MSC : A Marine Wildlife Video Dataset with Grounded Segmentation and Clip-Level Captioning

Quang-Trung Truong[1] Wong Yuk Kwan[1] Vo Hoang Kim Tuyen Dang [2] Rinaldi Gotama[3]
Duc Thanh Nguyen[4] Sai-Kit Yeung[1]
[1]HKUST [2]HCMUS [3]Indo Ocean Project [4]Deakin University

Abstract

Marine videos present significant challenges for video understanding due to the dynamics of marine objects and the surrounding environment, camera motion, and the complexity of underwater scenes. Existing video captioning datasets, typically focused on generic or human-centric domains, often fail to generalize to the complexities of the marine environment and gain insights about marine life. To address these limitations, we propose a two-stage marine object-oriented video captioning pipeline. We introduce a comprehensive video understanding benchmark that leverages the triplets of video, text, and segmentation masks to facilitate visual grounding and captioning, leading to improved marine video understanding and analysis, and marine video generation. Additionally, we highlight the effectiveness of video splitting in order to detect salient object transitions in scene changes, which significantly enrich the semantics of captioning content.

1. Introduction

The recent advent of Artificial Intelligence (AI) has led to a new research capacity in life and environmental sciences, from on-ground animal biology [9, 34, 36] to marine biology, e.g., coral detection [46], marine visual analysis [25, 44].

The application of AI methods to ocean-related research is challenging due to the requirement of significant domain expertise and engineering work. Biologists have to collect data and manually label the data for specific marine species and tasks, then find and train a suitable model for each task. For real-world problems, off-the-shelf models often struggle to maintain comparative results. For example, visual grounding models, e.g., Grounding DINO [23] and SAM2 [30], are reliant on COCO's predefined classes [20], and limited to user-defined input (e.g., customized text). Due to the presence of hundreds of marine species, these approaches are not applicable to the marine domain.

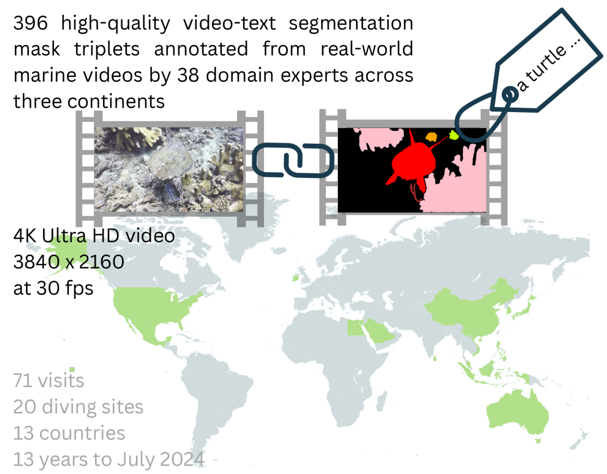

Figure 1. The MSC dataset is recorded from 13 different countries, over 24.8 hours of marine video content. The dataset is associated with fine-grained annotations including clip-level textual descriptions provided by 18 biologists and pixel-level segmentation masks provided by 20 professionals.

While Large Language Models (LLMs) can generate data at scale, they are prone to hallucination. Therefore, even with LLMs, human intervention is still needed to refine the synthetic data generated by the LLMs, especially in data-scarce domains. To address this issue, we propose a new marine video dataset, where the segmentation masks of marine objects are provided by annotators but the textual descriptions of the videos are generated automatically. Note that the MVK dataset introduced in [35] also used LLMs to generate image-level captions. Compared with MVK, our MSC provides more fine-grained and richer textual captions, generated from shorter temporal video segments (called clips) and validated by human experts. In addition, MSC is two times larger than MVK. Video Browser Showdown utilized MVK for known item search in [31, 37], finding MVK a challenging benchmark dataset for visual KIS tasks.

Natural questions arise: *Why do we create a new video captioning dataset? and why do we focus on clip-level*

