

MarineEVT: Advancing Event-Centric Marine Video Understanding via Visual Tool Reasoning

Tuan-An To¹, Yuk-Kwan Wong¹, Tuan-Anh Vu²,
Ziqiang Zheng^{†1,3}, and Sai-Kit Yeung¹

Project website: <https://marineevt.hkustvgd.com>; † : zhengziqiang1@gmail.com

¹ The Hong Kong University of Science and Technology, Hong Kong, China

² University of California, Los Angeles CA, USA

³ University of Electronic Science and Technology of China, China

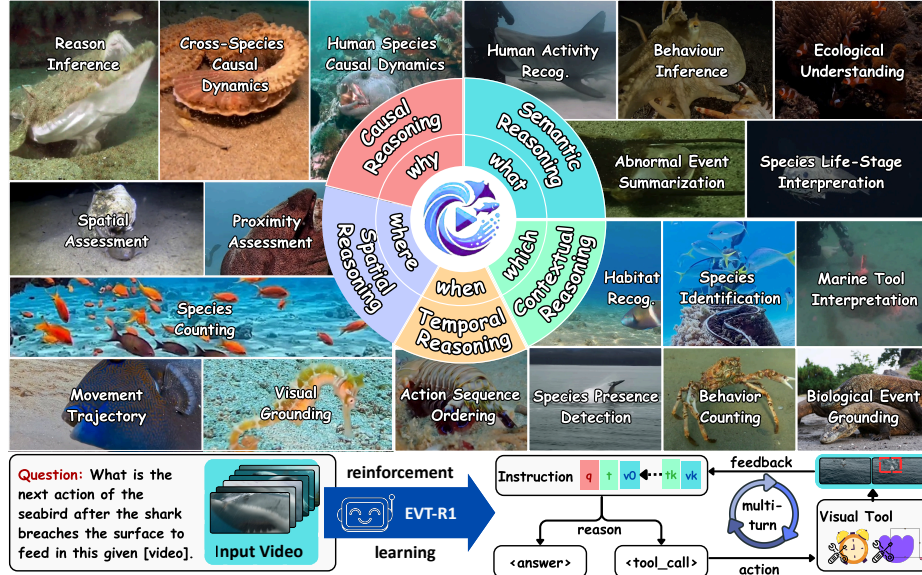


Fig. 1: We propose MarineEVT, the first hierarchical and comprehensive event-centric marine video dataset. Based on MarineEVT, we propose EVT-R1, integrating visual tool reasoning into the VLM for more reliable marine video understanding.

Abstract. Recent Vision-Language Models (VLMs) have achieved remarkable success in visual understanding, driven by the growing availability of high-quality image-text pairs. However, the performance of VLMs often degrades in the video domain due to the essential need for temporal understanding and the scarcity of large-scale annotated video data. In this work, we focus on marine video understanding, which brings further challenges: first, it requires substantial domain expertise; and video VLMs usually struggle with localizing and interpreting critical information from marine videos, as the informative events are typically *sparse*, *unpredictable*, and *unevenly distributed*. To address these challenges, we carefully curate the first event-centric marine video understanding dataset called **MarineEVT**, which features 20K multi-task,

video-level visual question-answering pairs spanning multiple dimensions of marine understanding and analysis. Meanwhile, based on MarineEVT, we decompose marine video understanding as an Event-centric Visual Tool-integrated Reasoning process (**EVT-R1** for short), where we leverage powerful visual tools to drive the model to localize and interpret critical information aligned with visual questions and human intent. To demonstrate its effectiveness, we compare EVT-R1 against 11 SOTA VLMs in different settings. EVT-R1 outperforms the top open-source and top commercial models by 5.22 and 11.09, respectively. MarineEVT and EVT-R1 lay the foundation for ecological discovery and marine education, fostering the development of VLMs capable of interpreting marine dynamics, reasoning about ecological interactions, and supporting sustainable ocean video understanding and analysis.

Keywords: Marine Video Understanding · Event Understanding · VLM · Visual Tool-Integrated Reasoning · Video Dataset and Benchmark

1 Introduction

Marine understanding stands as a pivotal frontier in biological and environmental science, shaping our ability to study and preserve the vast, complex ecosystems that cover over 70% of Earth yet remain little explored. Uncovering its secrets [74, 82–84] is crucial for both advancing scientific understanding and tackling global challenges such as biodiversity loss [28], climate change [90], and sustainable ocean management [66]. In recent years, this domain has experienced remarkable advancements, particularly in single-image analysis [71, 88], where state-of-the-art methods have demonstrated outstanding performance across a broad spectrum of visual tasks, including marine object detection [18, 65, 67], instance segmentation [39], instance-level captioning [88], and related applications [26, 71].

However, despite these remarkable advances in image-level analysis, progress in comprehensive marine video understanding remains considerably constrained [71, 74]. Unlike image-level visual understanding [83, 89], video understanding is inherently more challenging, as it demands **event-level interpretation**. The complex intrinsic of marine exploration challenges effective marine video understanding: 1) marine videos are often **long** and **tedious**, making it difficult to localize and interpret meaningful events; 2) comprehending the marine videos (especially videos with ecological traits) requires **significant domain expertise**; 3) marine video understanding tailored to ecological and educational purposes sets it apart from general-purpose applications.

In marine videos, informative events are often *sparse*, *ephemeral*, and *unevenly distributed*, posing significant challenges for existing VLMs [8, 9, 58] to effectively localize and interpret these sparse but crucial events. These episodic observations carry immense scientific value [16, 22], offering critical insights for ecological dynamics [21, 47], species behavior [29, 31], biodiversity evolution [28], and the impacts [44, 90] of human activities or climate change. However, current

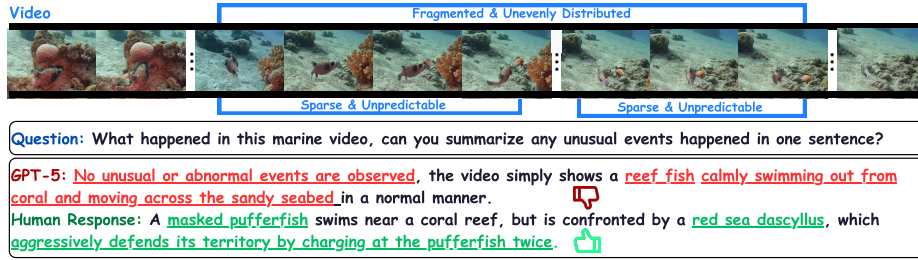


Fig. 2: An example question for evaluating event summarization tested on GPT-5.0 [53] with human response provided for comparison.

general-purpose VLMs primarily emphasize scene summarization, overlooking the fine-grained visual dynamics that better align with domain requirements. Thus, marine video understanding demands a significant shift toward precise, event-centric descriptions that capture specific entities (*e.g.*, marine organisms, divers, instruments) and their subtle and dynamic interactions. This gap underscores the need for specialized, domain-adaptive video VLMs capable of capturing informative visual dynamics for event-centric marine video understanding.

Motivated by such demand, we carefully construct an event-centric dataset and benchmark, **MarineEVT**, as shown in Fig. 1, to drive VLMs toward deeper temporal reasoning and domain-aware understanding of dynamic marine events. MarineEVT comprises 20,000 richly annotated underwater video question-answer pairs spanning 20 fine-grained dimensions, including *marine species*, *human activities*, *environmental conditions*, *behavioral interactions*, and *rare ecological events*, structured to support **semantic**, **contextualized**, **spatial-temporal**, and **causal** reasoning. During the construction of MarineEVT, we integrate marine domain expertise into *prompt design* and *annotation verification*, ensuring ecologically precise, context-aligned supervision that helps VLMs interpret biologically meaningful cues and minimize ambiguity. Lastly, unlike existing marine datasets [68, 74, 88, 89] that focus on global scene description, our dataset centers on localizing and understanding meaningful events.

