation of underwater reasoning abilities. Answer distribution emphasizes definitive responses with 87.1% single-word answers, 12.4% two-word phrases, and 0.5% three-plus words. Question types extend beyond conventional categories to include underwater-specific reasoning encompassing substrate identification, water quality assessment, organism-substrate relationships, and ecological context inference. We show the statistics of question-answer pairs in Figure 6.

V. UWBench Evaluation

A. Benchmark Overview

We construct three distinct evaluation tasks to comprehensively assess vision-language capabilities in underwater contexts. UWBench-Cap requires generating comprehensive descriptions for underwater images, capturing environmental characteristics, marine organisms, and their ecological relationships. UWBench-Ref involves identifying and localizing specific underwater objects based on textual descriptions, requiring precise understanding of morphological attributes and spatial relationships. UWBench-VQA aims to answer diverse questions about visual content in underwater scenes, spanning from basic object recognition to complex ecological reasoning.

To facilitate rigorous evaluation, we partition UWBench into non-overlapping training and test splits. The training set comprises 10,454 images with 10,454 captions, 10,654 object referring expressions, and 87,055 question-answer pairs. The test set contains 4,549 images with 4,549 captions, 4,627 object referring expressions, and 37,928 question-answer pairs. This partition ensures that evaluated models demonstrate genuine generalization capability rather than memorization. For this benchmark release, we evaluate state-of-the-art vision-language models exclusively on the test set, providing standardized protocols for fair comparison across different approaches.

We benchmark multiple leading vision-language models including closed-source systems such as GPT-4o, GPT-5, GPT-5-mini, and Gemini 2.5 Flash, alongside open-source models including Qwen2.5-VL series spanning 3B to 72B parameters, InternVL3.5 series ranging from 1B to 241B parameters, Qwen3-VL 30B in both instruction-following and reasoning modes, and GLM-4.1V and GLM-4.5V series. This diverse model selection enables comprehensive analysis of how model architecture, scale, and training methodology affect
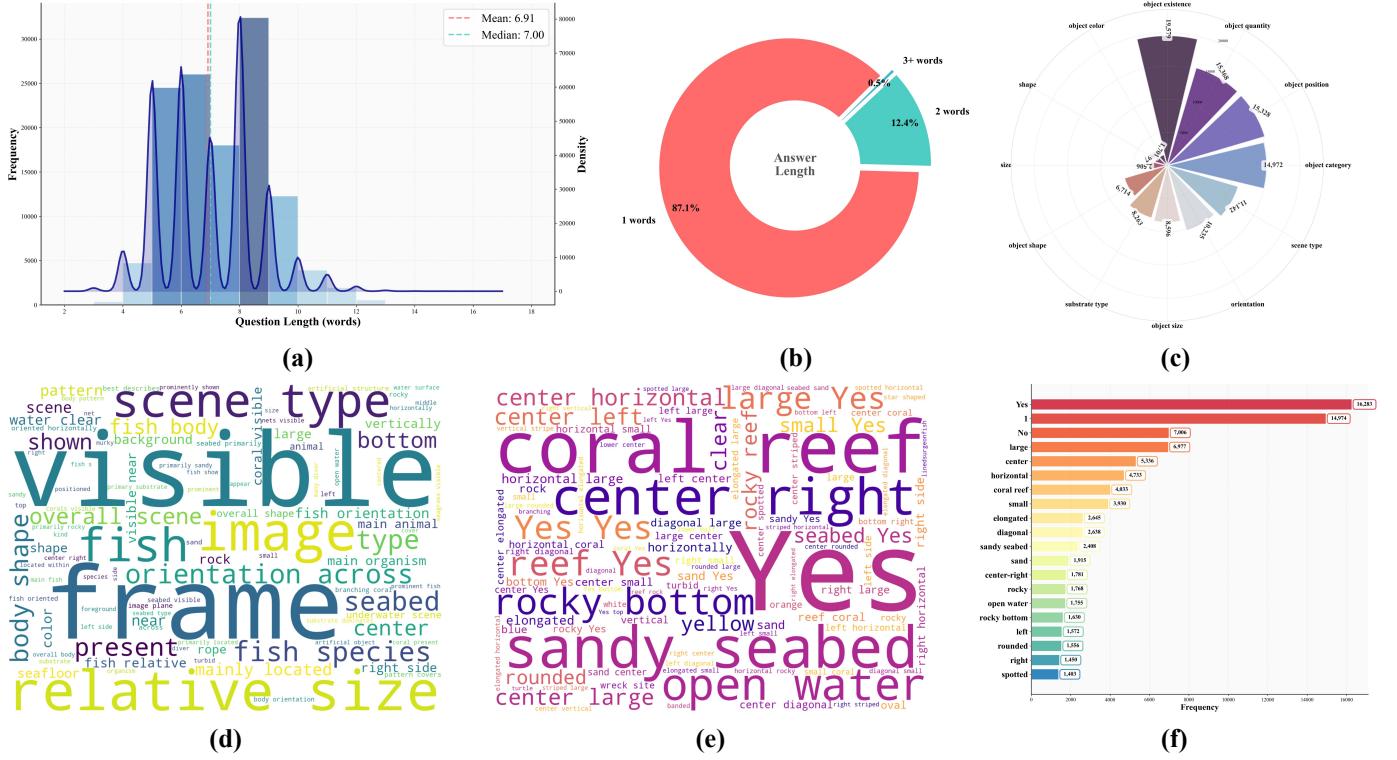
Figure 6: Statistics of visual question answer pairs in UWBench. (a) Question Length Distribution with Density Estimation. (b) Answer Length Distribution. (c) Question Types - Radial Distribution. (d) Question Words Cloud. (e) Answer Words Cloud. (f) Top 20 Most Frequent Answers.

performance on underwater vision-language understanding tasks.

B. Detailed Image Caption

1) Settings: We employ two complementary evaluation approaches to assess caption quality. Traditional automatic metrics provide quantitative measurements of lexical and semantic similarity, while GPT-based evaluation captures higher-level semantic alignment between generated and reference captions.