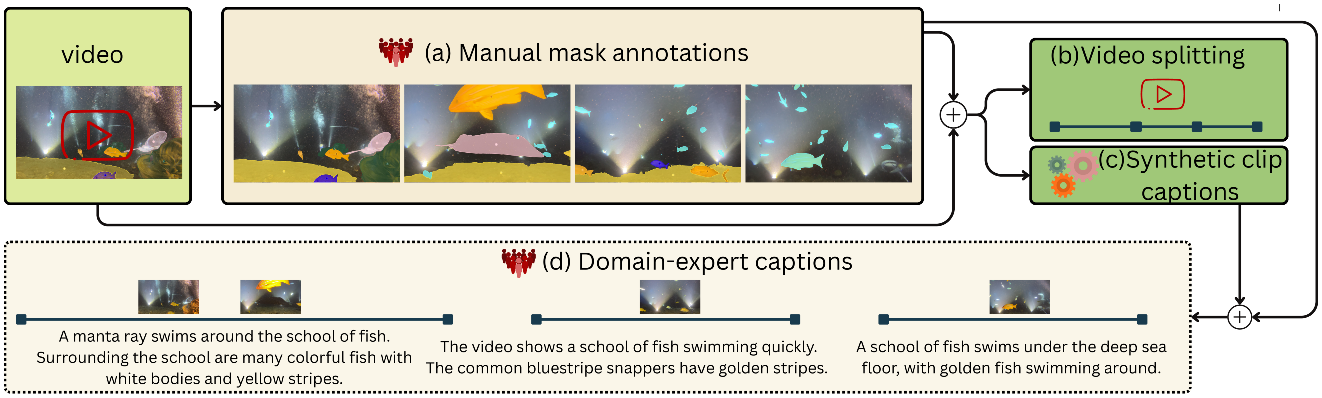

Figure 2. Annotation Generation. The dataset curation consists of two tasks: creating segmentation masks and captioning. (a) We manually label marine videos to identify the target objects. We aim to build a fine-grained marine wildlife video-text dataset, performing the following stages: (b) Extracting clips from marine videos using annotated images. (c) Using a VLM to generate synthetic clip-level captions. (d) Refining the synthetic captions by domain experts. Please zoom-in for the best view.

captioning? We argue that a clip-level video captioning dataset is essential for capturing richer semantic information, where each clip is defined as a semantically coherent segment. This leads to the development of our multimodal dataset comprising video-segmentation mask-text triplets, which explicitly link visual information with detailed textual descriptions and hold potential for multitask learning. Furthermore, we address the issue of hallucination prevalent in SOTA models, e.g., GPT-4.1 [2], suggesting that fine-grained, visually grounded data are crucial for generating accurate and reliable video captions.

In summary, our work makes the following contributions:

- We propose MSC , the first real-world and large-scale marine video dataset, captured in various environments (see Figure 1).
- We provide high-quality video-segmentation mask-text triplets for short-time video clips using a two-stage video annotation pipeline (see Figure 2).
- We provide benchmark results on MSC across applications, including video captioning, plot/clip-level captioning, video generation, and visual grounding.

2. Related Work

Underwater Video Datasets. Marine videos have recently attracted considerable attention from the computer vision community. MarineInst [45] introduced a large-scale marine dataset with instance segmentation masks, facilitating image-level visual analysis tasks. Only 10% of the instance masks of MarineInst are annotated by humans while the remainder is generated by a segmentation model (e.g., SAM [15]). CoralSCOP [46] demonstrates the effectiveness of SAM in coral image segmentation, specifically addressing over-segmentation issues. SAM has been commonly used to create marine image instance segmentation datasets such as Watermask [17] and USIS10K [18].

Video-Text Datasets. Recently, numerous large-scale video-text datasets, e.g., Koala-36M [39], HOIGen-1M [22], MiraData [13], have leveraged automatic captioning systems (e.g., GPT-4V) to generate video captions. MovieBench [42] provides video-, scene-, shot-level captions, also using GPT-4V, by incorporating additional semantic information. ViCaS [6] paves the way for new video-text datasets designed to evaluate video understanding tasks, including video captioning and visual grounding. Several datasets, e.g., BOVText [41], How2 [32], and VALUE [16] are utilized for tasks such as video retrieval, captioning, video text spotting, and video question answering. Furthermore, domain-specific datasets, e.g., BASKET [27], are useful in applications such as classification and video generation. Although existing video-text datasets play an important role for training downstream tasks, e.g., video generation, they rely on LLMs that often produce hallucinated outcomes human interventions.

