UVLM: Benchmarking Video Language Model for Underwater World Understanding

Xizhe Xue[1*], Yang Zhou[1*], Dawei Yan[1], Lijie Tao[1], Junjie Li[1], Ying Li[1], Haokui Zhang[1†], Rong Xiao[2]

[1]Northwestern Polytechnical University
[2]Intellifusion Inc.

Abstract

Recently, video-language models (VidLMs) have gained widespread attention and adoption. However, existing works primarily focus on terrestrial scenarios, overlooking the highly demanding application needs of underwater observation. To overcome this gap, we introduce UVLM, an under water observation benchmark which is build through a collaborative approach combining human expertise and AI models. To ensure data quality, we have conducted in-depth considerations from multiple perspectives. First, to address the unique challenges of underwater environments, we selected videos that represent typical underwater challenges including light variations, water turbidity, and diverse viewing angles to construct the dataset. Second, to ensure data diversity, the dataset covers a wide range of frame rates, resolutions, 419 classes of marine animals, and various static plants and terrains. Next, for task diversity, we adopted a structured design where observation targets are categorized into two major classes: biological and environmental. Each category includes content observation and change/action observation, totaling 20 subtask types. Finally, we designed several challenging evaluation metrics to enable quantitative comparison and analysis of different methods. Experiments on two representative VidLMs demonstrate that fine-tuning VidLMs on UVLM significantly improves underwater world understanding while also showing potential for slight improvements on existing in-air VidLM benchmarks.

Code —
https://github.com/Cecilia-xue/UVLM-Benchmark
Extended version — https://arxiv.org/abs/2507.02373

Introduction

Video-language understanding (Xu et al. 2016; Anne Hendricks et al. 2017) stands at the forefront of multimedia research, empowering systems to interpret, reason about, and generate natural language descriptions of temporal and dynamic visual content (Patraucean et al. 2023). Recent advances (Patraucean et al. 2023; Li et al. 2024; Zhang et al. 2025a) in video-language models (VidLMs) have demonstrated impressive performance in tasks such as video captioning, temporal grounding, and visual question answering, focusing primarily on human-centric scenarios and common object interactions. Despite these achievements, a critical question remains: **Can current VidLMs effectively understand videos captured in special imaging conditions, such as underwater environments?**

This question is especially relevant as underwater environments constitute an uncharted domain with immense scientific value (e.g., marine biodiversity monitoring, ecosystem health assessment (Xue et al. 2024)) and substantial engineering applications (e.g., autonomous underwater vehicles). As in Figure 1, applying VidLMs to underwater content presents three challenges that distinguish this domain from conventional video-language tasks:

- **Degraded Visual Features Hindering Analytical Decisions.** The underwater domain introduces fundamental perceptual barriers that confound existing video understanding approaches. Underwater scenes exhibit variable illumination with rapid light attenuation at depth, wavelength-dependent color distortion that shifts the visual spectrum, fluctuating turbidity affecting visibility (Islam, Xia, and Sattar 2020; Akkaynak and Treibitz 2019). Standard VidLMs designed for terrestrial scenario environments struggle to perform effectively underwater due to these low-quality visual cues.
- **Lack of Scientific Domain Knowledge.** Underwater content requires specialized ecological expertise for accurate interpretation. Unlike common scenarios featuring familiar objects and actions, underwater videos capture complex species interactions, specialized behavioral patterns, and environmental relationships that demand expert knowledge to decode (Marks et al. 2022; Han et al. 2024). The interpretation of underwater footage requires understanding multiple layers of information, including taxonomic identification, morphological characteristics, behavioral states, and environmental contexts. Models must capture intricate ecological relationships between marine organisms' appearance, behaviors, and habitats, creating a significant knowledge gap for systems trained primarily on common objects and human activities.
- **Data Resource Scarcity.** The development of effective VidLMs for underwater environments is critically hindered by the absence of comprehensive