For automatic evaluation, we utilize established metrics including BLEU measuring n-gram precision with n values of 1 through 4, METEOR assessing semantic similarity through synonym matching and stemming, ROUGE-L computing longest common subsequence, CIDEr evaluating consensus-based image description through TF-IDF weighted n-gram matching, and SPICE measuring semantic propositional content through scene graph matching. These metrics collectively evaluate caption quality from multiple complementary perspectives encompassing lexical precision, semantic similarity, and structural alignment.

We further employ CLAIR score for GPT-based evaluation, which leverages large language models to assess semantic similarity between candidate and reference captions. CLAIR prompts GPT-4o-mini to evaluate on a scale from 0 to 100 how likely a candidate caption describes the same image as the reference caption, considering semantic meaning rather than exact lexical matching. This approach better captures high-level semantic alignment particularly important for detailed underwater captions where diverse valid phrasings exist for

describing complex scenes. We report mean CLAIR score, standard deviation, and median across the test set. Additionally, we compute average caption length and standard deviation to assess description comprehensiveness.

2) Results: Experimental results (Table III) demonstrate substantial variation in model performance across underwater image captioning. GPT-5 achieves the highest automatic metric scores with BLEU-4 of 14.90, METEOR of 27.08, and CIDEr of 66.10, substantially outperforming other approaches. GPT-4o and GPT-5-mini follow with competitive performance, while Gemini 2.5 Flash exhibits lower automatic metric scores despite generating longer captions averaging 86.05 words. Among open-source models, GLM-4.5V and Qwen3-VL-30B-Instruct demonstrate strong performance approaching closed-source systems, while smaller models including Qwen2.5-VL-3B and InternVL-3.5-1B show significant performance gaps.

As figure 7 shows, GPT-based CLAIR evaluation reveals consistently high semantic alignment across most models. GPT-5 achieves the highest CLAIR score of 90.69, followed closely by GPT-4o at 90.11 and GPT-5-mini at 89.55. Open-source models demonstrate competitive CLAIR performance with GLM-4.5V, InternVL-3.5-241B, and InternVL-3.5-38B all exceeding 88.5, indicating strong semantic understanding despite gaps in automatic metrics. Notably, Qwen3-VL-30B-Thinking exhibits lower CLAIR score of 85.38 with high standard deviation of 13.99, suggesting less consistent caption quality in reasoning mode. Analysis of caption length reveals interesting patterns. Closed-source models generally produce captions closer to reference length of 68.6 words, with GPT-4o

Table III: Detailed image caption performance on UWBench. Boldface indicates the best performance.

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr	SPICE
GPT-4o	40.24	22.51	12.95	7.97	22.98	26.79	33.10	28.73
GPT-5	**49.17**	**31.40**	**21.14**	**14.90**	**27.08**	**34.54**	**66.10**	**35.41**
GPT-5-mini	43.41	23.86	13.60	8.11	23.59	26.16	36.53	26.58
Gemini-2.5-Flash	33.87	16.90	8.72	4.83	23.01	22.53	10.21	24.53
Qwen2.5-VL-3B	30.26	16.24	9.43	5.91	17.34	22.84	8.53	24.04
Qwen2.5-VL-7B	32.60	17.31	9.83	6.06	18.17	22.85	11.87	24.72
Qwen2.5-VL-72B	40.07	22.09	12.77	7.90	21.15	25.14	31.24	27.40
InternVL-3.5-1B	31.73	17.68	10.40	6.57	17.89	22.92	11.16	24.44
InternVL-3.5-38B	36.94	20.21	11.58	7.14	20.48	24.38	26.38	26.00
InternVL-3.5-241B	35.88	19.87	11.50	7.13	20.23	24.20	23.60	26.18
Qwen3-VL-30B-Instruct	41.07	22.90	13.35	8.24	23.37	25.18	31.61	27.24
Qwen3-VL-30B-Thinking	40.82	22.64	13.28	8.19	21.74	24.86	29.22	25.79
GLM-4.1V-9B	40.28	22.21	12.87	7.87	21.15	25.02	26.40	23.74
GLM-4.5V-106B	41.96	23.41	13.67	8.41	22.45	25.46	31.48	26.68

generating 73.60 words and Qwen3-VL-30B-Instruct producing 74.10 words. Gemini 2.5 Flash generates substantially longer captions averaging 86.05 words, while smaller open-source models produce more concise descriptions ranging from 48.69 to 61.53 words. This suggests that caption comprehensiveness positively correlates with model scale and training methodology, with larger models better capturing the detailed environmental and biological information characteristic of underwater scenes.

C. Object Referring

Table IV: Object referring performance on UWBench. Boldface indicates the best performance.

Method	Acc@IoU_0.5	Acc@IoU_0.7	mIoU	Cum_IoU
GPT-4o	28.29	6.78	36.78	43.64
GPT-5	62.81	22.78	54.22	60.37
GPT-5-mini	70.37	24.83	56.20	62.37
Gemini-2.5-Flash	60.21	17.53	52.97	57.70
Qwen2.5-VL-3B	54.55	15.37	47.34	51.60
Qwen2.5-VL-7B	37.15	11.76	32.77	33.69
Qwen2.5-VL-72B	90.90	74.71	77.90	**81.50**
InternVL-3.5-1B	/	/	/	/
InternVL-3.5-38B	55.91	22.36	51.38	32.53
InternVL-3.5-241B	84.76	64.14	71.99	76.98
Qwen3-VL-30B-Instruct	**94.40**	**80.38**	**80.18**	51.67
Qwen3-VL-30B-Thinking	56.52	30.04	50.78	56.00
GLM-4.1V-9B	85.95	68.92	75.27	48.07
GLM-4.5V-106B	89.21	75.38	79.26	80.51

1) Settings: Visual grounding evaluation assesses model capability to localize underwater objects based on textual referring expressions. We focus on bounding box prediction using normalized coordinates where models must generate location specifications in the format of four corner coordinates representing top-left and bottom-right positions.