Language-guided Segmentation Datasets. Early referring video object segmentation (RVOS) datasets, e.g., A2D Sentences [11], J-HMDB Sentences [11], DAVIS16-RVOS [14], DAVIS17-RVOS [14], and Refer-Youtube-VOS [33], focus on single-object segmentation, where language prompts are used to describe a single target object. These datasets often feature expressions, describing static attributes like color and shape. The number of expressions per object varies across the datasets. For instance, A2D Sentences and J-HMDB Sentences typically have 1-2 expressions per object, while DAVIS17-RVOS boasts an average of 7.5 expressions per object, and MeViS [10] provides 3.5 expressions per object. However, it is observed that there is a lack of video grounding datasets. To our

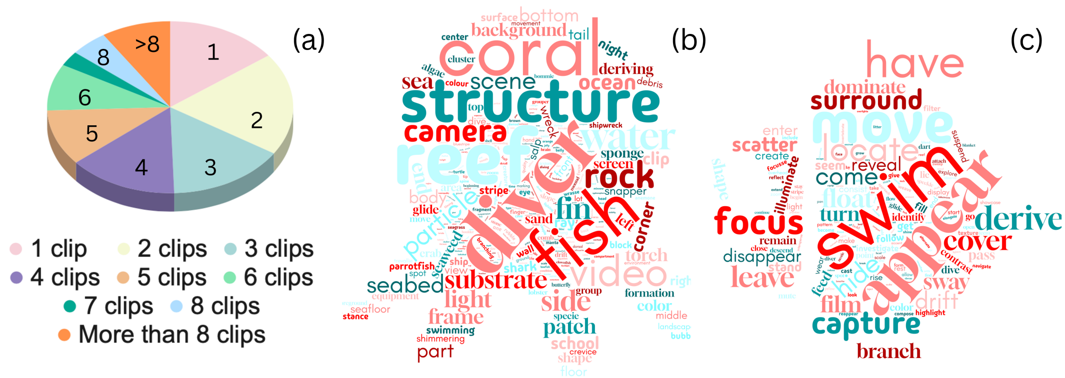

Figure 3. Overview of the MSC dataset. (a) The dataset includes videos and clips, ranging from 1 to 30 clips per video. (b) and (c) Word cloud for noun and verb of marine captions generated by domain experts, respectively.

knowledge, MSC is the first large-scale marine video text dataset for visual grounding. For visual grounding modeling, LLMs have been widely used to improve visual and language comprehension [12, 29]. Grounding DINO [23], built upon Transformers, is designed to generate multiple bounding boxes as prompts for SAM.

3. Our MSC

3.1. Video Filtering

To choose marine videos for annotation, we considered three criteria (validated by human annotators): clarity, complexity, and diversity. (1) *Clarity* means that only videos whose objects can be seen with moderate clarity are selected. (2) *Complexity* means that videos containing objects in three or more distinct object types (e.g., fish, plants, reefs, divers) are selected. (3) *Diversity* means that we choose videos with distinct scenes or objects among those of the same diving site. Applying the above criteria, we finally selected 396 videos out of 2,743 videos.

3.2. Video Annotation

The annotation process involves two steps: creating video segmentation masks and captioning short-time clips. For the first step, annotators utilize a GUI tool to segment marine objects. We then use the manually annotated segmentation masks to identify target objects. In the second step, we provide high-quality clip-level captions using segmentation masks returned in the first step.

Step 1: Instance-level Video Segmentation. We developed a web-based annotation tool for marine video object segmentation in Fig. 4. This annotation tool inherits from SAM [15] and [40] to produce high-quality pixel-wise segmentation masks (that we call pseudo-masks), and then allows annotators to refine the generated masks in an iterative manner. The annotators finally provide a category for each segmented mask. We focus on six categories, including fish, reefs, aquatic plants, wrecks, human divers, and sea floor.

Step 2: Captioning. We leveraged LLMs to generating captions for our collected videos. However, we observed that the generated captions of long-time videos are often superficial because of the lack of detailed descriptions for the events included in the videos. To address this issue, we split long-time videos into short-time clips, each of which captures a single-shot event. We found that this step helps enhance the semantics of the generated captions.

After splitting long-time videos into short-time clips, we used GPT-4.1 [2], Gemini-2.0 Flash-Lite [1], Qwen-VL [7] to generate a textual description for each clip. The generated descriptions were then refined by biologists to reflect the semantic content specified by the segmented objects in the corresponding clips. The descriptions were refined to include the visual attributes and behavior (e.g., feeding, resting, breathing, social interactions, defense) of segmented objects and the background (surrounding environment) in the clips. Finally, the biologists produce a comprehensive caption for each video by aggregating the refined clip-level descriptions for the clips in that video, providing a concise summary of the video content.