[*]These authors contributed equally.
[†]Corresponding author

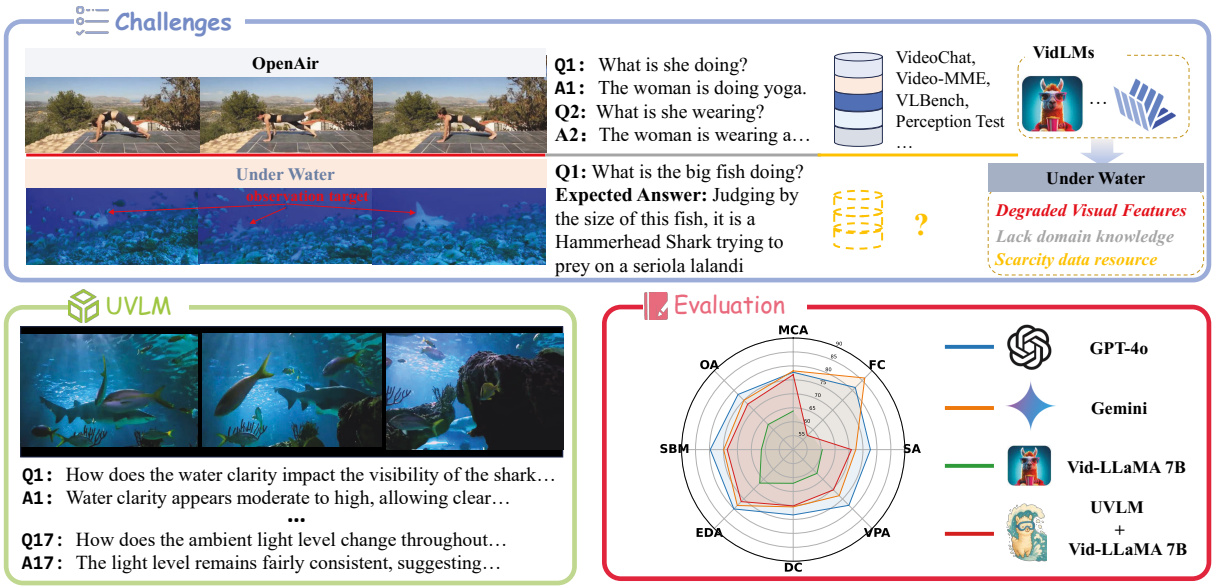

Figure 1: Challenges for VidLMs for understanding underwater videos. Our UVLM is proposed to overcome this gap and it enables the 7B VidLM to achieve performance comparable to closed-source models like GPT-4o and Gemini.

training resources. Current video-language datasets predominantly focus on everyday human activities (HowTo100M (Miech et al. 2019)), sports (Sports-1M (Karpathy et al. 2014)), or general-purpose actions (Kinetics (Carreira and Zisserman 2017)), offering minimal support for specialized scientific domains. While underwater-specific datasets exist, they typically prioritize narrow tasks such as object tracking (WebUOT (Zhang et al. 2024)) or instance segmentation (UIIS (Lian et al. 2023a)). These image-based datasets fail to equip models with scientific domain knowledge necessary for comprehensive understanding of marine organism behaviors in underwater videos. This data gap creates a major barrier to advancing VidLMs beyond common scenarios to specialized scientific contexts.

To bridge these gaps, a comprehensive Underwater Video-Language Multimodal (UVLM) dataset with professional annotations is presented. Compared to existing video language understanding benchmarks, UVLM demonstrates distinct differences in both content composition and construction methodology. First, in terms of content, UVLM is the first video-language benchmark specifically designed for underwater environments. To ensure the dataset accurately captures the distinctive characteristics of underwater settings, we carefully selected videos that encompass unique challenges of this domain, including low-light conditions, water turbidity, and the highly variable movement patterns of marine organisms. Second, regarding construction methodology, we adopted a structured framework combining human-AI collaboration. The annotation targets encompass both biological and environmental elements, covering static and dynamic scenarios. After frame-by-frame manual annotation, we leveraged these human-annotated labels to guide GPT-4o in generating diverse sample content. Finally, all data underwent manual verification, with inconsistent entries either re-annotated or supplemented using search engines for factual accuracy.

The final dataset comprises 0.9M frames, 419 distinct biological categories, and diverse underwater scenarios, encompassing 20 types of video-language understanding tasks. Additionally, we established eight specialized metrics for quantitative performance comparison on the dataset. UVLM advances video-language research through its comprehensive coverage of core technical challenges: **1) Temporal Understanding**: Supports development of models that can interpret continuous behavior sequences and environmental changes over time; **2) Fine-Grained Recognition**: Enables research on distinguishing subtle visual differences with significant scientific meaning. **3) Compositional Reasoning**: Facilitates the development of models that can decompose complex scenes into scientifically meaningful components; **4) Knowledge-Grounded Generation**: Provides a foundation for generating technically accurate language descriptions based on visual evidence.

In summary, UVLM represents a significant step toward extending the capabilities of VidLMs beyond everyday scenarios to specialized scientific domains. By providing richly annotated video-language pairs in underwater environments, UVLM enables the development of models that can interpret complex ecological dynamics and communicate this understanding through natural language, ultimately contributing to both multimedia research and marine science.

Related Work

Video Language Understanding Benchmark

In recent years, many video language benchmarks have emerged, each targeting specific themes and application domains. Some focus primarily on everyday life scenarios (Xu

Dataset	Venue	Im	Vid	Lang	OQA	BioInf	Seq	Frame	Task	Categ
LSUI (Peng et al. 2023)	TIP/23	✓	✗	✗	✗	✗	-	5k	IR	10
DRUVA (Varghese et al. 2023)	ICCV/23	✓	✓	✗	✗	✗	20	6K	DE, IR	20
IOCfish5K (Sun et al. 2023)	CVPR/23	✓	✗	✗	✗	✗	-	5K	OC	-
UIIS (Lian et al. 2023a)	ICCV/24	✓	✗	✗	✗	✗	-	4.6k	IS	7
VMAT (Cai et al. 2023)	IJCV/23	✓	✓	✗	✗	✗	33	57K	SOT	17
WebUOT (Zhang et al. 2024)	NeurIPS/24	✓	✓	✓	✗	✗	1500	1M	SOT	408
MarineInst (Zheng et al. 2024)	ECCV/24	✓	✗	✓	✗	✗	-	2.42 M	IS, CAP	-
USOD (Hong et al. 2025)	TIP/25	✓	✗	✗	✗	✗	-	10K	SOD	70
UVLM	-	✓	✓	✓	✓	✓	2111	0.86M	VU	419