Although MarineEVT fulfills the need for event-centric marine video datasets, we observe that general-purpose VLMs struggle with limited domain expertise and an inability to localize or retrieve critical information aligned with visual questions and human intent, as shown in Fig. 2. The advanced models like GPT-5 [53] fail to accurately summarize the happening event, a reef fish defending territory, highlighting the ecological significance. The abundance of redundant visual inputs across video, caused by slow or static underwater motion, obscures the boundaries of meaningful events and weakens the interest of frames that favor the domain requirements. These weaknesses particularly hinder the model’s ability to ground language to sparse, fragmented events, such as predator-prey interactions or rare species appearances, which are ecologically significant yet easily lost in the episodic videos. Even fine-tuning general-purpose VLMs on MarineEVT can enrich their domain knowledge, the fine-tuned VLMs will still struggle with robust temporal reasoning and event-level interpretation, since

there is no specific design to localize the sparse but critical events from marine videos with redundant frames.

In this work, we propose decomposing complex event-centric marine video understanding into **consecutive reasoning turns**. Our motivation is aligned with how humans solve complicated tasks (especially for the tasks that require deep expertise): they conduct tasks in multiple turns, in each turn they decide whether to use external tools for assistance and the intermediate outputs produced serve as extra information for the final outputs. Specifically, we leverage powerful visual tools to retrieve and localize critical information from redundant visual inputs. Meanwhile, the tool invocation also yields extra visual cues/guidance for condensing visual representations and driving the VLM towards more reliable answer generation. The proposed Event-centric Visual Tool-integrated Reasoning framework (**EVT-R1** for short) enables the VLM to focus on critical events and relevant entities, discarding irrelevant visual signals to enhance temporal grounding and dynamic context understanding.

Equipped with EVT-R1, we achieve better performance gains (+5.22) than supervised fine-tuning (+1.6), and also reinforcement learning (-8.09) for post-training only. Furthermore, we benchmark 11 SOTA models across a comprehensive suite of tasks, including multi-task spatial/temporal grounding, video question answering, and video summarization, under different settings. The experimental analysis provides insights regarding developing domain-adaptive VLMs and sparks a new direction for integrating visual tools for advancing event-centric marine video-language understanding. The main contributions of this paper are summarized as follows:

- We curate **MarineEVT**, a dataset of 20K video question-answer pairs spanning 20 distinct dimensions and reasoning tasks. To the best of our knowledge, this is the first dataset and benchmark for event-centric marine video understanding.
- We propose decomposing event-centric marine video understanding into a multi-turn **visual tool-integrated reasoning** process, leveraging powerful visual tools to localize and interpret critical information from redundant video frames with sparse and unevenly distributed events.
- We propose **EVT-R1**, a training paradigm that equips VLMs with multi-turn, event-grounded reasoning capabilities to better process complex marine video dynamics. It devises dual rewards to optimize tool use and answer generation, yielding more robust and interpretable domain-specific understanding.

2 Related Work

2.1 Event Understanding

Event understanding [30, 49, 50] involves comprehending how visual scenes evolve over time by interpreting interactions, transitions, and causal relationships between entities within a dynamic context [19, 56]. It extends beyond static visual perception to encompass the recognition [52], identification [11], and causal reasoning [79] of actions or scene changes over time. Through event understanding,

we can integrate spatial, temporal, and contextual information to form a coherent understanding of what is happening, why it occurs, and what may happen next. To achieve this, recent multimodal frameworks propose to incorporate linguistic cues to promote interpretability, such as VideoCLIP [72], MERLOT [80], InternVL series [14, 15, 60, 91], and QwenVL series [7–9, 58], which integrate visual dynamics with textual guidance to model temporal relationships and narrative semantics. Despite their effectiveness, these models heavily rely on large-scale pretraining corpora [61, 62] and often overlook fine-grained, domain-specific events (*e.g.*, surgical actions [33], scientific procedures [86], or ocean-monitoring [51] workflows), where data distributions and semantics differ substantially from those of internet-scale datasets. In detail, our daily activity videos [11, 17, 25] are typically structured, predictable, and even summarizable as a series of steps or procedures. In contrast, some domain-specific videos, such as surgical procedures [37, 81] or underwater exploration [74] recordings, exhibit complex and less predictable dynamic events.

To capture these dynamics, recent VLMs have incorporated event-centric vision encoders that enhance spatial-temporal representation learning. EventGPT [42] and EventVL [35] explicitly encode event structure into the visual backbone to improve temporal coherence. Reinforcement learning [46] provides a pathway toward interpretable, event-aware modeling. Recent advances [23, 48, 78, 87] enable LLMs to perform action-level reasoning under policy and reward constraints, fostering deliberate, stepwise inference. Specifically, VTool-R1 [69] employs reinforcement learning to fine-tune VLMs to explore flexible reasoning trajectories and learn to use visual editing tools effectively. Yet this paradigm remains underexplored in marine video understanding, where agents must recognize salient moments, infer causal interactions, and reason about underwater events as they explore.

2.2 Marine Understanding

Marine datasets and benchmarks have progressed rapidly, comprehensively addressing tasks such as instance segmentation [39, 45], object detection [26, 45, 67], and object tracking [6, 82]. However, perceptually complex tasks requiring semantic reasoning, contextual understanding, and domain-specific knowledge in vision-language understanding remain underexplored. MarineGPT [89] and MarineInst [88] introduce multimodal benchmarks that combine visual and linguistic understanding of marine imagery. More recently, CoralVQA [24] introduced a large-scale dataset specialized for coral reef understanding. MarineEval [68] introduced a multi-task framework evaluating marine intelligence across VQA, summarization, grounding, and completion for VLMs. These benchmarks remain constrained to static image-level assessment, overlooking the rich temporal dynamics. Because marine observations unfold through temporally and causally linked events, their interpretation demands domain expertise beyond static-image analysis; event-centric understanding is crucial for capturing ecological dynamics. Although UVLM [74] initiated marine video analysis, it remains

Table 1: Comparisons with existing marine (VLM & non-VLM) datasets.

Ⓐ: Attribute, Ⓑ: Behaviour, Ⓔ: Species, Ⓗ: Human, Ⓔ: Environment,
 Ⓔ: Static, Ⓓ: Dynamic, Ⓓ: Temporal, Ⓓ: Reason, Ⓓ: Outcome.

Dataset	V. Modality	Q. Format	Dataset Task	#Data	#Dimension	Tool	Semantic.	Contextual.	Spatial.	Temporal.	Causal.
UIIS [39]	Image	masks	Segmentation	5K	—	✗	—	Ⓔ	Ⓔ	—	—
FishNet [32]	Image	labels/bounding-boxes	Recog./Detect.	95K	—	✗	—	Ⓔ	Ⓔ	—	—
Ocean20K [34]	Image	masks	Segmentation	20K	—	✗	—	Ⓔ	Ⓔ	—	—
MarineInst [88]	Image	masks/open-end	Seg./Caption.	20M	—	✗	Ⓐ	Ⓔ	Ⓔ	—	—
NAUTILUS [73]	Image	masks/open-end	Seg./Caption.	1.5M	—	✗	Ⓐ	Ⓔ	Ⓔ	—	—
UTB180 [6]	Video	masks/bounding-boxes	Tracking	180	—	✗	—	Ⓔ	Ⓓ	—	—
WebUOT [82]	Video	masks/bounding-boxes	Tracking	1M	—	✗	—	Ⓔ	Ⓓ	—	—
MarineGPT [89]	Image	open-end	Species Und.	5M	—	✗	Ⓐ	—	—	—	—
MarineEval [68]	Image	multiple-format	General Und.	2K	20	✗	Ⓐ, Ⓑ	Ⓔ, Ⓔ	Ⓔ	—	—
CoralVQA [24]	Image	exact-match	Coral Und.	177K	16	✗	Ⓐ	Ⓔ	Ⓔ	—	—
UVBench [83]	Image	multiple-format	General Und.	125K	—	✗	Ⓐ	Ⓔ, Ⓔ	Ⓔ	—	—
UVLM [74]	Video	open-end	General Und.	2K	9	✗	Ⓐ	Ⓔ, Ⓔ	—	—	—
Ours	Video	multiple-format	Event Und.	20K	20	✓	Ⓐ, Ⓑ	Ⓔ, Ⓔ, Ⓔ	Ⓔ, Ⓓ	Ⓓ	Ⓓ, Ⓓ

limited to static attributes, such as species labels or isolated actions, overlooking the temporal and causal dynamics central to event-level video understanding explored in our work.