3.3. Analysis and Statistics

Diverse Lighting Conditions and Scenarios. Our MSC dataset was constructed by self-recording 71 visits across 20 distinct diving sites from 2011 to July 2024 (see Figure 1). Additionally, MSC is compiled with MVK [35], which encompasses recordings from various regions and under diverse illumination conditions. Data acquisition for this dataset was performed using GoPro cameras. Unlike common video datasets gathered through web crawling, our focus was specifically on marine organisms captured during

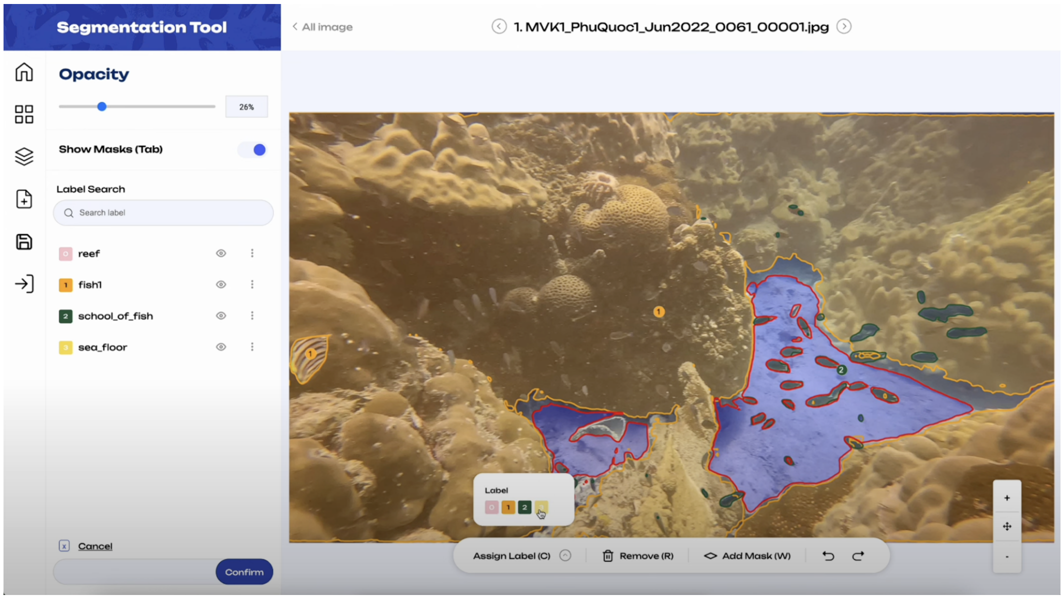

Figure 4. Visualization of our web-based segmentation annotation tool.

Table 1. Statistical overview of representative marine video datasets.

Dataset	Total Duration (hours)	Mean Duration (seconds)	#Video	#Country/Region
MVK [35]	12.3	29.9	1378	11
Our MSC	24.8	32.8	2743	20

dedicated diving expeditions. Importantly, a consistent data acquisition setup was maintained throughout the entire collection process. We show our data's directory structure in Figure 5.

Diverse Descriptions. We observed that textual descriptions obtained from LLMs pre-trained on large-scale datasets often suffer from hallucinations. To address this issue, we integrate the segmentation masks of target objects into textual descriptions, enabling the descriptions to be more focused on the target objects' behaviors and the surrounding background, thereby eliminating hallucinations. This approach allows to create high-quality captions not only for visual grounding tasks but for the Text-to-Video (T2V) task via LLMs video captioning.

To further refine our synthetic captions, we engaged a team of 18 biologists to accurately identify marine organisms in the video footage. Synthetic captions generated by LLMs, i.e., GPT4.1 [2], QWen[7], Gemini 2.0 Flash-Lite [1], are employed to reduce the workload for manual captioning. Each short-time clip in our dataset was annotated with 1 to 4 text descriptions, comprising 3 synthetic and 1 human-written caption. The distribution of annotated videos is illustrated in Figure 3 (a). The word cloud of text descriptions is illustrated in Figure 3 (b).

Challenges with Object Scale and Imbalance. As shown in Table 6 (a), while the numbers of fish and coral reef over 2,000 each, fish are predominantly small objects (0.67% of image area), whereas reefs are typically large 13.03%. Additionally, human divers are less frequent but consistently small objects (1.12% in Table 6 (b)). Conversely, wrecks are rare but large objects (9.89% in Table 6 (a, b)). This highlights an imbalance in both object quantity and scale in MSC dataset.

4. Challenges

4.1. Video-level Captioning

Video captioning aims at generating a descriptive text for an input video sequence.

Baselines: In this work, we evaluated several prevailing visual-language models (VLMs) for video-level captioning in the MSC dataset. These models include Qwen-VL-Chat [7], LLaVA [21], PLLaVA [43], Gemini [1], MovieBench [42]. The models were used to generate captions at the video level for our marine video collection.