Table 1: Comparison of recent underwater observation datasets. Im, Vid, Lang, OQA, BioInf, Seq, and Categ denote Image, Video, Language, Open-ended Question Answering, Taxonomic Classification Information in Biology, Sequence, and Category, respectively. In Task, IR, DE, OC, IS, SOT, CAP, SOD, and VU refer to Image Restoration, Depth Estimation, Object Counting, Instance Segmentation, Single Object Tracking, Captioning, Salient Object Detection, and Video Understanding, respectively.

et al. 2017; Xiao et al. 2021; Yu et al. 2019), while others emphasize human action or movie clips (Mangalam, Akshulakov, and Malik 2023; Song et al. 2024). More comprehensive datasets cover a wider range of categories, including knowledge, sports, and instructional videos (Li et al. 2024; Wu et al. 2024; Fu et al. 2025; Wang et al. 2025). Although these efforts have significantly advanced video-language research in terrestrial scenario contexts, the underwater environment remains largely uncharted, motivating our investigation of underwater video-language benchmark.

Underwater Observation Datasets

Underwater observation datasets and benchmarks have followed a clear progression, evolving from low-level tasks to high-level ones, from single-frame image analysis to video content analysis, and from unimodal to multimodal approaches. We summarize several key underwater image datasets and compare them with our UVLM in Table 1.

Low-level perception datasets focus on fundamental image enhancement and quality assessment, such as LSUI (Peng et al. 2023). These datasets provide essential resources for developing and validating algorithms to address the typical degradation in underwater imagery. More recent efforts have extended this work into the temporal domain with video datasets like DRUVA (Varghese et al. 2023), which contains 6,000 frames. Mid-level recognition tasks have also benefited from dedicated datasets. For instance, Wildfish (Zhuang, Wang, and Qiao 2018) is tailored for marine life recognition, while COU (Mukherjee et al. 2025) supports segmentation research. Additionally, a range of tracking datasets, such as VMAT (Cai et al. 2023) and WebUOT (Zhang et al. 2024) further advance object-centric underwater understanding. At the high-level, multimodal tasks are beginning to emerge. MarineInst (Zheng et al. 2024) marks a critical step forward by supporting advanced tasks such as image segmentation and captioning, thereby opening new avenues for comprehensive underwater analysis.

Although these advancements have significantly propelled multimodal underwater analysis, multimodal underwater video analysis remains largely unexplored. Underwater videos inherently capture rich temporal dynamics, such as marine life trajectories and multi-view morphological changes, which provide additional contextual cues. By jointly leveraging visual appearance, temporal evolution, and domain-specific textual knowledge, unique advantages emerge. This holistic approach holds considerable promise for enhancing marine life interpretation, ecological monitoring, and overall environmental understanding.

UVLM

Overview

Currently, research on VidLMs and the construction of corresponding benchmarks has been ongoing for several years, achieving significant progress. However, existing work primarily focuses on land scenarios, neglecting underwater environments. Extending VidLM technology to the underwater domain still faces challenges such as visual feature degradation caused by harsh aquatic conditions, the lack of domain knowledge due to domain gaps, and the scarcity of benchmarks due to difficulties in data acquisition.

In this paper, taking these challenges into consideration, we developed UVLM, the first underwater video-language benchmark. Firstly, we collect videos from typical underwater environments like oceans, lakes, and rivers, which exhibit different degradation problems such as water surface ripples, turbid water, and light scattering. Secondly, we select videos containing different content types, including static environmental features, static observations of organisms, dynamic scene variations, and dynamic biological behaviors, etc.

To systematically inject domain knowledge into the dataset, we design textual descriptions structured into 20 subtasks, reflecting VidLMs' capabilities across 9 dimensions, such as marine animal behavior, water bodies, and geological features. Thirdly, to maximize the scale of our dataset, we adopt a dual approach: systematically collecting data from online sources while strategically screening and re-annotating existing relevant datasets.