3 Methodology

3.1 MarineEVT Construction

Clarification with Existing Datasets and Benchmarks Marine videos are visually redundant and unevenly informative, making it hard to localize rare yet significant events and complex behaviors. To bridge this gap, we introduce MarineEVT, an event-centric dataset specifically designed to advance temporal and causal understanding of marine videos. It provides fine-grained, temporally grounded tasks to assess VLM reasoning in dynamic marine environments, going beyond *object-* or *image-* level benchmarks to enable deeper temporal and ecological understanding. In Table 1, we compare MarineEVT with existing marine datasets and benchmarks. MarineEVT precisely targets event-centric reasoning, encompassing *what*, *which*, *where*, *when*, and *why* of marine events. Meanwhile, it anchors its evaluation in four distinct question types grounded in domain knowledge, thereby assessing models’ ability to integrate semantic reasoning, temporal causality, and ecological understanding for autonomous marine observation, environmental monitoring, and scientific discovery.

MarineEVT Construction Pipeline. We propose a scalable pipeline with hierarchical multi-level verification for reliable data construction as shown in Fig. 3. We collect 7,300 marine videos from sources, including MBARI [3], Discovery Channel [1], National Geographic [4] from YouTube [5], and Instagram [2]. Considering video QA tasks introduce temporal complexities, we adopt a two-stage approach: *scene-level* pseudo-captions for context and *image-level* annotations for fine-grained grounding. During construction, we incorporate marine domain expertise into both **prompt design** and **data annotation verification**, ensuring that VLMs are guided by precise ecological terminology and contextually

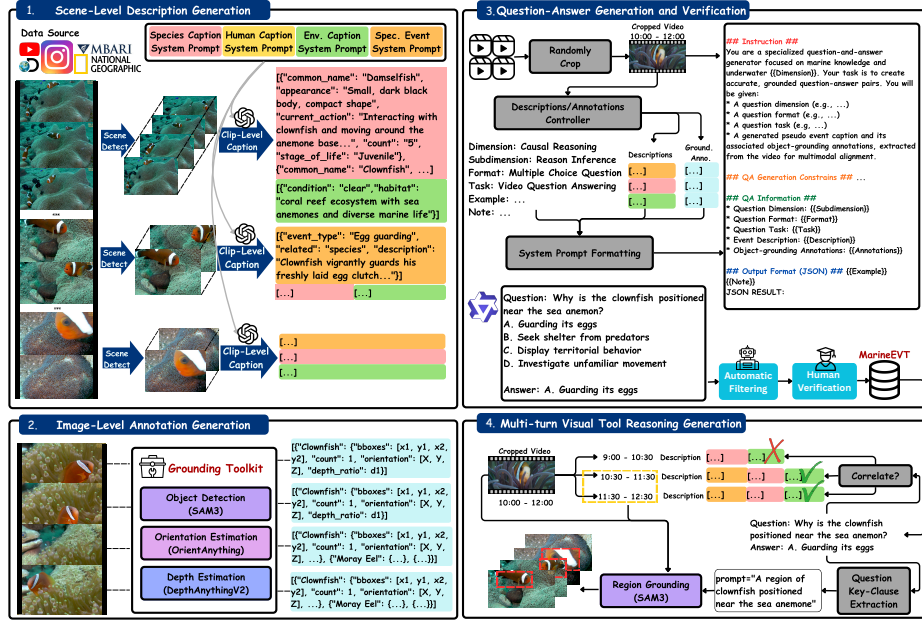


Fig. 3: We propose a data construction pipeline that systematically transforms publicly available marine videos into high-quality, human-verified annotations, ensuring reliability and versatility for diverse VLM training tasks.

aligned instructions. This expert integration enables the model to better interpret biologically meaningful cues and reduce ambiguity:

Two-stage data generation from coarse-descriptions to fine-annotation. Directly prompting an LLM to generate question-answer pairs from raw marine video is inevitably unreliable due to the knowledge gap, and the critical information is sparse and unevenly distributed in marine videos. To mitigate these, we adopt a two-stage coarse-to-fine framework. First, to address the intrinsic sparsity of marine events, we employ TransNet [54] to segment **7,300** videos into **97,284** scenes. For each scene, we use **domain-specialized prompts**, covering species, humans, environments, and notable events to guide GPT-5 [53] in generating coarse-grained, domain-specific scene-level descriptions. Second, for each image in the scene, we apply grounding toolkits to extract image-level grounding annotations for each entity driven from the descriptions. Toolkits include SAM3 [12] for detection, DepthAnythingV2 [76] for depth estimation, and OrientAnything [63] for orientation (totally **94,028** object-grounding annotations consist of **240,017** bounding boxes), enriching descriptions with spatial and geometric grounding. These intermediate descriptions and annotations establish a semantic foundation that enables the LLM to subsequently produce accurate, fine-grained question-answer pairs with improved grounding and consistency.

Question-answer generation with visual tool reasoning steps generation under rigorous verification. Following the previous generation stage, each scene

contains up to four domain-specific descriptions and N grounding annotations (where N is the number of frames). To augment data diversity, we randomly crop the original video into sequences containing k sub-scenes, yielding up to $4 \times k$ descriptions and $N \times k$ annotations. Combined with metadata (*e.g.*, question task (yes/no, MCQ, open-end, exact-match), reasoning dimension), these form structured prompts for QA generation via QwenVL-Max [8]. All generated pairs undergo **rigorous two-tier verification**: automated filtering by three LLMs, followed by validation from three expert human annotators, before final ingestion into MarineEVT. In addition, we **synthesize intermediate reasoning steps** to enable VLMs to localize critical information. GPT-5 [53] validates the correlation between sub-scene descriptions and QA pairs, introducing temporal grounding steps when necessary. Subsequently, SAM3 [12] generates **spatial bounding boxes** based on the question’s key intent. This process yields data ready for tool-integrated training.

3.2 Multi-Turn Visual Tool-Integrated Reasoning

We argue that it is really challenging to localize sparse but critical information within the marine videos, as supported by the above-mentioned challenges. In this work, we propose to integrate the visual tools for assisting event-centric video understanding. Specifically, we decompose a complicated video understanding problem into multiple steps, where in each step we can call corresponding expert models for generating intermediate outputs. Meanwhile, the intermediate outputs can provide extra visual cues and highlight the relevant information regarding user intents. Formally, we formulate the **multi-turn visual tool-integrated reasoning** as a sequential decision-making process, where a VLM π_θ , constructs a reasoning trajectory to solve a user-specified task. This trajectory consists of interleaved tool invocations, enabling the model to *dynamically plan, gather external evidence*, and *refine its internal representations* for robust and interpretable event-centric video understanding. Given query q , visual input V_0 , and toolbox $\mathcal{T}$, the VLM interacts over K steps. At step k , conditioned on context (q, V_k, H_k) , the model generates reasoning r_k , selects tools $\mathcal{T}_k \subseteq \mathcal{T}$, and invoke calls $\{c_{k,j}\}$. The calling tools return observations $\mathbf{o}_{k,j}$ that augment previous visuals, and this process continues until termination, yielding the final answer a . The trajectory is described as:

$$\zeta = \left((r_1, \mathcal{T}_1, \{c_{1,j}\}, \{\mathbf{o}_{1,j}\}), \dots, (r_K, \mathcal{T}_K, \{c_{K,j}\}, \{\mathbf{o}_{K,j}\}), a \right), \quad (1)$$

we optimize π_θ to maximize answer correctness and cumulative rewards:

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{n=1}^N \sum_{k=1}^K R(\pi_\theta(r_k, \mathcal{T}_k, \mathbf{o}_k \mid q_n, V_{n,k}, H_{n,k}), g_{n,k}) \right], \quad (2)$$

where R denotes a reward function that quantifies the validity and utility of the k -th reasoning step, including logical consistency r_k , appropriate tool $\mathcal{T}_k$, and the groundedness of the resulting observation $\mathbf{o}_k$ conditioned on the state k , which encapsulates the query q_n , visual input $V_{n,k}$, and the history $H_{n,k}$.