Qwen-VL uses Qwen as the language backbone and OpenCLIP ViT-bigG as the visual encoder, connected via a single-layer cross-attention module as the position-aware vision-language adapter. LLaVA employs a large language model, such as Vicuna, LLaMA, combined with a frozen vision encoder, such as CLIP ViT-L/14) via a single linear layer as a trainable projector. PLLaVA, an extension of LLaVA for videos, uses a parameter-free pooling strategy for video captioning. MovieBench is a data pipeline designed to create shot-level descriptions of scenes, with a particular emphasis on movie characters.

We set up to evaluate these VLMs with the guide prompt *"Describe the video by following guidelines: you should*

Table 2. Video-level captioning on MSC dataset.

Method	Year	BLEU↑	METEOR↑	ROUGE-L↑	CIDEr↑	SPICE↑
Qwen-VL-Chat 7B [7]	2023	0.0000	0.0759	0.1384	0.0148	0.0848
LLaVA 7B [21]	2024	**0.0125**	0.1241	0.1798	0.0000	0.0689
PLLaVA 7B [43]	2024	0.0000	0.1114	0.1443	0.0028	0.0533
Gemini-2.0 Flash-Lite [1]	2025	0.0000	**0.1251**	**0.1829**	0.0571	0.0870
MovieBench [42]	2025	0.0000	0.1213	0.1790	**0.0898**	**0.0914**

Table 3. Clip-level captioning on MSC dataset.

Method	Year	BLEU↑	METEOR↑	ROUGE-L↑	CIDEr↑	SPICE↑
Qwen-VL-Chat 7B [7]	2023	0.0000	0.0832	0.1541	0.0046	0.0696
LLaVA 7B [21]	2023	0.0000	0.1337	0.1678	0.0057	0.0638
Gemini-2.0 Flash-Lite [1]	2025	0.0931	0.1736	0.2890	0.3679	0.1647
GPT4.1 [2]	2025	**0.7196**	**0.5186**	**0.7844**	**4.9314**	**0.6139**

```
MSC /
├── MVK1
└── MVK2/
    ├── videos/
    |   ├── <location_time #1>/
    |   |   ├── <video #1.mp4>
    |   |   ├── <video #2.mp4>
    |   |   └── ...
    |   └── ...
    └── information/
        ├── metadata
        ├── highres_JPEGImages_1fps/
        |   ├── <location_time #1>_<video #1>/
        |   |   ├── 00001.jpg
        |   |   ├── 00002.jpg
        |   |   └── ...
        |   └── ...
        ├── JPEGImages_1fps/
        |   ├── <location_time #1>_<video #1>/
        |   |   ├── 00001.jpg
        |   |   ├── 00002.jpg
        |   |   └── ...
        |   └── ...
        ├── annotations/
        |   ├── masks/
        |   |   ├── <location_time #1>_<video #1>/
        |   |   |   ├── 00001.png
        |   |   |   ├── 00002.png
        |   |   |   └── ...
        |   |   └── ...
        |   ├── mask_vis/
        |   |   ├── MVK2_<location_time #1>_<video #1 00001.png>
        |   |   ├── MVK2_<location_time #1>_<video #1 00002.png>
        |   |   └── ...
        |   └── clip_descriptions/
        |       └── MVK2.0_metadata.csv
        └── splits/
            ├── train.txt
            ├── val.txt
            └── test.txt
```

Figure 5. The MSC 's data organization.

Table 4. Visual Grounding performance on MSC dataset.

Method	Year	Input	mIoU↑	Recall↑
GroundingDINO + SAM2	2024	Labels	0.6543	0.7135
GroundingDINO + SAM2	2024	Caption	0.4447	0.5244
GLaMM + SAM2	2024	Caption	0.6727	0.7255
VideoGLaMM	2025	Caption	**0.6812**	**0.7532**

give a paragraph with maximum 75 words; focus on the most obvious feature of the main objects in the initial frames; infer the behavior of the object (feeding, resting, breathing, social interactions, defense); and describe the background in about 10 words. Focus on fish, reefs, aquatic plants, wrecks, human divers, and sea floor. Omit the words 'underwater' and 'shows'."

We utilized the lightweight versions (7B parameters) of the models to fit with our hardware specification NVIDIA RTX-3090 GPUs. Additionally, we amended MovieBench to generate detailed scene descriptions. Specifically, we input a list of keyframes from a video sequence, along with the images of target objects, and use GPT-4.1 to describe the observable features and behaviors of the target objects in no more than 75 words. The images of the target objects are originated from the segmentation masks in our dataset. To optimize the image token usage, each frame is resized to 540×960. It is important to note that we omit the audio input required in the original MovieBench framework, as it is unavailable in our setup. We ultilised the captioning metrics, i.e., BLEU [28], METEOR [8], ROUGE-L [19], CIDEr [38], and SPICE [5].