The final dataset comprises approximately two thousand carefully selected video sequences and 0.9 million frames, covering 419 different marine organisms and about 40 thousand video-language pairs. Previous underwater observation datasets (Lian et al. 2023b; Zhao et al. 2021; Alawode et al. 2022, 2023; Zhang et al. 2025b; Fan et al. 2019; Huang,

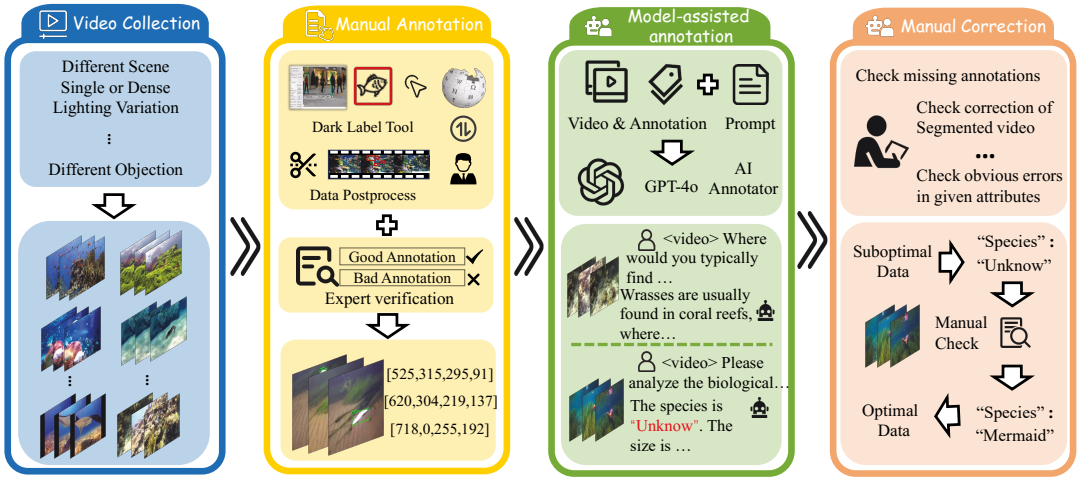

Figure 2: An overview of data preparation and generation pipeline for UVLM.

Zhao, and Huang 2021) typically focus on traditional computer vision tasks such as image segmentation and single object tracking, which are pure vision tasks. A few underwater observation datasets (Zheng et al. 2024; Li et al. 2025) also incorporate visual and language elements, but the tasks are still limited to typical applications such as captioning. Unlike these datasets, the proposed UVLM is designed to incorporate specific underwater domain knowledge and requirements. To achieve this, we designed 16 to 20 relative questions referring to underwater research topics such as marine organism recognition, behavioral analysis and prediction, habitat pattern characterization, subaquatic environmental monitoring, and ecosystem dynamics assessment, etc., covering both biological and environmental dimensions critical to underwater exploration and conservation. A comparison with previous underwater datasets is shown in Table 1.

Data Collection and Annotation

As shown in Figure 2, the benchmark construction consists for four major parts: video collection, manual annotation, model-assisted annotation and manual correction.

Video collection. Our objective extends beyond merely constructing an underwater video-language benchmark and training VidLMs on the data. Instead, we aim to build a dataset that captures the unique challenges of the underwater environment and enables VidLMs to operate in underwater contexts to support related research. To achieve this, video collection must satisfy two criteria: 1) the videos must capture distinctive underwater challenges; 2) the dataset must incorporate an adequate volume of representative cases to enable effective VidLM fine-tuning. To meet these requirements, we implement a dual-path video acquisition strategy.

The first path is collecting underwater videos from websites such as youtube and bilibili. We collect web-crawled data while enforcing two quality control principles during acquisition: 1) *Both camera and observation objects or scenes must be under water.* This selection rule is more aligned with real world underwater vehicle and robot application scenarios; 2) *The typical underwater scenarios and targets with different challenges and characteristics must be covered.* The videos should cover typical environments, like oceans, lakes, rivers and fish tanks, etc. Meanwhile, the chosen targets should include (but not limited to) common animals (fish, whales, prawns, tortoises, etc.), people and generic objects. With the principles, we selected about 400 videos, each of which contains 100 to 3000 frames, covering 53 classes of objects. The second path is re-annotating existing dataset. For scalability considerations, WebUOT (Zhang et al. 2024) are taken as primary data source.

Manual annotation. The substantial domain gap between terrestrial and underwater environments requires auxiliary annotation information to effectively leverage existing VidLMs for collaborative labeling. To facilitate granular analysis of underwater targets, such as studying the behavioral traits and ecological patterns of marine organisms, we annotate each frame with both bounding boxes and fine-grained taxonomic classifications aligned with marine biology standards. These annotations function as priors when fed into AI models for text generation.

For videos collected from websites, we employed 12 annotators annotate videos following 3 principles: 1) For occluded objects, only visible part is marked with a rectangular box; 2) For the sharply protruding part, such as the tail and mouth of fish, whether to contain it in the bounding box is determined based on the proportion of the target and background in the additional introduced area. In specific, if the object's protruding parts accounts for more than one-third of the additional introduced area, the sharply protruding part would be marked, otherwise it should not be marked; 3) Each frame is annotated with an axis aligned bounding box using the DarkLabel toolbox[1]. Each sequence is assessed by 3 domain experts to minimize annotation errors. Unqualified labels would be sent to other annotators for re-labeling.