We employ Intersection over Union based accuracy metrics to evaluate localization performance. Accuracy at threshold tau measures the proportion of predictions where IoU between predicted and ground truth bounding boxes exceeds tau. We report Acc@0.5 and Acc@0.7 representing accuracy at IoU

thresholds of 0.5 and 0.7 respectively, providing evaluation at both moderate and strict localization criteria. Additionally, we compute mean IoU across all predictions to assess average localization precision, and cumulative IoU measuring the average maximum IoU achievable, indicating the upper bound of model localization capability.

2) Results: Table IV reveal substantial performance variation across models in underwater object grounding. Qwen3-VL-30B-Instruct achieves the highest Acc@0.5 of 94.40% and Acc@0.7 of 80.38%, demonstrating exceptional localization precision. Qwen2.5-VL-72B and GLM-4.5V follow with strong performance exceeding 89% at IoU 0.5 threshold, while InternVL-3.5-241B and GLM-4.1V achieve competitive accuracy above 84%. Among closed-source models, GPT-5-mini surprisingly outperforms GPT-5 with 70.37% accuracy at threshold 0.5, while GPT-4o exhibits substantially lower performance at 28.29%, suggesting that grounding capability does not necessarily scale with general vision-language performance.

Mean IoU analysis reveals consistent patterns with accuracy metrics. Top-performing models including Qwen3-VL-30B-Instruct, GLM-4.5V, and Qwen2.5-VL-72B achieve mIoU exceeding 77%, indicating precise bounding box predictions. Mid-tier models range from 50% to 60% mIoU, while GPT-4o demonstrates significantly lower precision at 36.78%. Cumulative IoU measurements show similar trends, with best models exceeding 80% and representing strong upper bound performance potential.

Analysis indicates that open-source models generally outperform closed-source systems in visual grounding tasks on underwater imagery. This contrasts with caption generation results and suggests that grounding benefits particularly from domain-specific fine-tuning on object localization tasks. The substantial gap between GPT-5 and GPT-5-mini further indicates that general-purpose training may not optimally transfer to precise spatial reasoning in underwater contexts. Smaller models including Qwen2.5-VL-7B and InternVL-3.5-38B demonstrate moderate performance, highlighting the importance of model scale for complex spatial understanding
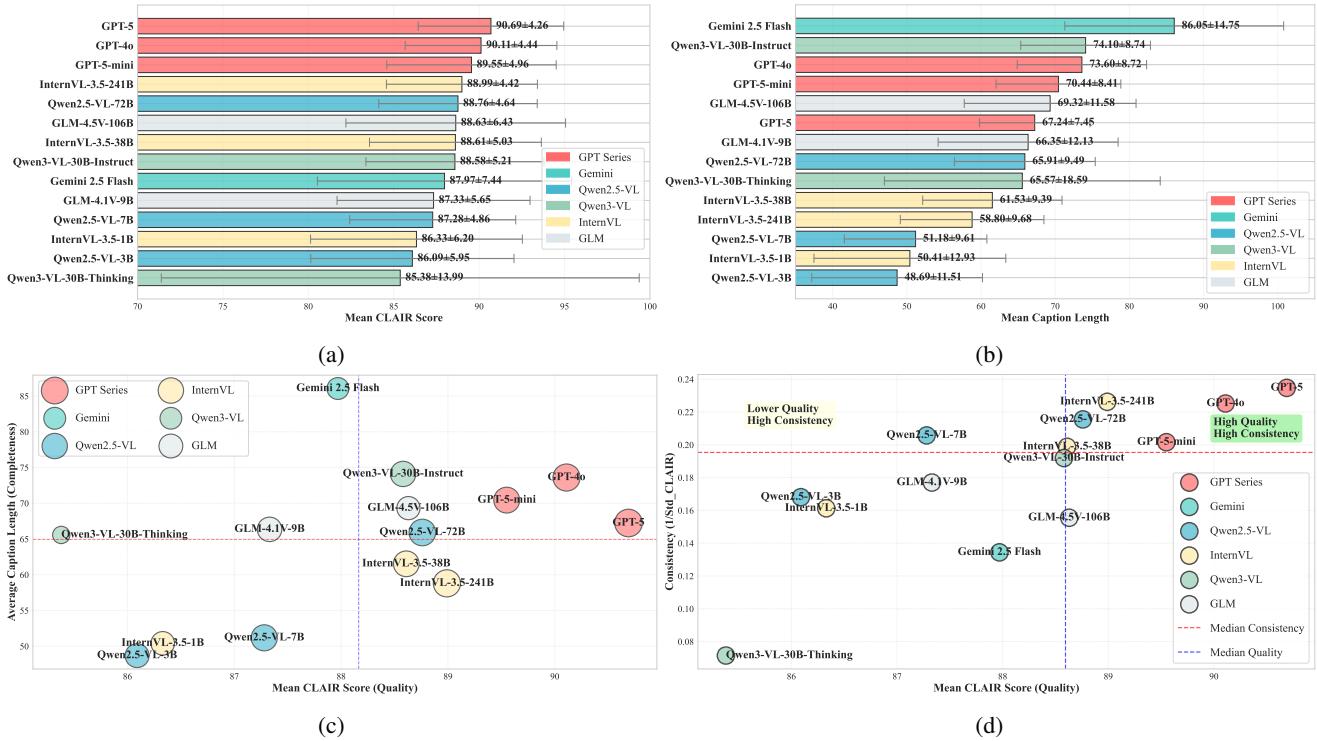
Figure 7: (a) Caption Quality: Mean CLAIR Score with Standard Deviation. (b) Caption Completeness: Mean Caption Length with Standard Deviation. (c) Caption Quality vs Length vs Consistency (Bubble size = Consistency). (d) Quality-Consistency Quadrant Analysis.

Table V: Visual question answering performance on UWBench. Boldface indicates the best performance.