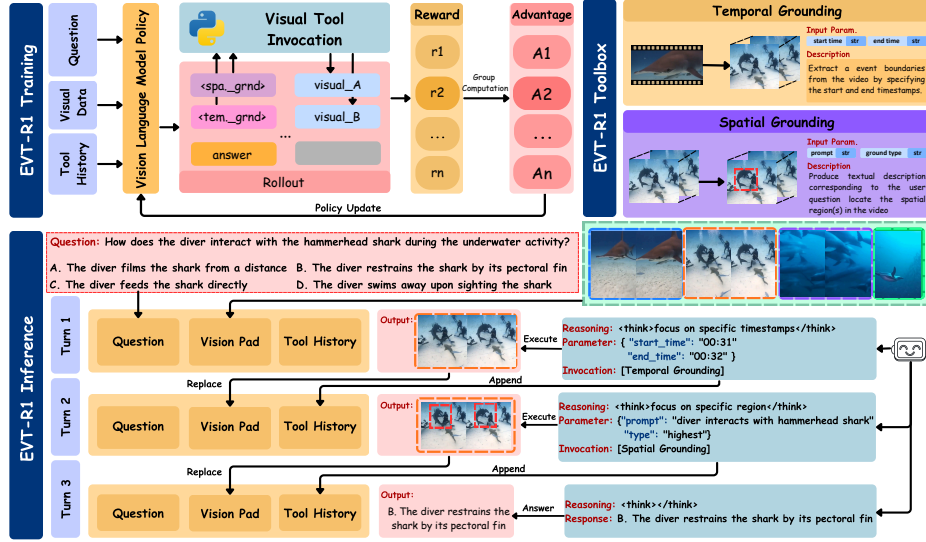


Fig. 4: The training and inference processes of EVT-R1.

3.3 EVT-R1 Overview

The proposed EVT-R1 leverages reinforcement learning (RL) to optimize VLMs for flexible reasoning and strategic visual tool invocation. As illustrated in Fig. 4, the policy model accesses temporal and spatial localization tools within toolbox $\mathcal{T}$. During inference, given a user query and video sequence, the policy π_θ dynamically decides whether to invoke a tool. Upon invocation, the tool modifies visual inputs (*e.g.*, via temporal or spatial grounding), replacing prior visual inputs in the dialogue history to remove outdated tokens and redundancy, which is essential for VLMs to gather critical and relevant information for generating reliable answers. Then the VLM updates its reasoning over updated visual inputs and tool-call history, stopping once the policy finds the evidence sufficient for a final answer. Meanwhile, during training, the VLM policy produces a group of responses, including tool invocations and new visual outputs, or answers only. These rollouts are evaluated by a **reward model** to guide VLM backpropagation. The optimization encourages the VLM to either invoke tools or directly output an answer at each turn. Finally, please note that EVT-R1 is different from chain-of-thought [64], which does not involve tool calling and visual updating.

3.4 Reward Model

Our reward model is modified from group relative policy optimization (GRPO) [23]. Differently, unlike existing works [23, 69], which assign rewards only based on the final output, our EVT-R1 devises separated rewards for **tool usage** and **answer accuracy** as shown in Fig. 5. Our design enables turn-level RL, guiding the model to produce correct answers and learn when and how to use visual

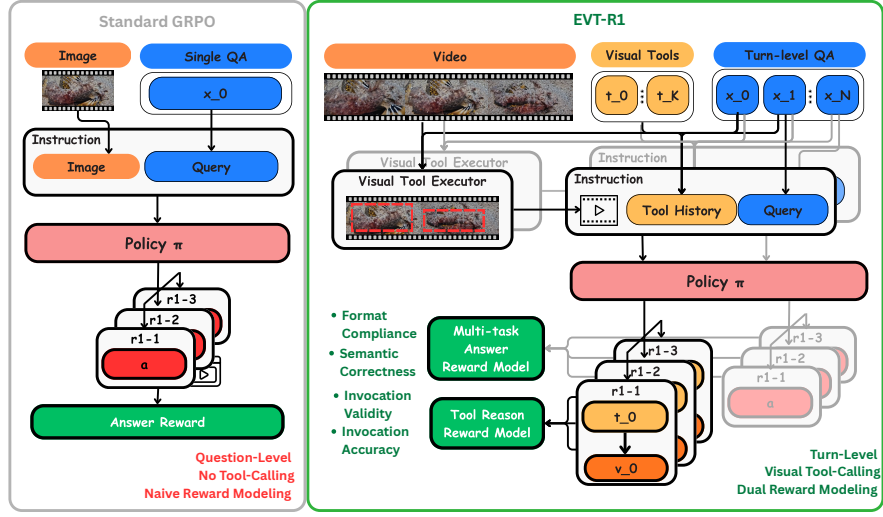


Fig. 5: Compared with GRPO, EVT-R1 devises separate rewards for tool usage and answer accuracy, providing more informative intermediate feedback.

tools effectively. Specifically, we propose a dual-component reward model that corrects final answers while explicitly encouraging effective intermediate tool-use decisions. The dual-reward model consists of: (i) a tool-reasoning reward R_{tool} that assesses whether invoking a tool was valid and accurate at each step, and (ii) a multi-task answer reward R_{ans} that evaluates the correctness of the final response. Please refer to our Supp. for details:

$$R(y_i, g_i) = \lambda \mathbb{I}[\text{tool_turn}](R_{\text{tool}}(y_i, g_i)) + (1 - \lambda) \mathbb{I}[\text{answer_turn}](R_{\text{ans}}(y_i, g_i)), \quad (3)$$

where we detail R_{tool} and R_{ans} as follows:

- **Tool-reasoning reward** R_{tool} evaluates two aspects: *Invocation Validity*, which awards a binary score (1/0) based on whether the tool invocation matches the ground truth at each turn; and *Invocation Accuracy*, an outcome-based metric scoring 1 if the tool’s visual output matches the ground truth, and 0 otherwise.
- **Multi-task answer reward** R_{ans} evaluates two criteria: *Format Compliance*, which checks adherence to the expected output format (e.g., JSON structure, long-short answer); and *Semantic Correctness*, computed via exact string matching for closed-form tasks (e.g., yes/no, MCQ, exact-match) or cosine similarity over embeddings for open-ended answers. Scores are normalized per group to down-weight hard-negative, low-similarity responses.