For videos from WebUOT, we followed several steps to refine video quantity. We first performed data cleaning to filter out videos with information interference, such as those

[1]https://github.com/darkpgmr/DarkLabel

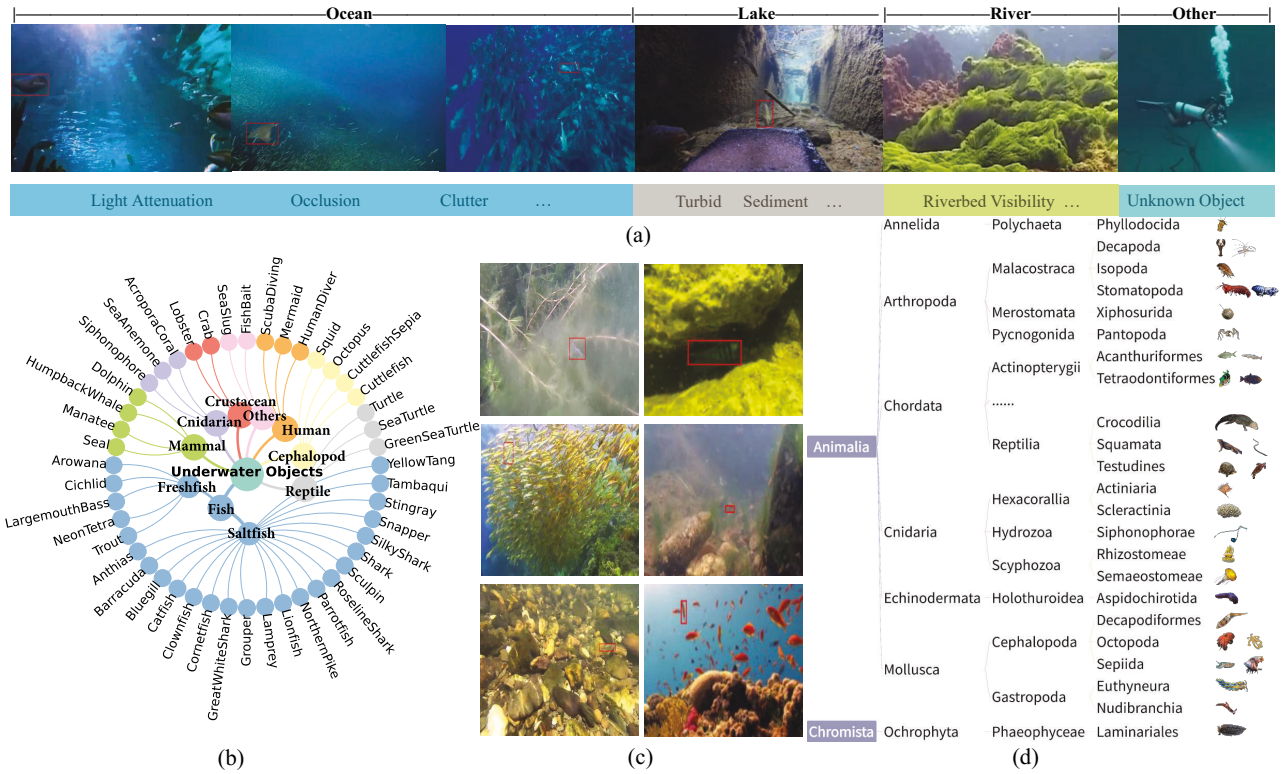

Figure 3: Statistics on UVLM. (a) Scene distribution; (b) Observation target distribution; (c) Several samples in UVLM; (d) Fine-grained taxonomic classification information (partial categories), from left to right: Kingdom, Phylum, Class, Order.

containing large areas of subtitles or watermarks that affect video quality. Then, we conducted scene cleaning to remove videos that were not in natural underwater environments, such as those filmed in aquariums, fish tanks, or simulated gaming scenes. Next, we adopted SAM (Kirillov et al. 2023) and LaMa (Suvorov et al. 2022) to remove small-area watermarks, subtitles, and other information interference. We use mouse clicks to select the area, SAM for segmentation, and then LaMa for inpainting. Next, to avoid challenges caused by excessively long videos and focus specifically on evaluating the impacts from underwater environments, We segmented some longer videos with fewer observed organisms into 300-600 frame clips to balance the sample distribution. Some videos collected from the internet even contain up to 3,000 frames, so the dataset can also evaluate the models' ability to understand extended underwater scenarios.

To achieve reliable and fine-grained taxonomic classification of marine animals in these videos, we implemented a structured three-phase annotation procedure. In the first phase, four annotators with extensive marine biology knowledge independently labeled each target within the video frames, providing both species-level identifications and detailed taxonomic classifications according to the classic five-kingdom system (kingdom, phylum, class, order, etc.) in biology (Whittaker 1969), supported by authoritative sources such as Wikipedia. In the second phase, annotations were cross-validated by annotator pairs. Any disagreements were resolved by consulting a third annotator for majority consensus, with unresolved cases flagged for expert review. In the third phase, a senior marine biology expert reviewed the validated annotations, marking questionable cases. Finally, all flagged annotations were collectively discussed by five experts to establish definitive classifications. This rigorous approach ensured high accuracy and consistency in fine-grained taxonomic annotation.