Method	Object Identification	Quantity & Existence	Position & Spatial Relations	Shape, Size & Form	Color, Texture & Pattern	Object Attributes & Features	Scene & Environment	Substrate & Materials	Water Quality & Visibility	Reasoning & Comparison	All
VQAs	4558	10618	8063	6102	1679	160	3540	2604	363	241	37928
GPT-4o	**99.98**	99.58	89.94	83.26	95.11	95.25	66.92	90.63	89.26	95.43	90.95
GPT-5	99.96	**99.61**	90.28	85.04	**97.32**	**98.75**	**85.48**	**93.78**	**95.04**	**97.51**	**93.44**
GPT-5-mini	97.28	96.96	85.02	79.24	93.45	93.75	82.09	90.74	86.50	93.77	89.50
Gemini-2.5-Flash	14.52	1.47	61.55	71.42	45.26	24.37	47.74	47.58	40.77	26.14	37.12
Qwen2.5-VL-3B	38.22	34.46	30.03	35.41	39.73	39.37	30.08	33.91	41.87	44.40	34.06
Qwen2.5-VL-7B	13.84	49.91	55.66	64.42	38.95	21.87	44.89	42.24	35.54	23.24	39.06
Qwen2.5-VL-72B	99.91	9.54	88.45	**93.22**	93.03	98.75	76.04	90.09	86.22	96.26	91.04
InternVL-3.5-1B	14.70	1.77	58.75	70.11	42.29	23.75	47.06	43.47	38.01	26.97	35.91
InternVL-3.5-38B	99.38	99.12	89.62	82.59	92.97	96.87	69.43	88.63	87.05	95.85	90.56
InternVL-3.5-241B	99.96	99.43	90.09	84.66	93.92	96.87	68.67	91.09	88.43	96.26	91.31
Qwen3-VL-30B-Instruct	99.96	99.16	**90.44**	85.17	95.12	96.87	71.89	89.82	88.71	96.27	91.66
Qwen3-VL-30B-Thinking	13.23	39.55	61.07	71.83	43.84	25.62	48.31	45.16	39.12	26.14	41.39
GLM-4.1V-9B	15.40	2.33	61.27	71.98	45.02	28.12	47.32	47.17	39.94	26.97	37.43
GLM-4.5V-106B	21.87	8.73	62.48	72.52	48.78	27.50	50.62	49.85	40.22	29.05	41.01

in challenging underwater conditions.

D. Visual Question Answering

1) Settings: VQA evaluation assesses model capability to answer diverse questions about underwater scenes. We employ GPT-based semantic matching to determine answer correctness, comparing predicted and ground truth answers through semantic similarity rather than exact string matching. This approach accommodates the semantic equivalence of synonymous terms common in underwater domain such as seabed and seafloor, or turbid and murky.

We utilize GPT-4o-mini to evaluate answer similarity through a carefully designed prompt that provides the question, ground truth answer, and predicted answer, instructing the model to determine whether predicted answer matches ground truth

considering semantic meaning. The model returns binary judgment of 1 for match or 0 for no match. To improve efficiency, we implement rule-based matching for straightforward cases including exact matches and simple containment relationships, reserving GPT evaluation for ambiguous cases requiring semantic reasoning. To enable fine-grained analysis, we categorize questions into ten major types reflecting diverse reasoning capabilities required for comprehensive underwater understanding. Detailed classification results are shown in the Appendix.

2) Results: Results in Table V reveal substantial variation across question categories and model capabilities. GPT-5 achieves the highest overall accuracy of 93.44%, demonstrating strong performance across all categories with particular strength in Object Identification at 99.96%, Quantity and Existence at 99.61%, and Color, Texture, and Pattern at 97.32%. GPT-4o
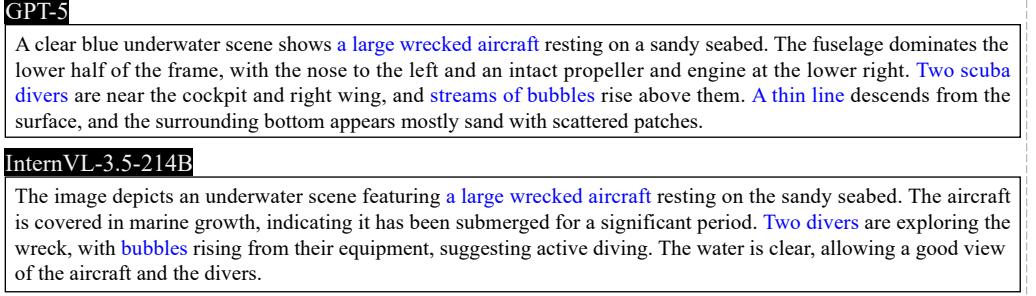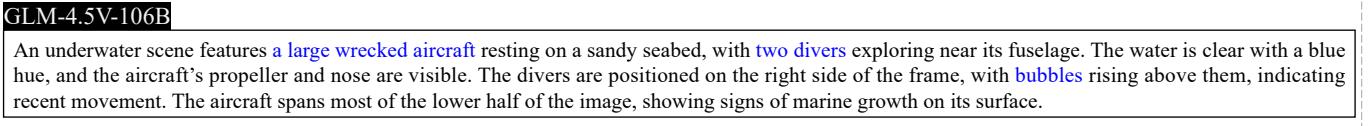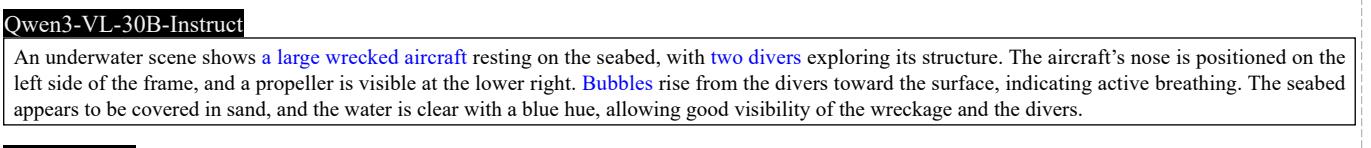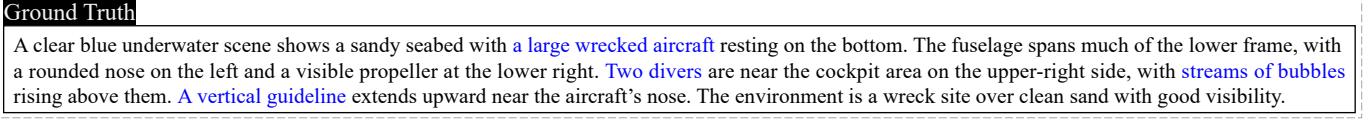
GPT-5

A clear blue underwater scene shows a large wrecked aircraft resting on a sandy seabed. The fuselage dominates the lower half of the frame, with the nose to the left and an intact propeller and engine at the lower right. Two scuba divers are near the cockpit and right wing, and streams of bubbles rise above them. A thin line descends from the surface, and the surrounding bottom appears mostly sand with scattered patches.