3.5 Training Objective and Pseudocode

Finally, we detail the whole training procedure of EVT-R1 in Algorithm 1. Concretely, given a multimodal input triplet $[q, V, H]$, the algorithm draws several

Algorithm 1 EVT-R1 Training

Require: Initial policy π_θ , reward R , dataset $\mathcal{D}$, group size G , clip parameters ϵ
Ensure: π_θ

- 1: **for** each training iteration **do**
- 2: Update old policy: $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
- 3: Sample a batch of input queries $\mathcal{Q}_b \sim \mathcal{D}$
- 4: **for** each query $q, V, H \in \mathcal{Q}_b$ **do**
- 5: Process V based on the current H step: $V' = \text{process}(V, H)$
- 6: Sampling G actions $\{o_t^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot \mid q, V', H)$
- 7: Calculate dual rewards R_{tool} or R_{ans} subjected to current turn g_i (Equ 3)
- 8: Calculate turn-level advantage $\hat{A}_t^i$
- 9: **end for**
- 10: Collect all turn-level rollouts into one batch
- 11: Update policy model π_θ by maximizing objective $\mathcal{L}_{\text{GRPO}}(\theta)$ (Equ 4)
- 12: **end for**

output sequences $\{o_i\}_{i=1}^G \sim \pi_{\text{old}}(\cdot \mid q, V, H)$ from the current policy, evaluates their relative quality, and updates the policy parameters to favor higher-scoring responses within each group, optimizing the objective $\mathcal{L}_{\text{GRPO}}(\theta)$ as follows:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \hat{A}_{i,t}, c(\epsilon, \hat{A}_{i,t}) \right] - \beta D_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}], \quad (4)$$

where ϵ, β are hyperparameters, c is the clip advantage, D_{KL} denotes KL divergence, which is used as a regularization to avoid the unstable update of the new policy π_θ compared to the reference policy π_{ref} .

4 Experiments

4.1 Experimental Setting

Baselines. We evaluate general-purpose and domain-specific video VLMs. Due to limited marine models, we fine-tune open-source VLMs on curated marine data using three strategies: supervised fine-tuning (SFT), GRPO [23] for post-training only, and our EVT-R1. We leverage Qwen3-VL [8] as our baseline. We adopt a two-stage fine-tuning strategy: SFT warm-up with one epoch, followed by RL with 4 epochs. For a fair comparison, we also perform SFT and GRPO for 5 epochs. We compare against Video-LLaVA-7B [40], LLaVA-NeXT-Video-7B [85], VideoLLaMA3-7B [10], InternVL3-8B [91], Qwen3-VL-8B-Instruct [8], and closed-source models, Gemini-3.0-Flash [57], Grok4-1-FR [70], and GPT-5-Mini [53], in both tool-intergated and tool-free settings. **Training setup.** We use LORA [27] finetuning with AdamW [43] (lr= 5×10^{-6} , weight decay= 1×10^{-2}), micro-batch size 2 for long sequences (up to 50k tokens), optimizing on a single NVIDIA H800 GPU. Decoding uses temperature 0.1 in bf16 precision for consistency. **Datasets & metrics.** We construct a testing set consisting of 2,000 QA pairs from our MarineEVT for evaluation only (others are used for training) and report overall accuracy. Please refer to our Supp. for details.

Table 2: Experimental comparison between open-source and closed-source VLMs under various settings. († means post-training only, 1st, 2nd, 3rd).

Open-source VLMs w/o Tools								
Model	Training	Tools	SemR.	ConR.	SpaR.	TemR.	CasR.	Avg.
Video-LLaVA-7B	✗	✗	24.40	29.00	8.40	5.80	42.00	21.92
LLaVA-NeXT-Video-7B	✗	✗	35.40	34.33	9.00	6.75	42.67	25.63
VideoLLaMA3-7B	✗	✗	41.00	46.33	5.40	3.50	61.77	31.60
InternVL3-8B	✗	✗	53.40	50.33	17.20	10.33	71.77	40.61
Qwen3-VL-8B-Instruct	✗	✗	58.20	52.77	22.40	14.00	71.00	43.67
Avg. across models	—	—	42.48	42.55	12.48	8.08	57.84	32.69
Closed-source VLMs w/o Tools								
Grok-4-1-FR	✗	✗	41.20	37.33	17.40	6.25	50.67	30.57
Gemini-3.0-Flash	✗	✗	48.20	43.00	27.00	7.75	62.00	37.59
GPT-5-Mini	✗	✗	58.40	30.67	22.80	10.00	66.67	37.71
Avg. across models	—	—	49.27	37.00	22.40	8.00	59.78	35.29
Fine-tuning open-source VLMs w/ Tools								
Qwen3-VL-8B (GRPO†)	✓	✓	44.60	37.00	20.00	10.75	62.66	35.58
Qwen3-VL-8B (SFT)	✓	✓	61.40	53.33	22.60	15.00	74.00	45.27
Ours (EVT-R1)	✓	✓	65.80	53.33	30.60	20.75	74.00	48.89

4.2 Benchmarking SOTAs

Performance analysis. First, we benchmark *open-source* and *closed-source VLMs without tool invocation* and *finetuning VLMs with tool invocation* in Table 2. Among open-source models, Qwen3-VL-8B-Instruct [8] outperforms all competitors, achieving the highest average score of 43.67. This superiority is likely attributed to its dynamic-resolution visual encoder, which effectively handles sequences with varying frame sizes. However, a performance gap remains: open-source models underperform in *spatial reasoning* (avg: 12.48) and *temporal reasoning* (8.08). In contrast, while closed-source models also struggle with *temporal reasoning* (8.00), they demonstrate significantly stronger *spatial reasoning* capabilities (avg: 22.40). We provide qualitative comparisons of various algorithms in Fig. 6.

Meanwhile, we benchmark *closed-source models with and without tool invocation* as shown in Table 3. We observe that integrating tools improves the performance of GPT-5-Mini [53], boosting its average score from 37.71 to 40.35. It indicates that an appropriate reasoning process with tool invocation could lead to performance gains without any re-training. In contrast, using the external tools results in performance degradation for the other two closed-source models, likely due to their limited coordination between tool invocation and their internal reasoning. Notably, EVT-R1 surpasses the best closed-source model GPT-5-Mini [53] with tool invocation by +8.54, demonstrating the superior performance of our EVT-R1.

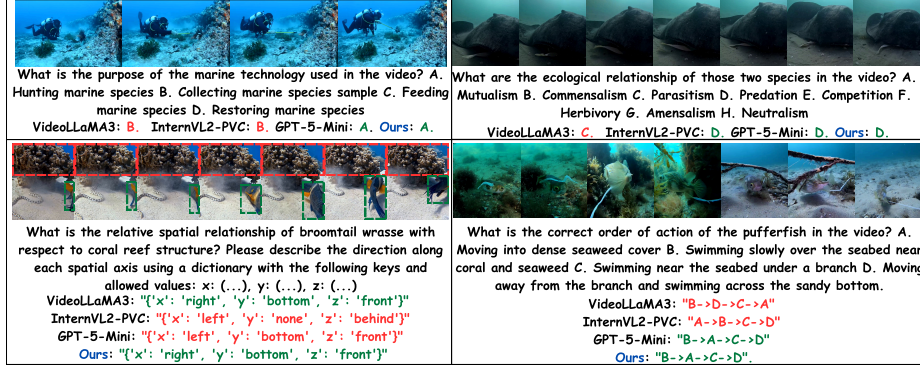


Fig. 6: Experimental result produced by open-source (general-purpose token-compression), closed-source, and our EVT-R1.

Table 3: Experimental comparisons of closed-source VLMs including GPT-5-Mini [53], Gemini-3.0-Flash [57], Grok-4-1-FR [70]: with and without tool invocation.

Model	Tools	SemR.	ConR.	SpaR.	TemR.	CasR.	Avg.
Grok-4-1-FR	✗	41.20	37.33	17.40	6.25	50.67	30.57
	✓	36.00	23.00	13.20	7.75	45.67	25.12 _{-5.45}
Gemini-3.0-Flash	✗	48.20	43.00	27.00	7.75	62.00	37.59
	✓	47.00	43.67	25.60	6.25	59.00	36.30 _{-1.29}
GPT-5-Mini	✗	58.40	30.67	22.80	10.00	66.67	37.71
	✓	58.40	43.67	22.40	13.00	64.33	40.35 _{+2.64}
Ours (EVT-R1)	✓	65.80	53.33	30.60	20.75	74.00	48.89

Table 4: EVT-R1 vs token compress.