Model-assisted annotation. After preparing the videos, we carefully designed prompts to guide GPT-4o in generating relevant questions and answers based on the input video content. To ensure the generated questions have practical domain relevance, we first conducted research on key topics in marine biology, such as observed species, organism behavior, and habitat characteristics covered in the videos. Guided by these findings, we designed prompts to instruct GPT-4o to generate content along two dimensions: 1) *Marine Organism Dimension*. Static aspects: Species identification, biological attributes (e.g., morphology, coloration). Dynamic aspects: Behavioral analysis (e.g., feeding, interactions), movement patterns. 2) *Underwater Environment Dimension*. Static aspects: Environmental features (e.g., substrate type, coral structures), habitat traits. Dynamic aspects: Light condition variations, visibility fluctuations, etc.

The questions include two formats: multiple-choice and open-ended. For instance, multiple-choice: *"What marine species is primarily observed in the images? A) Clownfish B) Chromis dimidiatus C) Parrotfish"*. Open-ended: *"What is the overall setting of the video, and how does it influ-*

Method	Objective Metrics		LLM-based Judgement Metrics					
	MCA	FGC	SA	DC	VPA	EDA	SBM	Overall Accuracy
Closed-source VidLMs and other Open Source Large VidLMs								
GPT-4o	77.72	81.47	77.67	73.40	78.23	80.07	79.73	77.95
Claude3.7-Sonnet	76.61	82.64	73.35	73.58	74.10	79.71	76.35	76.09
Gemini2.5-Flash	78.22	86.27	72.43	73.34	70.53	78.32	74.92	75.00
Qwen2.5VL-72B	75.97	80.57	74.22	71.94	74.85	78.45	77.40	75.49
Base VLM								
Qvis2.5-2B	61.87	29.85	52.88	49.70	54.66	55.81	50.62	53.06
Qvis2.5-2B + UVLM	72.89	59.25	67.64	62.39	65.30	69.05	66.29	68.85 (**+15.79**)
Base VidLMs								
InternVL2.5-1B	48.54	29.61	44.35	47.30	45.30	50.15	44.85	46.73
VideoLLaMA3-2B	57.17	31.25	55.89	58.48	58.53	63.15	55.74	58.39
Qwen2.5VL-2B	59.47	35.23	54.12	50.65	53.47	57.31	50.60	52.97
InternVL2.5-8B	57.64	36.63	57.70	59.75	59.45	63.48	61.05	60.15
VideoLLaMA3-7B	63.83	42.62	60.43	62.07	62.05	66.97	61.41	62.70
Qwen2.5VL-7B	66.22	48.36	66.67	59.98	61.74	68.41	61.79	63.57
InternVL2.5-1B + UVLM	64.52	45.37	54.48	56.55	55.78	65.16	58.74	59.14 (**+12.41**)
VideoLLaMA3-2B + UVLM	70.41	46.35	63.54	64.76	66.71	70.28	67.32	66.67 (**+8.28**)
Qwen2.5VL-2B + UVLM	62.38	56.08	60.47	55.81	57.93	61.45	58.26	58.44 (**+5.47**)
InternVL2.5-8B + UVLM	70.26	43.94	65.66	66.42	65.05	71.10	69.98	69.45 (**+9.30**)
VideoLLaMA3-7B + UVLM	76.85	57.17	70.88	70.17	70.40	76.35	73.66	73.04 (**+10.34**)
Qwen2.5VL-7B + UVLM	71.69	63.41	70.47	67.25	67.29	72.16	65.76	68.08 (**+4.51**)

Table 2: Performance comparison of different methods on UVLM test set. Metric abbreviations: MCA, FGC, SA, DC, VPA, EDA, SBM denote Multiple Choice Accuracy, Fine-grained Taxonomic Classification, Semantic Accuracy, Detail Completeness, Visual Perception Accuracy, Environmental Description Accuracy, and Species Behavior Matching, respectively.

ence Chromis dimidiatus' activities?". During the Q&A process, we implemented clear prompt constraints, including: 1) Topic-specific constraints (e.g., focusing on taxonomic details, observed behaviors, or physicochemical properties of the water body); 2) Style requirements (e.g., diversifying sentence structures and enriching question types). Each video generates 16 to 20 video-text pairs.

Manual correction. Despite the integration of domain experts and model-assisted generation in our annotation pipeline, the intrinsic complexity of underwater scenes and the limitations of automatic generation inevitably introduce semantic drift and consistency issues. To guarantee the ecological validity and scientific rigour of the final dataset, we therefore incorporated a dedicated manual-correction stage as the decisive quality-assurance step.

Stringent quality control was applied to all Q&A pairs generated by GPT-4o through a two-tier human review protocol. The first round, conducted by general reviewers, focused on detecting conflicts between the information supplied to GPT-4o and the content it produced. The second round involved senior experts performing in-depth edits to ensure factual precision and domain conformity. During both rounds, hallucinations or statements irrelevant to video context were either removed or rewritten. Whenever a model description diverged from ground truth, we leveraged the accompanying assistant information to cross-validate and revise the erroneous content, safeguarding dataset integrity.