Analysis: Gemini-2.0 and MovieBench perform well across the benchmarking metrics. Specifically, Gemini achieves 0.1251 (METEOR) and 0.1829 (ROUGE), while MovieBench achieves 0.0898 (CIDEr) and 0.0914 (SPICE). MovieBench is based on GPT-4.1, we observe that commer-

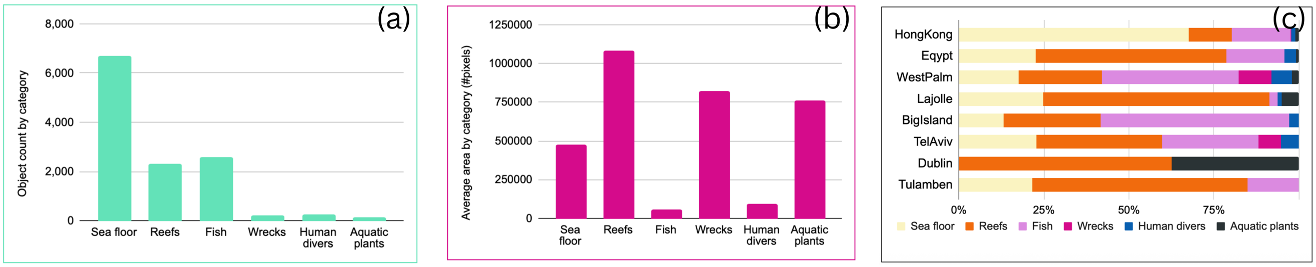

Figure 6. MSC Dataset. (a) The number of instances per category. (b) The average area of category. (c) Distribution of the number of instances by regions in the MSC dataset.

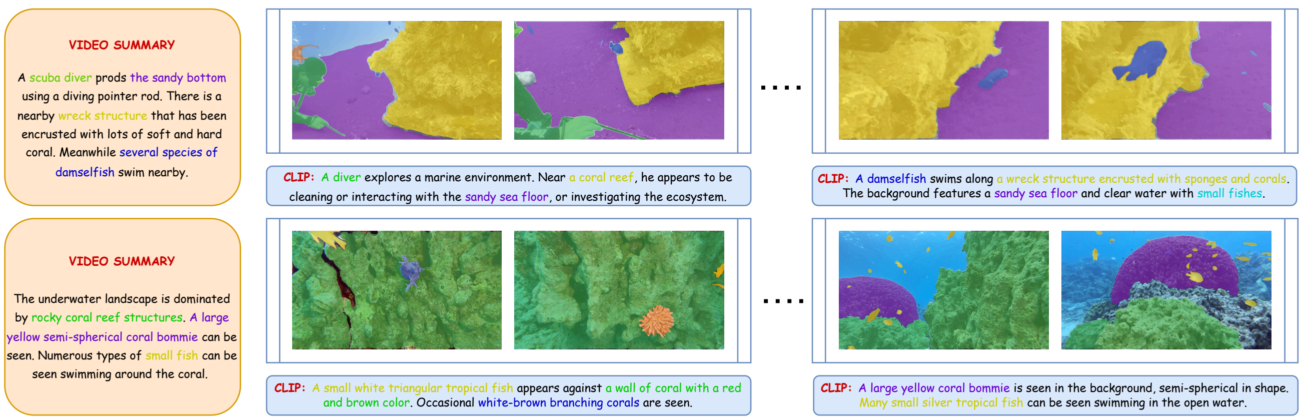

Figure 7. Visualization of video-text segmentation mask triplets.

cial models often produce comparative results.

4.2. Clip-level Captioning

Clip-level captioning aims to split a video into clips and then provide a caption for each clip.

Baselines: We also leverage various open-source and commercial VLMs like in the above video level captioning on our MSC dataset. With the similar setup to video captioning, we used the same prompt, but we input the frames from each clip rather than an entire video. We set the number of frames per clip to 10. To evaluate the clip-level captioning, we adopted the same metrics as in the Sec. 4.1.

Analysis: Regarding clip-level captioning benchmark, GPT-4.1 demonstrates superior performance, achieving the highest scores across all benchmark metrics. Gemini-2.0 ranks as the second-best performer in this comparison. Specifically, GPT-4.1 achieved scores of 0.7196 (BLEU), 0.5186 (METEOR), 0.7844 (ROUGE-L), 4.9314 (CIDEr), and 0.6139 (SPICE), while Gemini 2.0 yielded respective scores of 0.0931 (BLEU), 0.1736 (METEOR), 0.2890 (ROUGE-L), 0.3679 (CIDEr), and 0.1647 (SPICE). Note, we observe commercial models consistently achieve competitive performance in video and clip level captioning on this challenging domain in Table 3.

