UQ-Bench: A Benchmark for Evaluating Multimodal LLMs on Underwater Image Quality Assessment

Jingchao Cao[1], Guo An[1], Feng Gao[1], Ke Gu[2], Yutao Liu[1]*

[1]Ocean University of China
[2]Beijing University of Technology
caojingchao@ouc.edu.cn, anguo@stu.ouc.edu.cn, gaofeng@ouc.edu.cn, liuyutao@ouc.edu.cn, guke@bjut.edu.cn

Abstract

Despite the rapid progress of multimodal large language models (MLLMs), their capacity for low-level visual perception in underwater environments remains underexplored. To address this gap, we present UQ-Bench, the first systematically designed benchmark for evaluating the ability of MLLMs to perceive and assess underwater image quality at the low-level visual attribute level. UQ-Bench comprises three components: (1) UW-Perception, a dataset of 3,000 underwater images paired with targeted questions on key degradations such as color cast, blur, contrast, and exposure, covering both global and local perceptual dimensions; (2) UW-Describe, a dataset of 500 images with expert-annotated gold-standard descriptions for assessing the accuracy of model-generated text; and (3) UW-Eval, an evaluation protocol employing human mean opinion scores (MOS) for quantitative quality assessment. To ensure rigorous and reproducible benchmarking, we propose a GPT-assisted evaluation framework that aligns model outputs with expert references and enables fine-grained analysis of distortion perception. Experimental results demonstrate that while MLLMs exhibit preliminary competence in underwater low-level visual tasks, they still fall short in capturing subtle degradations and achieving human-level consistency, highlighting the need for further advances in foundation models for marine vision.

Code — https://github.com/ag-mercury/UQ-bench

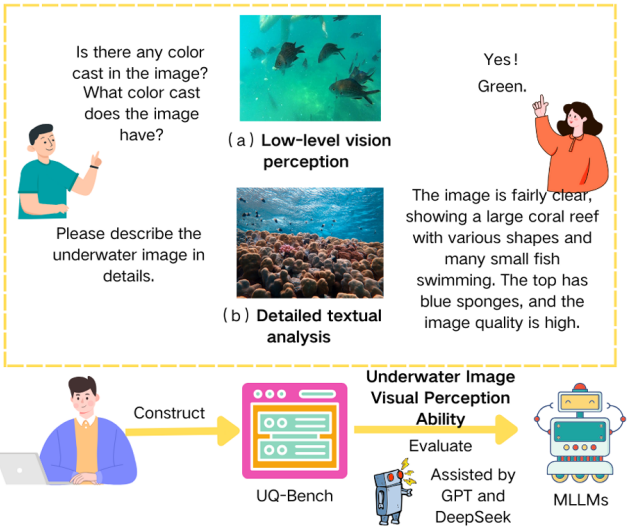

Figure 1: In UQ-Bench, we propose the first benchmark specifically designed to test the visual perception ability of MLLMs for underwater images, including both question-answering and description-based test formats. We also incorporate GPT and DeepSeek as auxiliary evaluation strategies.

Introduction

With the rapid development of large language models (LLMs) such as ChatGPT (OpenAI 2024) and DeepSeek (DeepSeek-AI 2025), alongside a variety of high-quality open-source systems, artificial intelligence is increasingly influential in both specialized and everyday contexts. Meanwhile, multimodal large language models (MLLMs), exemplified by LLaVA (Liu et al. 2023), Kosmos-2 (Peng et al. 2023), and MiniGPT-4 (Zhu et al. 2023), have demonstrated outstanding capabilities in tasks such as visual perception and multi-turn dialogue generation. Existing research has established that MLLMs can achieve strong performance on fundamental tasks, including visual question answering (QA) (Antol et al. 2015), image captioning (Chen et al. 2015), low-level visual recognition, and image classification (Lai et al. 2024). However, systematic studies of MLLMs' performance on underwater imagery remain scarce, even though underwater image quality assessment (UIQA) is becoming increasingly crucial in fields such as marine science. This gap underscores the pressing need to explore how MLLMs can be better adapted to underwater environments.

Underwater imaging differs significantly from natural-scene photography, with common issues such as color cast, contrast degradation, improper exposure, and blurriness or noise arising from the absorption and scattering of light underwater. These distortions pose considerable challenges for downstream tasks such as marine species recognition and seabed structure analysis, and they also place stringent demands on the perceptual and interpretive capabilities of machine learning models (Garg et al. 2019; Islam et al. 2020). To address these challenges, we propose UQ-Bench, the first

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

benchmark specifically designed to evaluate the low-level visual perception and assessment capabilities of MLLMs in underwater scenarios. Centered on four common distortion types (color cast, contrast, exposure, and blur/noise), UQ-Bench evaluates models across both linguistic and visual dimensions. This is accomplished through two dataset-driven tasks and one experimental evaluation:

#1. **Question-Answering on Underwater Distortions:** We systematically prompt models with questions targeting specific types of underwater degradations (e.g., "Is there any color cast in the image?") and evaluate their ability to accurately detect and categorize these distortions. The questions are presented in three formats: Yes-or-No, What, and How, and they address both global and local perceptual aspects.