The scene and targets distribution of UVLM are presented in Figure 3. The final dataset comprises 2,109 videos (0.86M frames total), with lengths ranging from 100 to 3,000 frames. Spanning 419 categories across 4 major underwater scenes, each video contains 16–20 video-language pairs, yielding approximately 40k video-text pairs in total.

Evaluation Metrics

UVLM's comprehensive evaluation employs two Objective Metrics and five LLM-based Judgement Metrics. Two objective metrics are: 1) Multiple Choice Accuracy, evaluate the model's performance on various common multiple-choice questions; 2) Fine-grained taxonomic classification, evaluate the model's performance on multiple-choice questions related to the professional biological taxonomy (Whittaker 1969). The LLM-based Judgement Metrics are inspired by LLM judge (Zheng et al. 2023) and MMDU (Liu et al. 2024), leveraging GPT-4o as the evaluation backbone. Specifically, they assess model performance from five aspects; 1) Semantic accuracy, assesses how closely the description matches the reference answer; 2) Visual perception accuracy, examines the correctness of interpretation of image content; 3) Detail completeness, measures the thoroughness of the provided information; 4) Environmental description accuracy, focuses on the correctness of underwater environmental descriptions; 5) Species behavior matching, evaluates the accuracy of prediction.

Experiments

Dataset Partition

UVLM contains 2109 videos, covering 419 categories. As shown in Figure 3 (b), the overall category distribution fol-

Figure 4: Visualization examples of fine-tuning with UVLM. Error and correct descriptions are marked in red and green.

lows a long-tail distribution and videos of many specific categories have relatively few observed targets. To ensure the training and test sets have consistent category distributions as much as possible, we sampled the test set proportionally. For categories with more than 5 videos, we randomly selected videos for the test set according to a predefined train-test split ratio. Then, we further randomly sampled videos from categories with fewer than 5 videos as test samples. Finally, 208 videos-text pairs were selected as test set.

Experimental Results on UVLM

Table 2 and Figure 4 present the results on UVLM. We selected three recently released VidLMs (InternVL (Chen et al. 2024), QwenVL (Bai et al. 2025), VideoLLaMA (Zhang et al. 2025a)) and a representative VLM (Qvis (Lu et al. 2024)) as baselines. According to the experimental results, we can draw several observations:

- **Underwater observation poses distinct challenges compared to in-air scenarios**. Even state-of-the-art closed-source models, GPT-4o (77.95) and Gemini (75.00), or open-source models like Qwen2.5VL-72B (75.49), achieve relatively limited performance. The gap highlights unique complexities of underwater scenarios, suggesting significant potential for further exploration.
- **Fine-tuning with UVLM significantly enhances VidLMs' underwater observation capabilities**. For example, VideoLLaMA3-7B achieves an accuracy gain exceeding 10 points, reaching 73.04, just 2.45 points behind the Qwen2.5VL-72B (75.49). Remarkably, this compact model also closely matches closed-source models like Gemini (75.00), despite their larger scales and proprietary datasets. These results indicate that the proposed human-AI collaborative annotation pipeline effectively distills knowledge from large models like GPT-4o and injects it into smaller models through fine-tuning. While maintaining acceptable performance trade-offs,it reduces model overhead and hardware requirements and expand model's applicability scope.
- **In highly specialized fields, simply increasing the amount of training data is insufficient to bridge the gap between small models and large models**. Differing from other six metrics, performance on the fine-grained taxonomic classification task requires complex biological domain knowledge. For this particular task, we observe a persistent performance gap between small and large models even after extensive fine-tuning. We attribute this phenomenon to the fundamental capacity differences between model scales. While simple QA tasks can be effectively addressed through data fine-tuning alone as they impose relatively low demands on model capacity, tasks requiring sophisticated domain expertise present a greater challenge. Small models, constrained by their limited capacity while simultaneously maintaining performance across multiple tasks, demonstrate particular difficulty in closing this performance gap through fine-tuning alone. This finding presents a new challenge for the field of underwater world understanding.

Conclusion

This paper presents the first multimodal VidLM benchmark for underwater environments. Combining human expertise with advanced AI annotation methods, UVLM encompasses extensive marine biodiversity, diverse video resolutions, and realistic underwater challenges, including varying illumination and water turbidity, thus providing authentic conditions for model training and evaluation. UVLM enables meaningful quantitative comparisons, fostering the development of accurate and reliable underwater observation systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant (62401471, 62271400); in part by 2024 Gusu Innovation and Entrepreneurship Leading Talents Program under Grant ZXL2024333. The authors sincerely thank Xiaoyue Yin and Ruikang Mao for their discussions and remarkable support in collecting the data.

References

Akkaynak, D.; and Treibitz, T. 2019. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1682–1691.