4.3. Visual Grounding

The visual grounding task in the MSC dataset involves linking a target model's response to a user-specific text query by identifying marine creatures, objects, and their behaviors. This task requires the model to understand both spatial and temporal aspects in visually complex marine scenes characterized by high object variability, occlusion, and challenging lighting.

Given a caption $\{CAPTION\}$ describing a clip, we prompt a target model using the format "$\{CAPTION\}$. *Please respond with segmentation masks*" to extract segmentation maps aligned with the textual description. The expected output highlights relevant spatial regions across frames, demonstrating the model's ability to understand both spatial details (what appears in each frame) and temporal dynamics (how things change over time). We evaluated the grounding quality using mIoU and Recall metrics across annotated frames.

Baselines: We first assessed the performance of an open-vocabulary model, GroundingDINO [23] combined with SAM2 [30]. For further validation, we benchmarked recent LLM-based visual grounding models, including VideoGLaMM [26] and GLaMM [29]. As GLaMM [29] is originally designed for image-based grounding, we ex-

Table 5. Evaluation results with open-source model and commercial models in video generation.

Method	Venue/Year	CLIP-T↑	Temp Consistency↑	FID↓	FVD↓
Latte[24]	2025	0.3189	0.993	76.91	3123.05
Hailuo [3]	-	**0.3236**	0.9934	83.68	**2007.37**
Kling 1.5 [4]	2024	0.3148	**0.997**	**71.90**	2820.24

tended it to the video domain by incorporating temporal capabilities through SAM2 [30].

Analysis: While GroundingDINO + SAM2 show promise, their reliance on COCO-style categories limits their generalization ability to the marine domain. Performance notably degrades with natural, unconstrained captions and only improves when explicit label names are extracted for prompting. In contrast, LLM-based models better support spatio-temporal reasoning, establishing a more robust foundation for grounding in complex marine environments.

4.4. Text-to-video Generation

Our goal is to benchmark T2V models in generation of marine videos using clip-level captions from a video-text dataset. Pre-trained models were used in this experiment.

Baselines: We compared the open-source model Latte[24] and commercial models, i.e., Hailuo [3], Kling 1.5 [4], in marine video generation task. For the experimental setup, we used 50 clip-level caption prompts as input for the T2V models. **Analysis:** We observe commercial models often perform well on video generation across T2V metrics in Table 5. Specifically, Hailuo gains the superior result with 0.3236 (CLIP-T) while Kling 1.5 gains 0.997 on temperal consistancy metric. Pre-trained T2V models exhibit suboptimal performance in FID and FVD metrics due to insufficient diversity of underwater visual content in existing training datasets.

5. Conclusion

This paper introduces MSC , the first large-scale video dataset of marine wildlife. The dataset contains fine-grained annotations, including object segmentation masks, clip-level captions, and video summaries. We developed an effective two-stage data annotation pipeline minimizing hallucinations by LLMs. Our associated benchmark includes video captioning, visual grounding, and text-to-video generation. We believe that MSC will contribute to facilitate research in marine video understanding.

References

[1] Gemini. https://gemini.google.com/. 3, 4, 5

[2] GPT-4.1. https://openai.com/index/gpt-4-1/. 2, 3, 4, 5

[3] Hailuo. https://hailuoai.video. 7

[4] Kling 1.5. https://klingai.com/. 7

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 382–398. 5

[6] Ali Athar, Xueqing Deng, and Liang-Chieh Chen. 2024. ViCaS: A Dataset for Combining Holistic and Pixel-level Video Understanding using Captions with Grounded Segmentation. *arXiv preprint arXiv:2412.09754* (2024). 2

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023). 3, 4, 5

[8] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72. 5

[9] Otto Brookes, Maksim Kukushkin, Majid Mirmehdi, Colleen Stephens, Paula Dieguez, Thurston C Hicks, Sorrel Jones, Kevin Lee, Maureen S McCarthy, Amelia Meier, et al. 2025. The PanAf-FGBG Dataset: Understanding the Impact of Backgrounds in Wildlife Behaviour Recognition. *arXiv preprint arXiv:2502.21201* (2025). 1

[10] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. 2023. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2694–2703. 2

[11] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5958–5966. 2

[12] Shuting He and Henghui Ding. 2024. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13332–13341. 3