#2. **Overall Description of Underwater Distortions:** In this task, models are required to generate free-form descriptions summarizing the overall visual quality of underwater images, including both global and localized distortions. This assesses not only the model's detection abilities but also its capacity to integrate observations into coherent textual explanations. (As shown in Figure 1(b))

#3. **Real-Image Quality Assessment (UW-Eval):** Beyond perception and description, we introduce an experimental task to evaluate whether MLLMs can judge the perceptual quality of real underwater images. We construct a test set with human-provided Mean Opinion Scores (MOS) and prompt models to generate corresponding quality assessments. This task probes how closely model-generated quality judgments align with human perception.

To support these tasks, we construct two complementary datasets: UW-Perception and UW-Describe. UW-Perception consists of 3,000 underwater images spanning various scenes and degradation levels. Each image is paired with a question about a low-level quality attribute, a ground truth answer, and several distractors. These questions are categorized by distortion type and perception scope (global or local). UW-Describe contains 500 underwater images annotated with expert-written gold-standard descriptions. To evaluate the quality of model-generated outputs, we employ single-modal language models such as GPT and DeepSeek to assess each response in terms of completeness, specificity, and relevance. For the real-image assessment task (UW-Eval), we use a similar GPT-assisted framework to compare model predictions against MOS-based ground truth.

The proposed UQ-Bench makes the following key contributions:

- **First dedicated benchmark for underwater MLLM evaluation:** We introduce the first systematically designed benchmark specifically targeting low-level visual distortions in underwater imagery, addressing a critical gap in existing benchmarks that primarily focus on natural or synthetic scenes.
- **Comprehensive multi-perspective evaluation protocol:** The proposed benchmark integrates both visual question answering and descriptive tasks to assess the perceptual capabilities of MLLMs across global and local dimensions.
- **Expert-annotated reference dataset:** We curate a high-quality reference dataset with expert-written textual descriptions, enabling reliable and interpretable comparisons between model outputs and human ground truth.
- **Scalable and automated evaluation pipeline:** Leveraging GPT and DeepSeek as judgment backends, we design a robust and scalable evaluation framework for quantitative assessment of MLLM performance.

Database Construction and Analysis

General Principles

Focusing on Low-Level Visual Perception in Underwater Image Quality. The proposed UQ-Bench is the first benchmark specifically focused on low-level visual perception in underwater imagery, setting it apart from general-purpose multimodal benchmarks designed to assess the overall capabilities of MLLMs. Underwater environments present distinct visual distortions such as color casts, reduced contrast, blurring, and noise (Zhang et al. 2023, 2025b; Schwenker 2013). These distortions pose substantial challenges to visual perception, making robust perceptual understanding critical for effective information extraction (Li et al. 2021; Jian et al. 2021). Therefore, evaluating the perceptual abilities of MLLMs in underwater scenarios requires dedicated methodologies that account for these environmental characteristics, rather than directly applying conventional benchmarks designed for natural scenes (Wang et al. 2004; Li et al. 2020).

Based on these considerations, we established two principal evaluation criteria. The first is *intuitiveness*, defined as the model's ability to directly and accurately recognize key visual features in underwater scenes. This criterion requires the model to exhibit rapid perception and precise representation (Hu et al. 2024; Li et al. 2019a). The second criterion assesses the models' ability to perceive and interpret underwater-specific low-level distortions, such as color casts, insufficient contrast, improper exposure, blurring, and noise(Zhang et al. 2025a). This dimension demands not only standard recognition abilities but also effective analysis and judgment of underwater-specific visual distortions. The constructed database, encompassing diverse underwater scenes and various visual degradations, effectively reflects real-world underwater visual conditions for evaluating MLLMs' low-level perception capabilities. This database provides comprehensive visual information and serves as a robust foundation for future research on MLLM applications in specialized scenarios, thereby contributing to advancements in underwater image processing and marine science (Ogunsina et al. 2024).

Diversity in Underwater Scenes and Distortions. All images in the proposed UW-Perception database are sourced from existing underwater visual research. We carefully curated 3,000 underwater images covering a wide spectrum of scenes, including aquatic animals, plants, divers, man-made objects, and underwater wreckage (Liu et al. 2024).

Additionally, these images exhibit multiple common underwater distortions, including color casts, low contrast, blurriness, and noise, ensuring authenticity and diversity(Cao et al. 2021).

Figure 2: Representative samples from the proposed UQ-Bench, illustrating diverse scenes, varying levels of visual quality, and multiple types of low-level distortions.

During image collection, we categorized the dataset into three main groups. The first group, *animals*, comprises mobile aquatic creatures (e.g., turtles, crabs, fish, and shrimp). The second group, *plants*, includes stationary underwater organisms (e.g., seaweed, sea urchins, and anemones). The third group, *humans and man-made artifacts*, covers divers, shipwrecks, submarines, and underwater robots. Each category consists of 1,000 images, which are carefully balanced in quality and uniformly distributed across distortion types. This design enhances the representativeness and reliability of subsequent evaluations. Figure 2 showcases representative examples from the proposed UW-Perception database, highlighting the diversity in underwater scenes, quality levels, and visual distortion types.

Low-Level Visual Perception Benchmark

The initial task evaluates MLLM's perceptual capabilities regarding low-level visual features in underwater images (Mazel 2005; Nomura, Sugimura, and Hamamoto 2018), focusing specifically on performance in addressing simple queries about visual quality attributes. From the existing research database, we selected 3,000 underwater images encompassing diverse scenes and varied degradation types (I). For each image, we formulated a simple question (Q) related to low-level visual quality, accompanied by one correct answer (C) and one to three incorrect distractors (F). Questions span four common distortion categories, evaluating both local and global visual contexts through three distinct question formats (Li et al. 2016). Image-question-answer triplets (I, Q, C, F) were input into MLLMs, and their outputs verified using GPT-assisted evaluations for correctness. Figure 3 presents a visual summary of the UW-Perception dataset structure, including the three question types and the four low-level distortion categories used to construct the benchmark (Tsimpoukelli et al. 2021).