Alawode, B.; Dharejo, F. A.; Ummar, M.; Guo, Y.; Mahmood, A.; Werghi, N.; Khan, F. S.; Matas, J.; and Javed, S. 2023. Improving Underwater Visual Tracking With a Large Scale Dataset and Image Enhancement. arXiv:2308.15816.

Alawode, B.; Guo, Y.; Ummar, M.; Werghi, N.; Dias, J.; Mian, A.; and Javed, S. 2022. UTB180: A High-quality Benchmark for Underwater Tracking. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 3326–3342.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report.

Cai, L.; McGuire, N. E.; Hanlon, R.; Mooney, T. A.; and Girdhar, Y. 2023. Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *International Journal of Computer Vision*, 131(6): 1406–1427.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6299–6308.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. 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*, 24185–24198.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5374–5383.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis.

Han, Y.; Chen, K.; Wang, Y.; Liu, W.; Wang, Z.; Wang, X.; Han, C.; Liao, J.; Huang, K.; Cai, S.; et al. 2024. Multi-animal 3D social pose estimation, identification and behaviour embedding with a few-shot learning framework. *Nature Machine Intelligence*, 6(1): 48–61.

Hong, L.; Wang, X.; Zhang, G.; and Zhao, M. 2025. USOD10K: A New Benchmark Dataset for Underwater Salient Object Detection. *IEEE Transactions on Image Processing*, 34: 1602–1615.

Huang, L.; Zhao, X.; and Huang, K. 2021. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1562–1577.

Islam, M. J.; Xia, Y.; and Sattar, J. 2020. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2): 3227–3234.

Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1725–1732.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

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

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Lian, S.; Li, H.; Cong, R.; Li, S.; Zhang, W.; and Kwong, S. 2023a. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1305–1315.

Lian, S.; Li, H.; Cong, R.; Li, S.; Zhang, W.; and Kwong, S. 2023b. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1305–1315.

Liu, Z.; Chu, T.; Zang, Y.; Wei, X.; Dong, X.; Zhang, P.; Liang, Z.; Xiong, Y.; Qiao, Y.; Lin, D.; et al. 2024. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. In *Advances in Neural Information Processing Systems*.

Lu, S.; Li, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Ye, H.-J. 2024. Ovis: Structural embedding alignment for multimodal large language model.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36: 46212–46244.

Marks, M.; Jin, Q.; Sturman, O.; von Ziegler, L.; Kollmorgen, S.; von der Behrens, W.; Mante, V.; Bohacek, J.; and Yanik, M. F. 2022. Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nature machine intelligence*, 4(4): 331–340.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2630–2640.

Mukherjee, R.; Singh, S.; McWilliams, J.; and Sattar, J. 2025. The Common Objects Underwater (COU) Dataset for Robust Underwater Object Detection. arXiv:2502.20651.

Patraucean, V.; Smaira, L.; Gupta, A.; Recasens, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Malinowski, M.; Yang, Y.; Doersch, C.; et al. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761.

Peng, L.; Zhu, C.; Bian, L.; and Bian, L. 2023. U-Shape Transformer for Underwater Image Enhancement. *IEEE Transactions on Image Processing*, 32: 3066–3079.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Sun, G.; An, Z.; Liu, Y.; Liu, C.; Sakaridis, C.; Fan, D.-P.; and Van Gool, L. 2023. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13791–13801.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Varghese, N.; Kumar, A.; Rajagopalan, A. N.; and Varghese. 2023. Self-supervised Monocular Underwater Depth Recovery, Image Restoration, and a Real-sea Video Dataset. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 12214–12224.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Ding, M.; Gu, X.; Huang, S.; Xu, B.; et al. 2025. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22958–22967.

Whittaker, R. H. 1969. New Concepts of Kingdoms of Organisms: Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science*, 163(3863): 150–160.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xue, X.; Wei, G.; Chen, H.; Zhang, H.; Lin, F.; Shen, C.; and Zhu, X. X. 2024. Reo-vlm: Transforming vlm to meet regression challenges in earth observation. *arXiv preprint arXiv:2412.16583*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding.

Zhang, C.; Liu, L.; Huang, G.; Wen, H.; ZHOU, X.; and Wang, Y. 2024. WebUOT-1M: Advancing Deep Underwater Object Tracking with A Million-Scale Benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhang, C.; Liu, L.; Huang, G.; Zhang, Z.; Wen, H.; Zhou, X.; Ge, S.; and Wang, Y. 2025b. Underwater Camouflaged Object Tracking Meets Vision-Language SAM2. arXiv:2409.16902.

Zhao, Z.; Liu, Y.; Sun, X.; Liu, J.; Yang, X.; and Zhou, C. 2021. Composited FishNet: Fish Detection and Species Recognition From Low-Quality Underwater Videos. *IEEE Transactions on Image Processing*, 30: 4719–4734.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 46595–46623.

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

Zhuang, P.; Wang, Y.; and Qiao, Y. 2018. Wildfish: A large benchmark for fish recognition in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, 1301–1309.