[13] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems* 37 (2024), 48955–48970. 2

[14] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. 2019. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*. Springer, 123–141. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026. 2, 3

[16] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. 2021. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632* (2021). 2

[17] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. 2023. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1305–1315. 2

[18] Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Laurence Tianruo Yang, Sam Kwong, and Runmin Cong. 2024. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In *Proceedings of the 41st International Conference on Machine Learning*. 29545–29559. 2

[19] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81. 5

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the the European Conference on Computer Vision*. 740–755. 1

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306. 4, 5

[22] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. 2025. HOIGen-1M: A Large-scale Dataset for Human-Object Interaction Video Generation. *arXiv preprint arXiv:2503.23715* (2025). 2

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *arXiv preprint arXiv:2303.05499*. 38–55. 1, 3, 6

[24] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2025. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research* (2025). 7

[25] Kazi Sajeed Mehrab, M Maruf, Arka Daw, Abhilash Neog, Harish Babu Manogaran, Mridul Khurana, Zhenyang Feng, Bahadir Altintas, Yasin Bakis, Elizabeth G Campolongo, et al. 2024. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images. *arXiv preprint arXiv:2407.08027* (2024). 1

[26] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. 2024. VideoGLaMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos. *arXiv preprint arXiv:2411.04923* (2024). 6

[27] Yulu Pan, Ce Zhang, and Gedas Bertasius. 2025. BASKET: A Large-Scale Video Dataset for Fine-Grained Skill Estimation. *arXiv preprint arXiv:2503.20781* (2025). 2

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318. 5

[29] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13009–13018. 3, 6

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024). 1, 6, 7

[31] Luca Rossetto, Klaus Schoeffmann, Cathal Gurrin, Jakub Lokoč, and Werner Bailer. 2024. Results of the 2024 Video Browser Showdown. *arXiv preprint arXiv:2502.15683* (2024). 1

[32] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347* (2018). 2

[33] Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Com-*

puter Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer, 208–223. 2

[34] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. 2024. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19412–19424. 1

[35] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue-Him Wong, Ajay Joneja, and Sai-Kit Yeung. 2023. Marine Video Kit: A new marine video dataset for content-based analysis and retrieval. In *International Conference on Multimedia Modeling*. Springer, 539–550. 1, 3, 4

[36] Devis Tuia. 2025. MammAlps: A multi-view video behavior monitoring dataset of wild mammals in the Swiss Alps. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025*. 1

[37] Lucia Vadicamo, Rahel Arnold, Werner Bailer, Fabio Carrara, Cathal Gurrin, Nico Hezel, Xinghan Li, Jakub Lokoc, Sebastian Lubos, Zhixin Ma, et al. 2024. Evaluating performance and trends in interactive video retrieval: Insights from the 12th vbs competition. *IEEE Access* (2024). 1

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575. 5

[39] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. 2024. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260* (2024). 2

[40] Yuk-Kwan Wong, Ziqiang Zheng, Mingzhe Zhang, David Suggett, and Sai-Kit Yeung. 2024. CoralSCOP-LAT: Labeling and Analyzing Tool for Coral Reef Images with Dense Mask. arXiv:2410.20436 [cs.CV] https://arxiv.org/abs/2410.20436 3

[41] Weijia Wu, Yuanqiang Cai, Debing Zhang, Sibo Wang, Zhuang Li, Jiahong Li, Yejun Tang, and Hong Zhou. 2021. A bilingual, openworld video text dataset and end-to-end video text spotter with transformer. *arXiv preprint arXiv:2112.04888* (2021). 2

[42] Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. 2024. MovieBench: A Hierarchical Movie Level Dataset for Long Video Generation. *arXiv preprint arXiv:2411.15262* (2024). 2, 4, 5

[43] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994* (2024). 4, 5

[44] Ziqiang Zheng, Yiwei Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. 2024. Marineinst: A foundation model for marine image analysis with instance visual description. In *European Conference on Computer Vision*. Springer, 239–257. 1

[45] Ziqiang Zheng, Yiwe Chen, Huimin Zeng, Tuan-Anh Vu, Binh-Son Hua, and Sai-Kit Yeung. 2024. Marine-Inst: A Foundation Model for Marine Image Analysis with Instance Visual Description. In *Proceedings of the European Conference on Computer Vision*. Springer. 2

[46] Ziqiang Zheng, Haixin Liang, Binh-Son Hua, Yue Him Wong, Put Ang, Apple Pui Yi Chui, and Sai-Kit Yeung. 2024. CoralSCOP: segment any coral image on this planet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28170–28180. 1, 2