Question Types. To emulate the diversity of human questioning behavior, we designed three distinct types of queries in the proposed UW-Perception dataset: binary judgment, descriptive inquiries, and fine-grained 'how' questions. Each query format targets specific aspects of visual perception and is intended to evaluate MLLMs from complementary perspectives(Li et al. 2024, 2023). Binary questions assess a model's ability to make categorical decisions (e.g., presence or absence of distortion), descriptive questions test its capacity to generate natural language explanations of visual content, and 'how' questions probe its understanding of distortion severity or quality levels. Together, these formats offer a more comprehensive and balanced evaluation framework (Lu et al. 2024).

Local and Global Context Queries. Recent studies indicate human visual perception integrates both local details and global contextual information. Reflecting this understanding, we differentiate visual tasks into global perception tasks, focusing on structural and holistic scene features, and local contextual tasks, emphasizing detailed interpretations of complex scenes. This bifurcation reveals multilayered perceptual dynamics, providing theoretical foundations for developing and optimizing MLLMs (Li et al. 2019b; Wang et al. 2021; Wu et al. 2022).

Evaluation Protocol. To evaluate low-level visual perception in MLLMs, we constructed a benchmark dataset paired with expert-designed questions and multiple-choice answers. For example: "How is the color cast of the lower left corner? [IMAGE TOKEN] Choose one: [Normal (Correct), Slight (Wrong), Severe (Wrong)]." To reduce position bias, answer options were randomly shuffled.

Since MLLMs often produce free-form responses like "No color cast" or "The color cast is normal" that complicate automatic evaluation, we employed GPT and DeepSeek as evaluators. Each MLLM response and its reference answer were judged independently by both models (Krishna et al. 2017; Liang et al. 2023).

To ensure robustness amid evaluator variability, we adopted a five-round voting strategy. Each prompt was assessed five times by both evaluators, and the final result was determined by aggregating their judgments (Schwenker 2013).

Underwater Image Description Benchmark

In our second task, we evaluate the descriptive capabilities of MLLMs for underwater imagery. Specifically, MLLMs generate natural language descriptions of both low-level visual features and overall content within underwater images, with a particular focus on low-level visual attributes. To accurately assess these capabilities, we compiled a dataset comprising 500 underwater images. Each image was paired with an expert-curated detailed description (approximately 50 words), termed the *golden description*. Leveraging these comprehensive descriptions, we systematically evaluated the precision, completeness, and relevance of MLLMs' descriptions regarding low-level visual features.

Definition of Golden Descriptions for Underwater Images. MLLM-generated descriptions for underwater images should encapsulate as many low-level visual attributes as possible while thoroughly conveying the overall visual information(Li et al. 2025). Our golden descriptions average around 50 words, meticulously detailing four primary

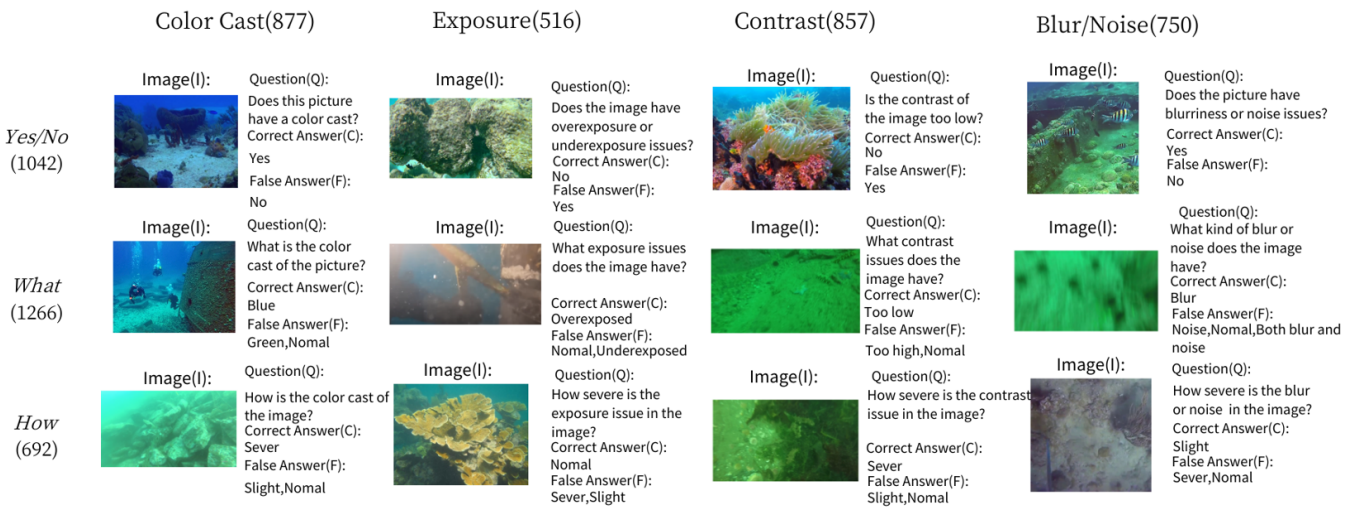

Figure 3: Overview of the UW-Perception dataset. The figure illustrates the three types of question formats and the four low-level distortion quadrants used in our benchmark. Each image in the dataset is paired with the most appropriate question and corresponding answer candidates. The numbers in parentheses denote the number of images associated with each question type.

distortion dimensions (color cast, contrast, exposure, and blur/noise), along with additional content-specific details. The images selected span diverse underwater scenarios and tasks, representing a broad spectrum of quality levels.