InternVL-3.5-214B

The image depicts an underwater scene featuring a large wrecked aircraft resting on the sandy seabed. The aircraft is covered in marine growth, indicating it has been submerged for a significant period. Two divers are exploring the wreck, with bubbles rising from their equipment, suggesting active diving. The water is clear, allowing a good view of the aircraft and the divers.

GLM-4.5V-106B

An underwater scene features a large wrecked aircraft resting on a sandy seabed, with two divers exploring near its fuselage. The water is clear with a blue hue, and the aircraft's propeller and nose are visible. The divers are positioned on the right side of the frame, with bubbles rising above them, indicating recent movement. The aircraft spans most of the lower half of the image, showing signs of marine growth on its surface.

Qwen3-VL-30B-Instruct

An underwater scene shows a large wrecked aircraft resting on the seabed, with two divers exploring its structure. The aircraft's nose is positioned on the left side of the frame, and a propeller is visible at the lower right. Bubbles rise from the divers toward the surface, indicating active breathing. The seabed appears to be covered in sand, and the water is clear with a blue hue, allowing good visibility of the wreckage and the divers.

Ground Truth

A clear blue underwater scene shows a sandy seabed with a large wrecked aircraft resting on the bottom. The fuselage spans much of the lower frame, with a rounded nose on the left and a visible propeller at the lower right. Two divers are near the cockpit area on the upper-right side, with streams of bubbles rising above them. A vertical guideline extends upward near the aircraft's nose. The environment is a wreck site over clean sand with good visibility.

Figure 8: Selected examples of detailed image caption results. We highlight pivotal information in blue.

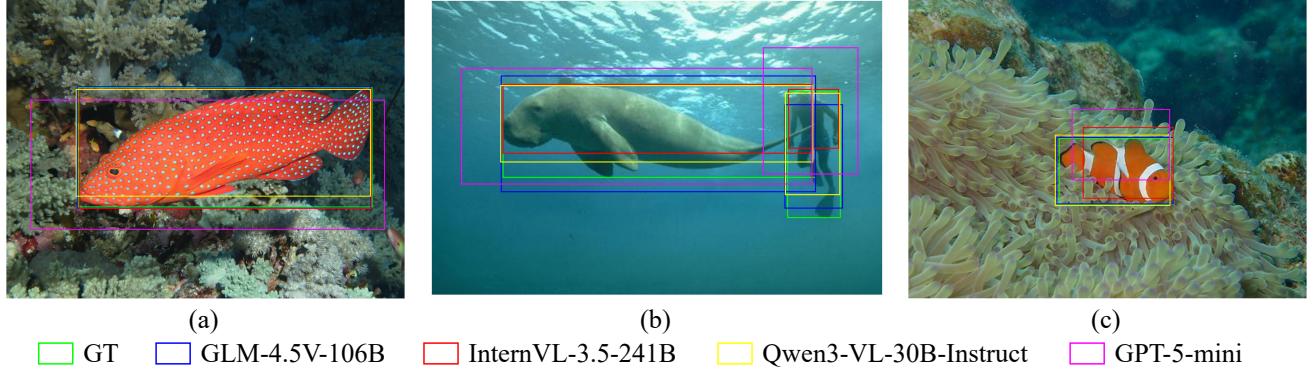|				
□ GT	□ GLM-4.5V-106B	□ InternVL-3.5-241B	□ Qwen3-VL-30B-Instruct	□ GPT-5-mini

Figure 9: Selected examples of referring object. (a) The large elongated coralhind covered in small pale spots extends horizontally across the middle of the image above the corals. (b) The dugong stretching horizontally across the middle of the image just below the bright water surface. & The lone diver on the right side of the frame, positioned to the right of the dugong. (c) The small ocellarisclownfish partly hidden among pale tubular anemone tentacles in the center-right of the image.
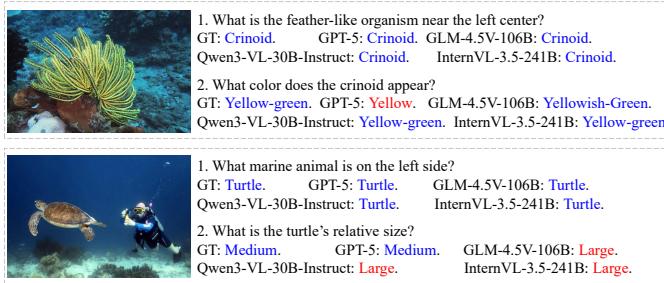

Figure 10: Selected examples of VQA results. Correct answers are shown in blue and incorrect answers are shown in red.

follows closely with 90.95% overall accuracy, while GPT-5-mini achieves competitive 89.50%. Among open-source models, performance varies dramatically. InternVL-3.5-241B achieves the highest accuracy at 91.31%, closely followed by Qwen3-VL-30B-Instruct at 91.66% and Qwen2.5-VL-72B at 91.04%, approaching closed-source performance. However, substantial performance gaps emerge across question categories. All top-performing models exceed 99% on Object Identification and Quantity tasks but show significant challenges on Scene and Environment questions where accuracy drops to 68% to 86%, indicating difficulty in holistic scene reasoning.