Model	Compress.	SemR.	ConR.	SpaR.	TemR.	CasR.	Avg.
LLaVA-1.5-7B [41]	VisionZip [77]	37.00	18.33	10.80	6.00	46.67	23.76
Queen2.5-VL-7B [9]		26.60	13.00	12.40	2.75	47.33	20.42
InternVL2-8B [91]	PVC [75]	45.60	24.67	9.40	2.50	69.00	30.23
Avg. of three	—	36.40	18.67	10.87	3.58	54.33	24.80
Ours (EVT-R1)	—	65.80	53.33	30.60	20.75	74.00	48.89

Table 5: EVT-R1 vs temporal algo.

Model	SemR	ConR	SpaR	TemR	CauR	Average
<i>Key-frame selection algorithm</i>						
MaxInfo [36]	48.00	23.67	19.20	11.00	50.00	30.37
AKS [55]	45.60	20.33	17.20	12.50	48.00	28.72
<i>Frameworks for temporal localization and event-centric video reasoning</i>						
VideoITG [59]	47.40	48.67	20.20	14.25	67.33	39.57
Chain-of-Frames [20]	47.00	52.00	6.00	5.75	68.67	35.88
Ours (EVT-R1)	65.80	53.33	30.60	20.75	74.00	48.89

4.3 Ablation Studies

Comparison with token compression algorithms is first included since this line of algorithms is specifically designed to address the redundancy challenge in temporal sequences. We evaluate three open-source VLMs by using two distinct compression methods: VisionZip [77] and PVC [75]. Experimental results are reported in Table 4, which reveals a significant gap between token compression models and our EVT-R1 (+18.66 over the best-performing token compression algorithm). We attribute the poor performance of token compression algorithms to their inability to effectively localize critical tokens that convey key information, and in some cases, to the loss of such information during compression.

Comparison with temporal-centric algorithms. Table 5 compares our method against state-of-the-art key frame selection and temporal localization approaches. EVT-R1 outperforms the best existing baselines by +18.52 and +9.32, respectively. These results highlight the inability of current temporal-centric models to adequately capture sparse events across frame and temporal dimensions.

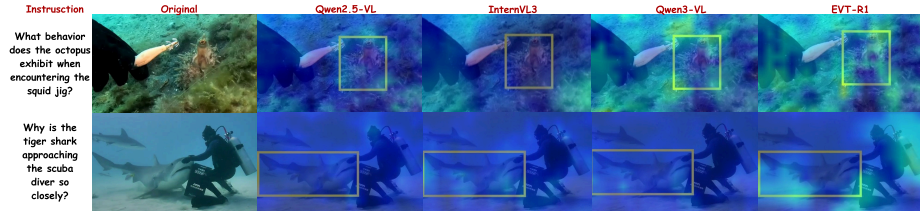


Fig. 7: Attention activation visualization of various models produced by TAM [38].

Table 6: Ablation studies on tool-calling behavior and spatio-temporal accuracy on testing set.

Method	Invocation Validity		Invocation Accuracy	
	Correct	Tool Correct Step	Spatial (IoU ≥ 0.5)	Temporal (IoU ≥ 0.5)
Qwen3-VL (SFT-only)	65.55	84.40	80.09	14.88
EVT-R1 (SFT+RL)	70.00(+4.45)	87.35(+2.95)	83.84(+3.75)	15.00(+0.12)

Table 7: Average accuracy (Mean_{std}) of 3 trials.

Question Task	Blank	Adversarial
Open-end	0.00 _{0.00}	0.00 _{0.00}
MCQ	43.62 _{0.67}	68.86 _{0.44}
Yes-No	40.33 _{1.55}	65.67 _{1.68}
Exact-Match	1.48 _{0.03}	10.04 _{0.85}
Avg. of tasks	21.36 _{0.75}	36.14 _{0.99}

Does RL training improve tool-use behavior? Table 6 evaluates whether EVT-R1 learns better tool-calling behavior. RL training significantly improves the model’s accuracy in selecting the correct tools at the right steps, yielding gains of +**3.75** in spatial IoU and +**0.12** in temporal IoU. Additionally, TAM-generated [38] attention maps, shown in Fig. 7, reveal that this improved spatial localization enables the model to attend more precisely to critical frame tokens.

Does visual input matter? Existing analysis [13] pointed out that VLMs may directly discard visual inputs and yield responses based on language priors. We conduct similar analysis on MarineEVT under two settings: blank or semantically meaningless inputs and adversarial inputs with temporal/semantic inversions, evaluating whether models rely on visual evidence or default to language priors when cues are absent or misleading. To avoid contamination, we directly use the best-performing commercial VLM GPT-5-Mini [53] to do the evaluation since EVT-R1 was optimized on MarineEVT. We observe consistently poor performance in Table 7. Crucially, performance on open-ended/exact-match tasks remains low under adversarial inputs. These results reveal that GPT-5-Mini overly relies on language priors and thus struggles on MarineEVT, highlighting the necessity of VLMs to extract visual cues for answer generation.

Why post-training only does not work. In Table 2, utilizing GRPO for post-training only leads to degraded performance. We manually verified the model outputs and found instability and reward overfitting on long, multi-turn tasks, frequently triggering **infinite reasoning loops**. It may be caused by the lack of domain knowledge, leading to a weak ability to discriminate when to perform tool invocation or yield final answer. Thus, we first fine-tune the VLM on our MarineEVT for one epoch to alleviate the knowledge gap, followed by GRPO for RL training with a cold start. Such a training strategy leads to an observable performance gain as shown in Table 8, revealing that the SFT is an essential

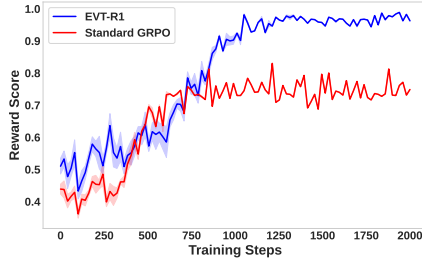


Fig. 8: We compare the reward score curves of GRPO and EVT-R1.

Table 8: GRPO vs. EVT-R1.

Method	Setting	SemR.	ConR.	SpaR.	TemR.	CasR.	Avg.
GRPO	RL only	44.60	37.00	20.00	10.75	62.66	35.58
EVT-R1		44.80	39.00	21.40	13.75	63.33	36.52
GRPO	SFT+RL	63.40	53.33	29.60	18.00	74.67	48.20
EVT-R1		65.80	53.33	30.60	20.75	74.00	48.89

Table 9: Different λ coefficient values.

Coefficient	SemR.	ConR.	SpaR.	TemR.	CasR.	Avg.
$\lambda = 0.00$	44.80	38.67	18.80	14.50	65.00	36.35
$\lambda = 0.25$	54.40	37.33	19.40	12.50	66.33	37.99
$\lambda = 0.50$	54.00	38.33	18.20	12.25	66.67	37.89
$\lambda = 0.75$	55.40	41.33	20.80	13.50	68.33	39.87
$\lambda = 1.00$	55.00	37.00	13.80	14.00	67.00	37.36

step for adapting general-purpose VLMs to specific domains, analogous to that students should have basic knowledge to determine their learning actions.

Comparison with GRPO. Under the same experimental settings: RL only and SFT+RL, our EVT-R1 outperforms GRPO by **+0.94** and **+0.69**, respectively, demonstrating the effectiveness of the proposed dual-reward design. Meanwhile, we provide the reward curve of GRPO and EVT-R1 in Fig. 8. Our reward function produces higher rewards and more stable convergence compared with GRPO, *e.g.*, a lower variance at later stages (steps 1000–2000), reflecting stable policy updates and effective reward assignment. Finally, we ablate the reward coefficient λ in Table 9. $\lambda = 0.75$ achieves the best performance (avg. 39.87).