MLLMs Description Procedure for Underwater Images. The explicit prompt used for eliciting MLLM descriptions is as follows: *Describe the quality and low-level appearance of the image in detail*. Importantly, we did not specifically instruct the MLLMs to focus exclusively on color cast, contrast, exposure, and blur/noise dimensions. This approach evaluates whether the models can inherently recognize and describe underwater-specific distortions and degradations at a level comparable to human descriptions.

Single-Modal Evaluation using GPT and DeepSeek. Recent research has demonstrated the effectiveness of single-modal GPT as a reliable evaluation tool for purely linguistic tasks, with DeepSeek similarly recognized as an exemplary tool in this domain (Zheng et al. 2023). Following the MLLMs' description generation for our 500 underwater images, we compared their outputs against our curated golden descriptions. The evaluation encompassed three dimensions:

- **Precision:** Evaluates the accuracy of descriptions regarding specific low-level visual distortions such as color cast, contrast, exposure, blur/noise, and assesses the model's ability to correctly identify content under common underwater distortions.
- **Completeness:** Measures the extent to which the description captures all relevant low-level visual attributes, as well as comprehensive details of the underwater image content.
- **Relevance:** Assesses the relevance of descriptions specifically related to low-level visual features, including distortion attributes such as blur/noise, exposure issues, color accuracy, lighting conditions, focus, and composition.

Benchmarking the Precise Quality Assessment Capability of MLLMs

In this task, we benchmark the ability of MLLMs to assess the low-level visual quality of underwater images. Unlike the previous tasks, the focus here is on whether models can mimic human subjective judgments and produce reliable quality scores. We use several widely adopted underwater IQA datasets with subjective MOSs, covering various degradations such as color cast, blurriness, and contrast loss.

Since MLLM outputs lack explicit quantitative structure, extracting continuous scores directly is challenging. To overcome this, we introduce a softmax-based probabilistic scoring method. By treating "good" and "poor" as anchor tokens, we compute their relative probabilities from output logits to derive a continuous quality score. This approach is simple, general, and improves consistency with human ratings across different models.

We propose a unified evaluation framework to fairly compare multiple MLLMs on underwater image quality perception. To ensure consistency across models and datasets, we design standardized prompts that guide all models to assess images based on the same criteria, considering factors such as color distortion, contrast loss, exposure issues, noise, and blurriness.

We explore two prompt styles: (1) classification prompts, where models choose between "good" and "poor," and (2) scoring prompts, where models rate image quality from 1 to 5. While classification prompts work well across various models, scoring prompts often lead to biased outputs toward extreme values (e.g., mostly "1" or "5"), reducing evaluation reliability.

To address this, we introduce a softmax-based scoring

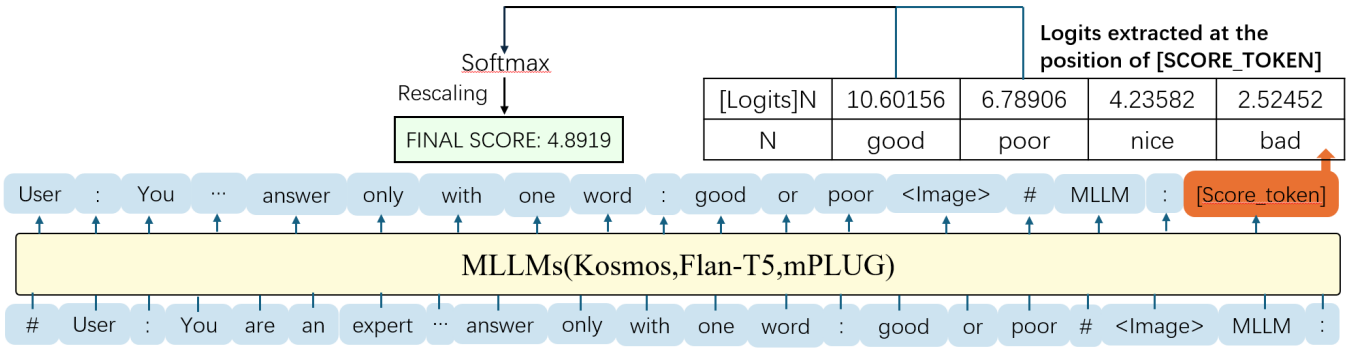

Figure 4: Overview of the softmax-based evaluation strategy using MLLMs.

method that avoids relying on the model's final answer. By computing the relative probabilities of "good" and "poor" tokens from output logits, this method yields more stable and quantifiable quality scores.

A Softmax-Based Quantitative Evaluation Strategy. To overcome the challenges of subjective bias and the limitations of discrete quality scales, we propose a softmax-based quantitative evaluation strategy. As illustrated in Figure 4, this approach leverages the output logits of multi-modal large language models (MLLMs), such as Kosmos, Flan-T5-XL, and mPLUG, to produce a continuous quality score for underwater images.