Smaller open-source models demonstrate markedly lower performance. GLM-4.5V achieves 41.01% overall accuracy, while GLM-4.1V and Qwen3-VL-30B-Thinking reach approximately 37% to 41%. These models exhibit severe difficulty on Quantity and Existence questions, with accuracy dropping below 10% in some cases, while achieving moderate performance exceeding

60% on Position and Shape categories. This suggests smaller models struggle particularly with numerical reasoning and abstract existence verification but retain reasonable capability for concrete spatial and morphological understanding. Category-specific analysis reveals consistent patterns. Object Identification proves relatively easier with most models exceeding 95% except smaller variants. Quantity and Existence shows the largest performance gap, with top models approaching perfect accuracy while smaller models fail dramatically, likely due to the precise numerical reasoning required for counting underwater organisms. Position and Spatial Relations demonstrates moderate difficulty with top models achieving approximately 90% while smaller models reach 60%. Shape, Size, and Form follows similar patterns. Water Quality and Visibility shows intermediate difficulty around 86% to 95% for strong models, indicating that environmental assessment benefits from both visual perception and domain knowledge.

E. Visualization

We have visualized the three UWBench tasks, demonstrating the typical performance of mainstream models in underwater scene vision-language understanding:

For image caption (Figure 8), the generated results from various models are generally highly consistent with the manually annotated reference descriptions. They all accurately capture key scene elements, such as the wreckage of the sunken aircraft, the sandy bottom, the diver's bubbles, and the clear blue water. They also describe some spatial structure (such as the nose orientation, propeller position, and diver distribution). However, there are subtle differences in performance between different models: For example, InternVL focuses on describing biological traces attached to the aircraft, while GLM and Qwen3-VL-30B-Instruct provide more complete depictions of details and visible structures, with clearer spatial relationships. GPT-5's descriptions are the most comprehensive and closely match the human-annotated descriptions.

For object referring (Figure 9), the models showed good consistency in selecting and localizing objects (e.g., aircraft, propellers, divers, etc.). Visualization results show a high degree of overlap in localization or highlighting across both the ground truth and the models, demonstrating that mainstream models possess accurate underwater object understanding and spatial localization capabilities. Fine-grained differences primarily manifest in accuracy under complex structures or occlusion.

For vqa task (Figure 10), the models achieved high agreement with the ground truth answers for most questions, particularly for specific organisms (e.g., the feathery creature is a crinoid, the animal on the left is a sea turtle), demonstrating strong species recognition. They also performed well on color and relative size questions. For example, with the exception of GPT-5, the large models' judgment of the "yellow-green" color scheme was highly consistent with the ground truth answer. While some models exhibited slight bias on individual questions, overall, all models provided reasonable and scientific answers. This demonstrates that the large models have achieved a high level of fine-grained attribute recognition and reasoning capabilities in underwater scenes.

VI. CONCLUSION

We introduce UWBench, the first large-scale comprehensive benchmark specifically designed for underwater vision-language understanding. The benchmark comprises 15,003 high-resolution underwater images with 15,003 human-verified detailed captions, 15,281 object referring expressions, and 124,983 visual question-answer pairs spanning 158 underwater object categories. Through a rigorous semi-automatic construction pipeline combining GPT-5 assisted generation with multi-stage expert verification, we ensure scientific accuracy and ecological validity across all annotations. UWBench facilitates comprehensive evaluation across three interconnected tasks including detailed image captioning, visual grounding, and visual question answering, providing standardized protocols for systematic assessment of vision-language models in challenging aquatic environments. Our extensive evaluation of state-of-the-art vision-language models reveals that underwater image understanding remains highly challenging even for the most advanced systems. These findings highlight the critical need for specialized approaches tailored to underwater contexts, motivating continued research in this important yet underexplored domain.

Limitations and Future Work: Current vision-language models exhibit common weaknesses including difficulty with precise numerical reasoning, limited understanding of complex ecological relationships, and challenges in integrating domain-specific knowledge with visual perception. Our ongoing efforts include fine-tuning specialized vision-language models on UWBench training data to develop systems with enhanced underwater understanding capabilities.

REFERENCES

[1] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.

[3] Z. Li, X. Wu, H. Du, H. Nghiem, and G. Shi, "Benchmark evaluations, applications, and challenges of large vision language models: A survey," *arXiv preprint arXiv:2501.02189*, vol. 1, 2025.

[4] Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. Wang, "Figstep: Jailbreaking large vision-language models via typographic visual prompts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 951–23 959.

[5] L. Zhu, D. Ji, T. Chen, P. Xu, J. Ye, and J. Liu, "Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1624–1633.

[6] H. Li, H. Wang, Y. Zhang, L. Li, and P. Ren, "Underwater image captioning: Challenges, models, and datasets," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 220, pp. 440–453, 2025.

[7] S. Sarto, M. Cornia, R. Cucchiara *et al.*, "Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives," in *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, 2025.

[8] H. Hua, Q. Liu, L. Zhang, J. Shi, S. Y. Kim, Z. Zhang, Y. Wang, J. Zhang, Z. Lin, and J. Luo, "Finecaption: Compositional image captioning focusing on wherever you want at any granularity," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 24 763–24 773.

[9] C. Huang, B. Maneechotesuwan, S. Chopra, and Z. Kira, "Frames-vqa: Benchmarking fine-tuning robustness across multi-modal shifts in visual question answering," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3909–3918.

[10] O. Mañas, B. Krojer, and A. Agrawal, "Improving automatic vqa evaluation using large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4171–4179.

[11] L. Xiao, X. Yang, X. Lan, Y. Wang, and C. Xu, "Towards visual grounding: A survey," *arXiv preprint arXiv:2412.20206*, 2024.

[12] D. Liu, Y. Liu, W. Huang, and W. Hu, "A survey on text-guided 3-d visual grounding: Elements, recent advances, and future directions," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[13] S. Nahavandi, R. Alizadehsani, D. Nahavandi, S. Mohamed, N. Mohajer, M. Rokonuzzaman, and I. Hossain, "A comprehensive review on autonomous navigation," *ACM Computing Surveys*, vol. 57, no. 9, pp. 1–67, 2025.