5 Conclusion and Acknowledgment

Conclusion. In this work, we have introduced the first event-centric marine video understanding dataset, MarineEVT, highlighting the specific and intrinsic challenges of marine video: the deep domain expertise requirement and the difficulty to localize and understand the sparse, unpredictable, and unevenly distributed marine events. Besides MarineEVT, we also introduce EVT-R1, the first event-centric visual tool-integrated reasoning framework, where we decompose the video understanding task into the multi-turn tool-integrated reasoning. EVT-R1 demonstrates a stronger ability to localize critical information spatially and temporally than existing algorithms. It also introduces a new direction of using visual tools for the complicated video understanding tasks.

Acknowledgement. This project was partially supported by Bridging Horizons: An AI-Powered STEM Learning Initiative in Space and Marine Education under the EdUHK–HKUST Joint Centre for Artificial Intelligence and the Marine Conservation Enhancement Fund MCEF22112, and an internal grant from HKUST (R9429). We would also like to express our sincere gratitude to the “Sustainable Smart Campus as a Living Lab” (SSC) program at HKUST for its vital support. The program and its dedicated staff not only contributed essential funding and coordination but also fostered the integration of sustainability into campus operations, providing a real-world demonstration of the principles that underpin this research.

References

1. Discovery video-source webpage, <https://www.discovery.com/> 6
2. Instagram video-source webpage, <https://www.instagram.com/> 6
3. Mbari video-source webpage, <https://www.mbari.org/> 6
4. National geographic video-source webpage, <https://www.nationalgeographic.com/> 6
5. Youtube video-source webpage, <https://www.youtube.com/> 6
6. Alawode, B., Guo, Y., Umamar, M., Werghi, N., Dias, J., Mian, A., Javed, S.: Utb180: A high-quality benchmark for underwater tracking. In: ACCV (2022) 5, 6
7. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 5
8. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 2, 5, 8, 11, 12
9. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 2, 5, 13
10. Boqiang Zhang, Kehan Li, Z.C.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025), <https://arxiv.org/abs/2501.13106> 11
11. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Nibbles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015) 4, 5
12. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025), <https://arxiv.org/abs/2511.16719> 7, 8
13. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? (2024), <https://arxiv.org/abs/2403.20330> 14
14. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 5
15. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning

- for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 5
16. Contributors, V.: Computer Vision Across the Marine Sciences. Open Textbook (2024), <https://oceancv.org/> 2
 17. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset (2018), <https://arxiv.org/abs/1804.02748> 5
 18. Fan, B., Chen, W., Cong, Y., Tian, J.: Dual refinement underwater object detection network. In: European Conference on Computer Vision (ECCV). pp. 275–291. Springer (2020) 2
 19. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition (2019), <https://arxiv.org/abs/1812.03982> 4
 20. Ghazanfari, S., Croce, F., Flammarion, N., Krishnamurthy, P., Khorrami, F., Garg, S.: Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning (2026), <https://arxiv.org/abs/2506.00318> 13
 21. Gong, Z., et al.: CLIBD: Bridging vision and genomics for biodiversity monitoring at scale. ICLR / Nature Methods (2024), <https://openreview.net/forum?id=d5HUnyByAI> 2
 22. González-Sabbagh, S.P., Robles-Kelly, A.: A survey on underwater computer vision. ACM Computing Surveys 56(4) (2023). <https://doi.org/10.1145/3578516>, <https://dl.acm.org/doi/full/10.1145/3578516> 2
 23. Guo, D., et al., Y.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature 645(8081), 633–638 (Sep 2025). <https://doi.org/10.1038/s41586-025-09422-z>, <http://dx.doi.org/10.1038/s41586-025-09422-z> 5, 9, 11
 24. Han, H., Wang, W., Zhang, G., Li, M., Wang, Y.: Coralvqa: A large-scale visual question answering dataset for coral reef image understanding (2025), <https://arxiv.org/abs/2507.10449> 5, 6
 25. Han, S., Huang, W., Shi, H., Zhuo, L., Su, X., Zhang, S., Zhou, X., Qi, X., Liao, Y., Liu, S.: Videoesspresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection (2024), <https://arxiv.org/abs/2411.14794> 5
 26. Hong, L., Wang, X., Zhang, G., Zhao, M.: Usod10k: a new benchmark dataset for underwater salient object detection. IEEE Transactions on Image Processing (TIP) (2023) 2, 5
 27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), <https://arxiv.org/abs/2106.09685> 11
 28. Hughes, T.P., Kerry, J.T., Simpson, T.: Large-scale bleaching of corals on the great barrier reef. Ecology 99(2) (2018) 2
 29. Jalal, A., et al.: Fish detection and species classification in underwater environments using deep learning with temporal information. IEEE Transactions on Image Processing (2023), <https://ieeexplore.ieee.org/document/XXXXXXX> 2
 30. Jang, J., Park, J., Kim, J., Kwon, H., Sohn, K.: Knowing where to focus: Event-aware transformer for video grounding. In: IEEE/CVF International Conference on Computer Vision. pp. 13846–13856 (2023) 4
 31. Katona, Z., et al.: MARINE: A computer vision model for detecting rare predator–prey interactions in animal videos. In: ECCV Workshops / arXiv preprint (2025), <https://www.researchgate.net/publication/389540473> 2
 32. Khan, F.F., Li, X., Temple, A.J., Elhoseiny, M.: Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In: Pro-

- ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20496–20506 (October 2023) [6](#)
33. Kim, G., Jeong, T.K., Park, J.: Surgical video understanding with label interpolation. arXiv preprint arXiv:2509.18802 (2025) [5](#)
 34. Li, B., Huo, T., Zhang, D., Zhao, Z., Gao, J., Li, X.: Exploring the underwater world segmentation without extra training. arXiv preprint arXiv:2511.07923 (2025) [6](#)
 35. Li, P., Lu, Y., Song, P., Li, W., Yao, H., Xiong, H.: Eventvl: Understand event streams via multimodal large language model (2025), <https://arxiv.org/abs/2501.13707> [5](#)
 36. Li, P., Abdullaeva, I., Gambashidze, A., Kuznetsov, A., Oseledets, I.: Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding (2025), <https://arxiv.org/abs/2502.03183> [13](#)
 37. Li, Y., Yang, X., Xu, D., Yu, Y., Zhao, L., Hu, X., Li, J., Heng, P.A.: Surgpub-video: A comprehensive surgical video dataset for enhanced surgical intelligence in vision-language model (2025), <https://arxiv.org/abs/2508.10054> [5](#)
 38. Li, Y., Wang, H., Ding, X., Wang, H., Li, X.: Token activation map to visually explain multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 48–58 (October 2025) [14](#)
 39. Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S.: Watermask: Instance segmentation for underwater imagery. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1305–1315 (2023) [2](#), [5](#), [6](#)
 40. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection (2024), <https://arxiv.org/abs/2311.10122> [11](#)
 41. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) [13](#)
 42. Liu, S., Li, J., Zhao, G., Zhang, Y., Meng, X., Yu, F.R., Ji, X., Li, M.: Eventgpt: Event stream understanding with multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29139–29149 (2025) [5](#)
 43. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), <https://arxiv.org/abs/1711.05101> [11](#)
 44. Minglong, W., et al.: A machine learning-driven framework for enhancing underwater ecological monitoring. *Frontiers in Environmental Science* (2026), <https://www.frontiersin.org/journals/environmental-science/articles/10.3389/fenvs.2025.1689855/full> [2](#)
 45. Mukherjee, R., Singh, S., McWilliams, J., Sattar, J.: The common objects underwater (cou) dataset for robust underwater object detection (02 2025). <https://doi.org/10.48550/arXiv.2502.20651> [5](#)
 46. Murphy, K.: Reinforcement learning: An overview (2025), <https://arxiv.org/abs/2412.05265> [5](#)
 47. Pantazis, O.: Data-Efficient Computer Vision for Biodiversity Monitoring. Ph.D. thesis, University College London (2023), <https://discovery.ucl.ac.uk/id/eprint/10200299/> [2](#)
 48. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), <https://arxiv.org/abs/2305.18290> [5](#)