The central idea is to designate the tokens "good" and "poor" as the polar ends of the quality spectrum. By extracting their corresponding logits from the model's output and applying a softmax function, we compute their relative probabilities. This yields a continuous prediction score reflecting the image's perceived quality. The scoring formula is defined as:

$$\text{Pred} = \frac{e^{z_{\text{good}}}}{e^{z_{\text{good}}} + e^{z_{\text{poor}}}}. \tag{1}$$

This predicted probability can then be rescaled to a desired quality range, enabling seamless integration with existing human rating scales (e.g., 1–5).

By applying this strategy across multiple MLLMs, we can extract robust, model-agnostic quality scores that align closely with human perception. Experimental results confirm that this method significantly improves both the accuracy and stability of underwater image quality assessments, making it a simple yet effective tool for enhancing the evaluation capabilities of MLLMs.

Experiments and Analysis

In our zero-shot experimental setting, we evaluate a total of 10 MLLMs, including 8 widely used open-source models and 2 representative closed-source commercial systems: GPT-4o (developed by OpenAI) and Moonshot Kimi (developed by Moonshot AI). The primary objective of this study is to assess these models' low-level visual perception capabilities in underwater imagery. Specifically, we focus on their ability to recognize and reason about image clarity, color fidelity, contrast, and detail preservation. Unlike traditional vision-language tasks, low-level perception requires models to capture fine-grained, low-level features, imposing greater demands on the underlying multimodal fusion mechanisms.

All evaluations are conducted under a zero-shot setting, with no task-specific fine-tuning or additional training. This ensures that the results reflect the models' inherent generalization abilities when applied directly to UIQA. The selected open-source models include some of the most influential multimodal architectures in the community, such as Kosmos-2 (Peng et al. 2023), LLaVA (Liu et al. 2023), InstructBLIP (Dai et al. 2023), and others, widely used in image captioning and visual QA. GPT-4o and Moonshot serve as advanced closed-source references, providing strong baselines for industry-level performance in low-level perception.

We evaluate MLLMs on the underwater visual QA task from four perspectives: question type, low-level distortions, perception scope, and overall performance. As shown in Table 1, all models outperform random guessing, confirming their capability to extract semantic and structural cues from underwater images. However, their performance varies across different dimensions.

For question types, all models exceed chance levels. GPT-4o achieves the best accuracy on Yes-or-No questions (84.45%), indicating strong binary reasoning, while Moonshot leads on What questions (83.10%), reflecting superior semantic comprehension. How questions yield more varied results, revealing challenges in fine-grained quality assessment.

Regarding low-level distortions (color cast, contrast, exposure, blur/noise), open-source models like VisualGLM-6B and Qwen-VL show strengths on contrast and blur, whereas GPT-4o maintains consistently high scores, suggesting robust perceptual ability.

In global vs. local perception, GPT-4o performs best (79.23% and 81.44%), showing balanced holistic and detail understanding. VisualGLM-6B also performs competitively among open-source models.

Overall, closed-source models, especially GPT-4o, lead in performance, while some open-source models excel in specific aspects. These results reveal that performance differ-

Models (variant)	Question Types			Low-level Underwater Visual Attributes				Perception Scope		
	Yes-or-No↑	What↑	How↑	Color Cast↑	Contrast↑	Exposure↑	Blur/Noise↑	Global↑	Local↑	Overall↑
random guess	50.00%	31.78%	32.21%	37.21%	37.73%	37.57%	33.17%	36.29%	36.35%	36.17%
InstructBLIP (Flan-T5-XL)	56.39%	64.45%	51.73%	73.39%	55.19%	56.56%	47.07%	59.95%	49.72%	58.72%
IDEFICS-Instruct (LLaMA-7B)	55.09%	54.90%	45.95%	62.23%	49.24%	60.27%	41.17%	54.11%	44.04%	61.40%
Kosmos-2	57.29%	55.85%	55.03%	64.51%	62.89%	39.53%	48.21%	57.75%	40.72%	55.70%
LLaVA-v1.5 (Vicuna-v1.5-7B)	63.15%	74.72%	62.28%	**91.24%**	56.36%	68.88%	52.86%	67.83%	67.87%	67.83%
mPLUG-Owl (LLaMA-7B)	74.38%	65.80%	53.90%	79.75%	71.88%	55.77%	50.33%	66.73%	60.94%	66.03%
Qwen-VL (QwenLM)	71.59%	78.12%	63.15%	82.71%	71.06%	67.12%	65.47%	73.32%	65.65%	72.40%
InstructBLIP (Vicuna-7B)	58.50%	72.81%	56.36%	83.16%	57.71%	60.47%	51.33%	66.05%	49.30%	64.04%
VisualGLM-6B (GLM-6B)	83.12%	74.66%	68.26%	83.10%	**80.23%**	64.13%	71.17%	76.51%	72.41%	76.07%
GPT-4o (Closed-Source)	**84.45%**	77.73%	**75.29%**	80.89%	79.58%	**79.45%**	**77.82%**	**79.23%**	**81.44%**	**79.50%**
Moonshot (Closed-Source)	67.85%	**83.10%**	74.57%	89.08%	77.60%	68.69%	63.21%	76.77%	68.98%	75.83%

Table 1: Performance of MLLMs on the underwater visual QA task. The table reports scores across three question types (Yes-or-No, What, How), four underwater low-level distortion attributes (color cast, contrast, exposure, blur/noise), and two perception scopes (global and local). All results are reported as accuracy (%). Bold and underline indicate best and second-best results.

ences stem from variations in visual encoder design, data diversity, and multimodal fusion, and that task-specific alignment can partially compensate for smaller model capacity.