[14] J. Lee, M. Bjelonic, A. Reske, L. Wellhausen, T. Miki, and M. Hutter, "Learning robust autonomous navigation and locomotion for wheeled-legged robots," *Science Robotics*, vol. 9, no. 89, p. eadi9641, 2024.

[15] M. Kolla, S. Salunkhe, E. Chandrasekharan, and K. Saha, "Llm-mod: Can large language models assist content moderation?" in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–8.

[16] K. Palla, J. L. R. García, C. Hauff, F. Fabbri, A. Damianou, H. Lindström, D. Taber, and M. Lalmas, "Policy-as-prompt: Rethinking content moderation in the age of large language models," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025, pp. 840–854.

[17] J. Yuan, J. Gupta, A. Padmanabha, Z. Karachiwalla, C. Majidi, H. Admoni, and Z. Erickson, "Towards an llm-based speech interface for robot-assisted feeding," in *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, pp. 1–4.

[18] A. Padmanabha, J. Yuan, J. Gupta, Z. Karachiwalla, C. Majidi, H. Admoni, and Z. Erickson, "Voicepilot: Harnessing llms as speech interfaces for physically assistive robots," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, pp. 1–18.

[19] H. Li, M. Xu, Y. Zhan, S. Mu, J. Li, K. Cheng, Y. Chen, T. Chen, M. Ye, J. Wang *et al.*, "Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 7752–7762.

[20] A. Albalak, D. Phung, N. Lile, R. Rafailov, K. Gandhi, L. Castricato, A. Singh, C. Blagden, V. Xiang, D. Mahan *et al.*, "Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models," *arXiv preprint arXiv:2502.17387*, 2025.

[21] Q. Wang, Y. Shi, J. Ou, R. Chen, K. Lin, J. Wang, B. Jiang, H. Yang, M. Zheng, X. Tao *et al.*, "Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8428–8437.

[22] W. Pei, X. Pei, Z. Xie, and J. Wang, "Research progress of marine anti-corrosion and wear-resistant coating," *Tribology International*, vol. 198, p. 109864, 2024.

[23] J. Chen, Y. Jia, Y. Sun, K. Liu, C. Zhou, C. Liu, D. Li, G. Liu, C. Zhang, T. Yang *et al.*, "Global marine microbial diversity and its potential in bioprospecting," *Nature*, vol. 633, no. 8029, pp. 371–379, 2024.

[24] A. K. Spalding and E. McKinley, "The state of marine social science: Yesterday, today, and into the future," *Annual Review of Marine Science*, vol. 17, 2025.

[25] J. Zhou, Z. He, D. Zhang, S. Liu, X. Fu, and X. Li, "Spatial residual for underwater object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[26] Z. Wang, L. Shen, M. Xu, M. Yu, K. Wang, and Y. Lin, "Domain adaptation for underwater image enhancement," *IEEE Transactions on Image Processing*, vol. 32, pp. 1442–1457, 2023.

[27] Z. Wu, Z. Wu, X. Chen, Y. Lu, and J. Yu, "Self-supervised underwater image generation for underwater domain pre-training," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–14, 2024.

[28] T. R. Fuad, S. Ahmed, and S. Ivan, "Aqua20: A benchmark dataset for underwater species classification under challenging conditions," *arXiv preprint arXiv:2506.17455*, 2025.

[29] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[30] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[31] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu *et al.*, "Qwen2. 5-coder technical report," *arXiv preprint arXiv:2409.12186*, 2024.

[32] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[33] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao *et al.*, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," *arXiv preprint arXiv:2406.12793*, 2024.

[34] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024.

[35] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[37] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 10 371–10 381.

[38] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[39] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[40] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 185–24 198.

[41] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[42] K. Zheng, Y. Zhang, W. Wu, F. Lu, S. Ma, X. Jin, W. Chen, and Y. Shen, "Dreamlip: Language-image pre-training with long captions," in *European Conference on Computer Vision*. Springer, 2024, pp. 73–90.

[43] W. Tang, L. Li, X. Liu, L. Jin, J. Tang, and Z. Li, "Context disentangling and prototype inheriting for robust visual grounding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3213–3229, 2023.

[44] S. Yang, W. Yu, W. Yang, X. Liu, H. Tan, L. Lan, and N. Xiao, "Wildvideo: Benchmarking lmms for understanding video-language interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[45] L. Hong, Z. Liu, W. Chen, C. Tan, Y. Feng, X. Zhou, P. Guo, J. Li, Z. Chen, S. Gao *et al.*, "Lvos: A benchmark for large-scale long-term video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[47] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[48] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[49] J. Chen, F. Wei, J. Zhao, S. Song, B. Wu, Z. Peng, S.-H. G. Chan, and H. Zhang, "Revisiting referring expression comprehension evaluation in the era of large multimodal models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 513–524.

[50] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[51] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[52] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE transactions on image processing*, vol. 29, pp. 4376–4389, 2019.

[53] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE robotics and automation letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[54] S. Raveendran, M. D. Patil, and G. K. Birajdar, "Underwater image enhancement: a comprehensive review, recent trends, challenges and applications," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 5413–5467, 2021.

[55] H. Wang, W. Zhang, L. Bai, and P. Ren, "Metalantis: A comprehensive underwater image enhancement framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–19, 2024.

[56] L. Shen, H. Xia, X. Zhang, Y. Zhao, N. Li, S. G. Kong, B. Wang, and Z. Li, "U^{2}pnet: An unsupervised underwater image-restoration network using polarization," *IEEE Transactions on Cybernetics*, vol. 54, no. 9, pp. 5164–5177, 2024.

[57] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE transactions on image processing*, vol. 32, pp. 3066–3079, 2023.

[58] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[59] C. Liu, H. Li, S. Wang, M. Zhu, D. Wang, X. Fan, and Z. Wang, "A dataset and benchmark of underwater object detection for robot picking," in *2021 IEEE international conference on multimedia & expo workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[60] M. Pedersen, D. Lehotský, I. Nikolov, and T. B. Moeslund, "Brackishmot: The brackish multi-object tracking dataset," in *Scandinavian Conference on Image Analysis*. Springer, 2023, pp. 17–33.