49. Ramanathan, V., Liang, P., Fei-Fei, L.: Video event understanding using natural language descriptions. In: IEEE International Conference on Computer Vision. pp. 905–912 (2013) [4](#)
50. Sanders, K., Van Durme, B.: A survey of video datasets for grounded event understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7314–7327 (2024) [4](#)
51. Shi, Z., Guan, C., Li, Q., Liang, J., Cao, L., Zheng, H., Gu, Z., Zheng, B.: Detecting marine organisms via joint attention-relation learning for marine video surveillance. IEEE Journal of Oceanic Engineering **47**(4), 959–974 (2022) [5](#)
52. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos (2014), <https://arxiv.org/abs/1406.2199> [4](#)
53. Singh, A., et al.: Openai gpt-5 system card (2025), <https://arxiv.org/abs/2601.03267> [3](#), [7](#), [8](#), [11](#), [12](#), [13](#), [14](#)
54. Souček, T., Moravec, J., Lokoč, J.: Transnet: A deep network for fast detection of common shot transitions. arXiv preprint arXiv:1906.03363 (2019) [7](#)
55. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding (2025), <https://arxiv.org/abs/2502.21271> [13](#)
56. Tang, Y.Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J., Xu, C.: Video understanding with large language models: A survey (2025), <https://arxiv.org/abs/2312.17432> [4](#)
57. Team, G.: Gemini: A family of highly capable multimodal models (2025), <https://arxiv.org/abs/2312.11805> [11](#), [13](#)
58. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) [2](#), [5](#)
59. Wang, S., Chen, G., an Huang, D., Li, Z., Li, M., Liu, G., Alvarez, J.M., Zhang, L., Yu, Z.: Videoitg: Multimodal video understanding with instructed temporal grounding (2026), <https://arxiv.org/abs/2507.13353> [13](#)
60. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) [5](#)
61. Wang, X., Wang, S., Tang, C., Zhu, L., Jiang, B., Tian, Y., Tang, J.: Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19248–19257 (2024) [5](#)
62. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023) [5](#)
63. Wang, Z., Zhang, Z., Pang, T., Du, C., Zhao, H., Zhao, Z.: Orient anything: Learning robust object orientation estimation from rendering 3d models. arXiv:2412.18605 (2024) [7](#)
64. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), <https://arxiv.org/abs/2201.11903> [9](#)
65. Wille, M., Fischer, T., Raine, S.: Are all marine species created equal? performance disparities in underwater object detection. arXiv preprint arXiv:2508.18729 (2025) [2](#)

66. Winther, J.G., Dai, M., Rist, T., Hoel, A.H., Li, Y., Trice, A., Morrissey, K., Juinio-Meñez, M.A., Fernandes, L., Unger, S., et al.: Integrated ocean management for a sustainable ocean economy. *Nature ecology & evolution* **4**(11), 1451–1458 (2020) [2](#)
67. Wong, Y.K., Liang, H., Ma, Z., Chen, Y., Zheng, Z., Gotama, R., Sebastian, P., Sparks, L.D., Yeung, S.K.: Orca: Object recognition and comprehension for archiving marine species. *arXiv preprint arXiv:2512.21150* (2025) [2](#), [5](#)
68. Wong, Y.K., To, T.A., Zhang, J., Zheng, Z., Yeung, S.K.: Marineeval: Assessing the marine intelligence of vision-language models. *arXiv preprint arXiv:2512.21126* (2025) [3](#), [5](#), [6](#)
69. Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., Nahrstedt, K.: Vtool-rl: Vllms learn to think with images via reinforcement learning on multimodal tool use (2025), <https://arxiv.org/abs/2505.19255> [5](#), [9](#)
70. xAI: Grok 4 - xai. <https://x.ai/news/grok-4> (July 2025), accessed: 2026-02-27 [11](#), [13](#)
71. Xie, Y., Kong, L., Chen, K., Zheng, Z., Yu, X., Yu, Z., Zheng, B.: Uveb: A large-scale benchmark and baseline towards real-world underwater video enhancement. In: *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)* (2024) [2](#)
72. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. pp. 6787–6800. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.emnlp-main.544>, <https://aclanthology.org/2021.emnlp-main.544/> [5](#)
73. Xu, W., Wang, C., Liang, D., Zhao, Z., Jiang, X., Zhang, P., Bai, X.: Nautilus: A large multimodal model for underwater scene understanding. *arXiv preprint arXiv:2510.27481* (2025) [6](#)
74. Xue, X., Zhou, Y., Yan, D., Tao, L., Li, J., Li, Y., Zhang, H., Xiao, R.: Uvllm: Benchmarking video language model for underwater world understanding (2025), <https://arxiv.org/abs/2507.02373> [2](#), [3](#), [5](#), [6](#)
75. Yang, C., Dong, X., Zhu, X., Su, W., Wang, J., Tian, H., Chen, Z., Wang, W., Lu, L., Dai, J.: Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 24939–24949 (2025) [13](#)
76. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. *arXiv:2406.09414* (2024) [7](#)
77. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467* (2024) [13](#)
78. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.Y., Zhang, Y.Q., Yan, L., Qiao, M., Wu, Y., Wang, M.: Dapo: An open-source llm reinforcement learning system at scale (2025), <https://arxiv.org/abs/2503.14476> [5](#)
79. Zang, C., Wang, H., Pei, M., Liang, W.: Discovering the real association: Multimodal causal reasoning in video question answering. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 19027–19036 (June 2023) [4](#)

80. Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J.S., Cao, J., Farhadi, A., Choi, Y.: MERLOT: multimodal neural script knowledge models. CoRR **abs/2106.02636** (2021), <https://arxiv.org/abs/2106.02636> 5
81. Zeng, Z., Zhuo, Z., Jia, X., Zhang, E., Wu, J., Zhang, J., Wang, Y., Low, C.H., Jiang, J., Zheng, Z., Cao, X., Ban, Y., Dou, Q., Liu, Y., Jin, Y.: Surgvbm: A large vision-language model and systematic evaluation benchmark for surgical intelligence (2025), <https://arxiv.org/abs/2506.02555> 5
82. Zhang, C., Liu, L., Huang, G., Wen, H., Zhou, X., Wang, Y.: Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark (2024), <https://arxiv.org/abs/2405.19818> 2, 5, 6
83. Zhang, D., Rong, C., Li, B., Wang, F., Zhao, Z., Gao, J., Li, X.: Uwbench: A comprehensive vision-language benchmark for underwater understanding (2025), <https://arxiv.org/abs/2510.18262> 2, 6
84. Zhang, P., Yan, T., Liu, Y., Lu, H.: Fantastic animals and where to find them: Segment any marine animal with dual sam (2024), <https://arxiv.org/abs/2404.04996> 2
85. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/> 11
86. Zhao, Y., Zhang, H., Xie, L., Hu, T., Gan, G., Long, Y., Hu, Z., Chen, W., Li, C., Xu, Z., et al.: Mmvu: Measuring expert-level multi-discipline video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8475–8489 (2025) 5
87. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., Zhou, J., Lin, J.: Group sequence policy optimization (2025), <https://arxiv.org/abs/2507.18071> 5
88. Zheng, Z., Chen, Y., Zeng, H., Vu, T.A., Hua, B.S., Yeung, S.K.: Marineinst: A foundation model for marine image analysis with instance visual description. ECCV (2024) 2, 3, 5, 6
89. Zheng, Z., Zhang, J., Vu, T.A., Diao, S., Tim, Y.H.W., Yeung, S.K.: Marinegpt: Unlocking secrets of ocean to the public (2023), <https://arxiv.org/abs/2310.13596> 2, 3, 5, 6
90. Zhong, J., Li, M., Zhang, H., Qin, J.: Combining photogrammetric computer vision and semantic segmentation for fine-grained understanding of coral reef growth under climate change. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 186–195 (2023) 2
91. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 5, 11, 13