Underwater Image Description Results

Table 2 compares MLLMs' performance across three key evaluation dimensions: completeness, precision, and relevance. These metrics collectively assess the MLLMs' ability to perceive and articulate low-level visual details, which is critical for accurate image quality assessment in challenging underwater scenarios. The following section provides a detailed analysis of performance patterns and key observations across different models.

Model Performance Analysis **GPT-4o (Closed-Source):** GPT-4o achieves the highest scores across all three evaluation dimensions: Completeness (1.63), Precision (1.23), and Relevance (1.70), yielding a total score of 4.56. This indicates its strong ability to capture comprehensive low-level visual details, describe them accurately, and maintain high alignment with actual image characteristics. Such balanced and robust performance is crucial for reliable image quality assessment, particularly in complex underwater environments.

Moonshot (Closed-Source): Moonshot ranks second with a total score of 3.89. It shows notable strength in completeness, indicating effective summarization of overall image quality. However, it falls slightly behind in precision and relevance, suggesting occasional misinterpretation of subtle distortions such as color shifts or minor blur, which may limit its diagnostic accuracy.

Open-Source Models (e.g., InstructBLIP, IDEFICS-Instruct (Laurençon et al. 2024), and VisualGLM-6B): These models generally score between 2.7 and 3.1. Their lower performance in precision and relevance indicates a weaker ability to identify and articulate underwater-specific degradations such as color cast and low contrast. Although mPLUG-Owl (Ye et al. 2024) shows relatively balanced scores, its overall performance (3.41) still lags behind closed-source counterparts, likely due to limited adaptation to underwater scenes during training or architecture constraints.

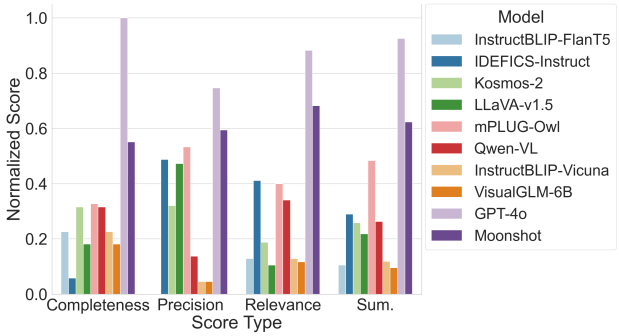

Figure 5: Visualization of MLLMs' performance on underwater image description across completeness, precision, and relevance.

Dimensions of Low-Level Visual Quality Assessment. The proposed evaluation metrics not only measure the descriptive ability of MLLMs but also indirectly reflect their effectiveness in underwater image quality assessment. Describing low-level attributes (e.g., color deviation, contrast, blur/noise) is essential for identifying key visual impairments and supporting downstream tasks. As shown in Figure 5, the qualitative visualization further confirms the quantitative trends in Table 2, illustrating how different models vary in capturing critical underwater distortions.

- **Completeness:** Capturing all relevant distortions is crucial in underwater imagery, where multiple quality issues often co-occur. A complete description enables accurate diagnosis and targeted enhancement.
- **Precision:** Accurate identification of specific issues ensures reliable evaluation. Misidentification or exaggeration can lead to incorrect assessments and suboptimal post-processing.
- **Relevance:** The semantic alignment between descriptions and actual impairments determines practical utility. High relevance ensures that model outputs align with human perception and real degradation patterns.

Across all models, GPT-4o and Moonshot achieve the most complete, precise, and relevant assessments, indicating

Models (variant)	Completeness				Precision				Relevance				Sum↑
	P_0	P_1	P_2	score↑	P_0	P_1	P_2	score↑	P_0	P_1	P_2	score↑	
InstructBLIP (Flan-T5-XL)	12.20%	80.90%	6.90%	0.94	35.08%	55.24%	9.68%	0.74	4.34%	88.50%	12.46%	1.06	2.74
IDEFICS-Instruct (LLaMA-7B)	23.08%	75.00%	1.92%	0.79	27.48%	38.40%	34.12%	1.06	3.48%	63.36%	33.16%	1.30	3.15
Kosmos-2	11.30%	74.92%	13.78%	1.02	25.30%	54.52%	20.18%	0.95	3.40%	83.40%	14.48%	1.11	3.08
LLaVA-v1.5 (Vicuna-v1.5-7B)	14.38%	80.78%	4.84%	0.90	19.42%	56.34%	24.24%	1.05	7.26%	81.16%	11.58%	1.04	2.99
mPLUG-Owl (LLaMA-7B)	3.32%	90.50%	6.18%	1.03	9.46%	72.00%	18.54%	1.09	1.16%	68.94%	29.90%	1.29	3.41
Qwen-VL (QwenLM)	11.76%	74.86%	13.38%	1.02	32.29%	51.58%	15.52%	0.83	2.74%	70.74%	26.52%	1.24	3.09
InstructBLIP (Vicuna-7B)	13.34%	79.46%	7.20%	0.94	34.84%	52.92%	12.24%	0.77	3.70%	86.14%	10.16%	1.06	2.77
VisualGLM-6B (GLM-6B)	15.96%	78.42%	5.62%	0.90	35.66%	51.72%	12.62%	0.77	4.92%	85.16%	9.92%	1.05	2.72
GPT-4o (Closed-Source)	1.18%	34.60%	64.22%	**1.63**	7.10%	62.72%	30.18%	**1.23**	0.30%	29.36%	70.34%	**1.70**	**4.56**
Moonshot (Closed-Source)	8.00%	61.26%	30.74%	1.23	20.32%	45.94%	33.74%	1.13	0.92%	44.60%	54.48%	1.53	3.89