[61] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2020, pp. 1769–1776.

[62] S. Lian, H. Li, R. Cong, S. Li, W. Zhang, and S. Kwong, "Watermask: Instance segmentation for underwater imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 1305–1315.

[63] L. Hong, X. Wang, G. Zhang, and M. Zhao, "Usod10k: a new benchmark dataset for underwater salient object detection," *IEEE transactions on image processing*, vol. 34, pp. 1602–1615, 2023.

[64] L. Hong, X. Wang, Y. Li, and X. Wang, "Usis16k: High-quality dataset for underwater salient instance segmentation," *arXiv preprint arXiv:2506.19472*, 2025.

[65] B. Alawode, Y. Guo, M. Ummar, N. Werghi, J. Dias, A. Mian, and S. Javed, "Utb180: A high-quality benchmark for underwater tracking," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 3326–3342.

[66] L. Cai, N. E. McGuire, R. Hanlon, T. A. Mooney, and Y. Girdhar, "Semi-supervised visual tracking of marine animals using autonomous underwater vehicles," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1406–1427, 2023.

[67] C. Zhang, L. Liu, G. Huang, H. Wen, X. Zhou, and Y. Wang, "Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50152–50167, 2024.

[68] H. Qin, X. Li, J. Liang, Y. Peng, and C. Zhang, "Deepfish: Accurate underwater live fish recognition with a deep architecture," *Neurocomputing*, vol. 187, pp. 49–58, 2016.

[69] F. F. Khan, X. Li, A. J. Temple, and M. 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*, 2023, pp. 20496–20506.

[70] N. Varghese, A. Kumar, and A. Rajagopalan, "Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 12248–12258.

[71] X. Xue, Y. Zhou, D. Yan, Y. Li, H. Zhang, and R. Xiao, "Uvlm: Benchmarking video language model for underwater world understanding," *arXiv preprint arXiv:2507.02373*, 2025.

[72] H. Han, W. Wang, G. Zhang, M. Li, and Y. Wang, "Coralvqa: A large-scale visual question answering dataset for coral reef image understanding," *arXiv preprint arXiv:2507.10449*, 2025.

[73] S. Sun, H. Wang, H. Zhang, M. Li, M. Xiang, C. Luo, and P. Ren, "Underwater image enhancement with reinforcement learning," *IEEE Journal of Oceanic Engineering*, vol. 49, no. 1, pp. 249–261, 2024.

[74] M. Elmezain, L. S. Saoud, A. Sultan, M. Heshmat, L. Seneviratne, and I. Hussain, "Advancing underwater vision: a survey of deep learning models for underwater object recognition and tracking," *IEEE Access*, 2025.

[75] J. Gao, Y. Li, Y. Chen, Y. He, and J. Guo, "An improved sac-based deep reinforcement learning framework for collaborative pushing and grasping in underwater environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–14, 2024.

[76] Y. Tian, K. Yao, and X. Yu, "An adaptive underwater image enhancement framework via multi-domain fusion and color compensation," *arXiv preprint arXiv:2503.03640*, 2025.

[77] D. Du, L. Si, F. Xu, J. Niu, and F. Sun, "A physical model-guided framework for underwater image enhancement and depth estimation," *arXiv preprint arXiv:2407.04230*, 2024.

[78] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation letters*, vol. 3, no. 1, pp. 387–394, 2017.

[79] D. Xu, J. Zhou, Y. Liu, and X. Min, "Underwater image enhancement based on hybrid enhanced generative adversarial network," *Journal of Marine Science and Engineering*, vol. 11, no. 9, p. 1657, 2023.

[80] M. Guan, H. Xu, G. Jiang, M. Yu, Y. Chen, T. Luo, and X. Zhang, "Diffwater: Underwater image enhancement based on conditional denoising diffusion probabilistic model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 2319–2335, 2023.

[81] M. Jian, N. Yang, C. Tao, H. Zhi, and H. Luo, "Underwater object detection and datasets: a survey," *Intelligent Marine Technology and Systems*, vol. 2, no. 1, p. 9, 2024.

[82] Z. Zheng, H. Liang, B.-S. Hua, Y. H. Wong, P. Ang, A. P. Y. Chui, and S.-K. Yeung, "Coralscop: Segment any coral image on this planet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28170–28180.

[83] Z. Zheng, Y. Chen, H. Zeng, T.-A. Vu, B.-S. Hua, and S.-K. Yeung, "Marineinst: A foundation model for marine image analysis with instance visual description," in *European Conference on Computer Vision*. Springer, 2024, pp. 239–257.

[84] Z. Zheng, J. Zhang, T.-A. Vu, S. Diao, Y. H. W. Tim, and S.-K. Yeung, "Marinegpt: Unlocking secrets of ocean to the public," *arXiv preprint arXiv:2310.13596*, 2023.

[85] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[86] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vaya, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical image analysis*, vol. 66, p. 101797, 2020.

[87] Z. Zhang, Y. Yu, Y. Chen, X. Yang, and S. Y. Yeo, "Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29744–29755.

[88] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[89] X. Li, J. Ding, and M. Elhoseiny, "Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 3229–3242, 2024.

[90] X. Guo, Z. Zhu, T. Yang, B. Lin, J. Huang, J. Deng, G. Huang, J. Zhou, and J. Lu, "Gait recognition in the wild: A large-scale benchmark and nas-based baseline," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[91] W. Luo, Y. Liu, B. Li, W. Hu, and S. Maybank, "Figvcl: Fine-grained benchmark and method for video copy localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[92] Y. Tang, A. Liu, J. Liu, S. Zhang, W. Dai, J. Zhou, X. Li, and J. Lu, "Flag3d++: A benchmark for 3d fitness activity comprehension with language instruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[93] K. Huang, C. Duan, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[94] H. Li, S. Lian, Z. Li, R. Cong, and S. Kwong, "Uwsam: Segment anything model guided underwater instance segmentation and a large-scale benchmark dataset," *arXiv preprint arXiv:2505.15581*, 2025.