Table 2: Evaluation of MLLMs on underwater image description. The table reports model-wise scores across three key dimensions: Completeness, Precision, and Relevance. P_i denotes the percentage of responses assigned a score of i, and the overall score is computed as the weighted average. Bold and underline indicate the best and second-best results.

more mature multimodal fusion and pretraining pipelines. Their advantages likely stem from large-scale, high-quality visual–text alignment data and extensive cross-domain tuning. In contrast, open-source models still struggle with subtle or compound distortions due to domain gaps and limited low-level supervision. These findings highlight the need for underwater-specific pretraining data, perceptual calibration, and adaptive alignment objectives tailored for low-level visual understanding.

Model	UID		UIQD	
	SRCC	PLCC	SRCC	PLCC
Kosmos-2	0.2101	0.2135	0.4389	0.4655
IDEFICS	-0.0230	-0.0409	-0.1808	-0.1608
mPLUG-Owl	0.2162	0.1941	**0.8579**	**0.8537**
Qwen-VL	-0.1067	-0.1088	0.0068	0.0247
LLaVA-v1.5	0.1279	0.1353	0.6115	0.5728
Vicuna-7B	0.1201	0.1328	0.7357	0.7295
Flan-T5-XL	0.1749	0.1383	0.4635	0.2956

Table 3: Performance of different multimodal models on UID and UIQD.

Analysis of MLLMs' Scoring Results

Table 3 reports the performance of various MLLMs on the UID (Hou et al. 2023) and UIQD datasets, evaluated using SRCC and PLCC metrics. We selected these two datasets to evaluate model performance across distinct quality scenarios: UID consists of enhanced underwater images, where the quality differences are often subtle and require fine-grained perceptual sensitivity; UIQD, on the other hand, includes real-world underwater images with significant degradations, such as blur, low contrast, and color cast. This dual-dataset setup allows us to assess not only a model's ability to detect distortion but also its understanding of quality improvement.

To derive predicted scores, we extracted logits corresponding to the "good" and "poor" labels from the models' responses and applied softmax normalization. The results show that most models achieve higher correlations on UIQD than UID, indicating that current MLLMs are more adept at identifying obvious degradations than evaluating subtle enhancements. This may be because real-world distortions are more visually salient and align better with the models' general visual-text pretraining.

Among all models, mPLUG-Owl performs best, especially on UIQD (SRCC: 0.8579, PLCC: 0.8537), suggesting it has a stronger capability to align with human perception. In contrast, models like IDEFICS and Qwen-VL perform poorly on both datasets, with near-zero or even negative correlations, reflecting weak sensitivity to visual quality. LLaVA-v1.5 and Vicuna-7B demonstrate moderate performance, indicating some ability to distinguish image quality, though still limited when compared to mPLUG-Owl.

In summary, while MLLMs show initial promise for no-reference image quality assessment, especially under severe distortion scenarios, their perceptual accuracy on fine-grained quality differences remains limited. These results highlight the importance of incorporating more quality-aware supervision and domain-specific tuning to improve their robustness across diverse image conditions.

Conclusion

We benchmarked ten MLLMs on underwater VQA, description, and quality scoring. All models demonstrate fundamental low-level perceptual abilities, yet closed-source systems consistently surpass open-source ones—likely due to larger model scales, broader multimodal alignment, and richer pretraining corpora. GPT-4o's dominance across distortion detection and descriptive completeness further illustrates the importance of tightly coupled vision-language fusion for robust perception. Nonetheless, UQ-Bench still faces limitations, such as a moderate dataset scale and incomplete coverage of rare underwater distortions. Addressing these issues through expanded data diversity and task-specific evaluation design will further strengthen its role as a rigorous benchmark for developing reliable, quality-aware underwater vision and perception systems.

Acknowledgments

This work was supported by the Natural Science Foundation of Shandong Province under grants ZR2023QF145,

ZR2022QF006, and ZR2024MF116, and the National Natural Science Foundation of China under grant 62201538.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Cao, J.; Wang, R.; Jia, Y.; Zhang, X.; Wang, S.; and Kwong, S. 2021. No-reference image quality assessment for contrast-changed images via a semi-supervised robust PCA model. *Information Sciences*, 574: 640–652.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Garg, S.; V, M. B.; Dharmasiri, T.; Hausler, S.; Suenderhauf, N.; Kumar, S.; Drummond, T.; and Milford, M. 2019. Look No Deeper: Recognizing Places from Opposing Viewpoints under Varying Scene Appearance using Single-View Depth Estimation. arXiv:1902.07381.

Hou, G.; Li, Y.; Yang, H.; Li, K.; and Pan, Z. 2023. UID2021: An Underwater Image Dataset for Evaluation of No-Reference Quality Assessment Metrics. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(4).

Hu, R.; Luo, T.; Jiang, G.; Lin, Z.; and He, Z. 2024. No-Reference Quality Assessment Based on Dual-Channel Convolutional Neural Network for Underwater Image Enhancement. *Electronics*, 13: 4451.

Islam, M. J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S. S.; and Sattar, J. 2020. Semantic Segmentation of Underwater Imagery: Dataset and Benchmark. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1769–1776.

Jian, M.; Liu, X.; Luo, H.; Lu, X.; Yu, H.; and Dong, J. 2021. Underwater image processing and analysis: A review. *Signal Processing: Image Communication*, 91: 116088.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1): 32–73.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9579–9589.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? arXiv:2405.02246.

Li, C.; Anwar, S.; Hou, J.; Cong, R.; Guo, C.; and Ren, W. 2021. Underwater Image Enhancement via Medium Transmission-Guided Multi-Color Space Embedding. *IEEE Transactions on Image Processing*, 30: 4985–5000.

Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; and Tao, D. 2019a. An Underwater Image Enhancement Benchmark Dataset and Beyond. arXiv:1901.05495.

Li, C.-Y.; Guo, J.-C.; Cong, R.-M.; Pang, Y.-W.; and Wang, B. 2016. Underwater Image Enhancement by Dehazing With Minimum Information Loss and Histogram Distribution Prior. *IEEE Transactions on Image Processing*, 25(12): 5664–5677.

Li, D.; Jiang, T.; Lin, W.; and Jiang, M. 2019b. Which Has Better Visual Quality: The Clear Blue Sky or a Blurry Animal? *IEEE Transactions on Multimedia*, 1221–1234.

Li, W.; Han, W.; Deng, L.-J.; Xiong, R.; and Fan, X. 2025. Spiking Variational Graph Representation Inference for Video Summarization. *IEEE Transactions on Image Processing*, 34: 5697–5709.

Li, W.; Ma, Z.; Deng, L.-J.; Fan, X.; and Tian, Y. 2023. Neuron-Based Spiking Transmission and Reasoning Network for Robust Image-Text Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7): 3516–3528.

Li, W.; Wang, P.; Xiong, R.; and Fan, X. 2024. Spiking Tucker Fusion Transformer for Audio-Visual Zero-Shot Learning. *IEEE Transactions on Image Processing*, 33: 4840–4852.

Li, X.; Hou, G.; Tan, L.; and Liu, W. 2020. A Hybrid Framework for Underwater Image Enhancement. *IEEE Access*, 8: 197448–197462.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C. A.; Manning, C. D.; Re, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; WANG, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N. S.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R. A.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Liu, Y.; Zhang, B.; Hu, R.; Gu, K.; Zhai, G.; and Dong, J. 2024. Underwater Image Quality Assessment: Benchmark Database and Objective Method. *IEEE Transactions on Multimedia*, 26: 7734–7747.

Lu, J.; Rao, J.; Chen, K.; Guo, X.; Zhang, Y.; Sun, B.; Yang, C.; and Yang, J. 2024. Evaluation and Enhancement of Semantic Grounding in Large Vision-Language Models. arXiv:2309.04041.

Mazel, C. H. 2005. Underwater fluorescence photography in the presence of ambient light. *Limnology and Oceanography: Methods*, 3(11): 499–510.

Nomura, K.; Sugimura, D.; and Hamamoto, T. 2018. Underwater Image Color Correction using Exposure-Bracketing Imaging. *IEEE Signal Processing Letters*, 25(6): 893–897.

Ogunsina, M.; Efunniyi, C.; Osundare, O.; Folorunsho, S.; and Akwawa, L. 2024. Robust Multimodal Perception in Autonomous Systems: A Comprehensive Review and Enhancement Strategies. *Engineering Science & Technology Journal*, 5: 2694–2708.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv:2306.14824.

Schwenker, F. 2013. Ensemble Methods: Foundations and Algorithms [Book Review]. *IEEE Computational Intelligence Magazine*, 8(1): 77–79.

Tsimpoukelli, M.; Menick, J.; Cabi, S.; Eslami, S. M. A.; Vinyals, O.; and Hill, F. 2021. Multimodal Few-Shot Learning with Frozen Language Models. arXiv:2106.13884.

Wang, Y.; Ke, J.; Talebi, H.; Yim, J. G.; Birkbeck, N.; Adsumilli, B.; Milanfar, P.; and Yang, F. 2021. Rich features for perceptual quality assessment of UGC videos. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wu, H.; Chen, C.; Liao, L.; Hou, J.; Sun, W.; Yan, Q.; Gu, J.; and Lin, W. 2022. Neighbourhood Representative Sampling for Efficient End-to-end Video Quality Assessment.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Zhang, X.; Cai, N.; Zhang, H.; Zhang, Y.; Di, J.; and Lin, W. 2023. AFD-Former: A Hybrid Transformer With Asymmetric Flow Division for Synthesized View Quality Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 3786–3798.

Zhang, X.; Ma, J.; Wang, G.; Zhang, Q.; Zhang, H.; and Zhang, L. 2025a. Perceive-IR: Learning to Perceive Degradation Better for All-in-One Image Restoration. *IEEE Transactions on Image Processing*, 1–1.

Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L.; and Du, B. 2025b. UniUIR: Considering Underwater Image Restoration as an All-in-One Learner. *IEEE Transactions on Image Processing*, 34: 6963–6977.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.