CUEBENCH: Advancing Unified Understanding of Context-Aware Video Anomalies in Real-World

Yating Yu*, Congqi Cao*†, Zhaoying Wang, Weihua Meng, Jie Li, Yuxin Li, Zihao Wei, Zhongpei Shen, Jiajun Zhang

Northwestern Polytechnical University, Xi'an Shaanxi, 710129, China
yatingyu@mail.nwpu.edu.cn, congqi.cao@nwpu.edu.cn, wangzhaoying@mail.nwpu.edu.cn

Abstract

How far are deep models from real-world video anomaly understanding (VAU)? Current works typically emphasize detecting unexpected occurrences deviating from normal patterns or comprehending anomalous events with interpretable descriptions. However, they exhibit only a superficial comprehension of real-world anomalies, with limited breadth in complex principles and subtle contexts that distinguish the anomalies from normalities, *e.g.*, climbing cliffs with safety gear *vs.* without it. To this end, we introduce **CUEBENCH**, the first of its kind Benchmark, devoted to Context-aware video anomalies within a Unified Evaluation framework. We comprehensively establish an event-centric hierarchical taxonomy that anchors two core event types: 14 conditional and 18 absolute anomaly events, defined by their refined semantics from diverse contexts across 174 scenes and 198 attributes. Based on this, we propose to unify and benchmark context-aware VAU with various challenging tasks across recognition, temporal grounding, detection, and anticipation. It also serves as a rigorous and fair probing evaluation suite for generalized and specialized vision-language models (VLMs) across both generative and discriminative paradigms. To address the challenges underlying CUEBENCH, we further develop **CUE-R1** based on R1-style reinforcement fine-tuning with verifiable, task-aligned, and hierarchy-refined rewards in a unified generative manner. Extensive results on CUEBENCH reveal that, existing VLMs are still far from satisfactory real-world anomaly understanding, while our CUE-R1 surpasses these state-of-the-art approaches by over 24% on average.

Code — https://github.com/Mia-YatingYu/Cue-R1
Extended version — https://arxiv.org/abs/2511.00613

1 Introduction

Video anomaly understanding (VAU) derived from general video understanding, emphasizes the automated comprehension of anomalous events in videos, which encompasses a diverse range of tasks including anomaly detection (Ristea et al. 2024; Cai et al. 2021; Cao, Lu, and Zhang 2024; Yan et al. 2023), recognition (Wu et al. 2024b; Yu et al. 2025b), and localization (Zhou et al. 2016). At its core, video anomaly detection (VAD) tends to detect deviations from the learned normal patterns (Zhu, Bao, and Yu 2022). Keeping pace with the development of VLMs (Radford et al. 2021; Bai et al. 2025; Yu et al. 2025a; Zhang et al. 2025a; Comanici et al. 2025; Hurst et al. 2024), a growing body of works has emerged to comprehend anomalies in open-vocabulary settings (Wu et al. 2024a; Li et al. 2025a; Zanella et al. 2024) and further in a VQA manner with interpretable explanations (Du et al. 2024b; Zhang et al. 2025c; Ye, Liu, and He 2025; Xu et al. 2025; Du et al. 2024a; Ma et al. 2025; Huang et al. 2025). Given that real-world anomalies are complex, diverse, and evolving, there is a need for a more **realistic** and **universal** comprehension that aligns with human experiences and societal norms. With current advancements, a natural question raises: *How far are current VLMs from truly understanding of real-world video anomalies?*

While existing works are appealing, they oversimplify the nature of real-world anomalies. Some studies have explored the role of contextual semantics in VAU (Wu et al. 2024a; Ma et al. 2025), but their focus has largely been on comprehending traditional *absolute anomaly events* (*e.g.*, "explosion", "car crash") or simple *deviations* (*e.g.*, "biking" instead of the expected "walking"), where contextual cues are not decisive in determining normality *vs.* anomaly. Recent efforts have drawn attention to scene dependencies underlying anomalies (Cao et al. 2023, 2025; Zhang et al. 2025b), yet the reliance on scene-only contexts and the sparsity of scene-dependent anomalies reveal a substantial gap in real-world VAU. In practice, the same event (*e.g.*, "climbing") could be interpreted as normal or abnormal depending on both scene and attribute context: "climbing cliffs with safety gear" is normal, whereas "climbing cliffs without any protection" is clearly abnormal, due to the inherent risks in cliff scenes and the need for additional precaution. Such *conditional anomaly events*, implying ambiguous boundaries and subtle context dependencies from both scenes and attributes, remain largely underexplored in existing works.

For a long period, VAU research has followed task-specific paradigms, designing specialized architectures and loss functions to cater to unique requirements of separate tasks and benchmarks. Despite the breadth, VAU has predominantly focused on specific capabilities like VAD, multi-modal retrieval and VQA, leading to fragmented and incom-

*These authors contributed equally.
†Corresponding author.

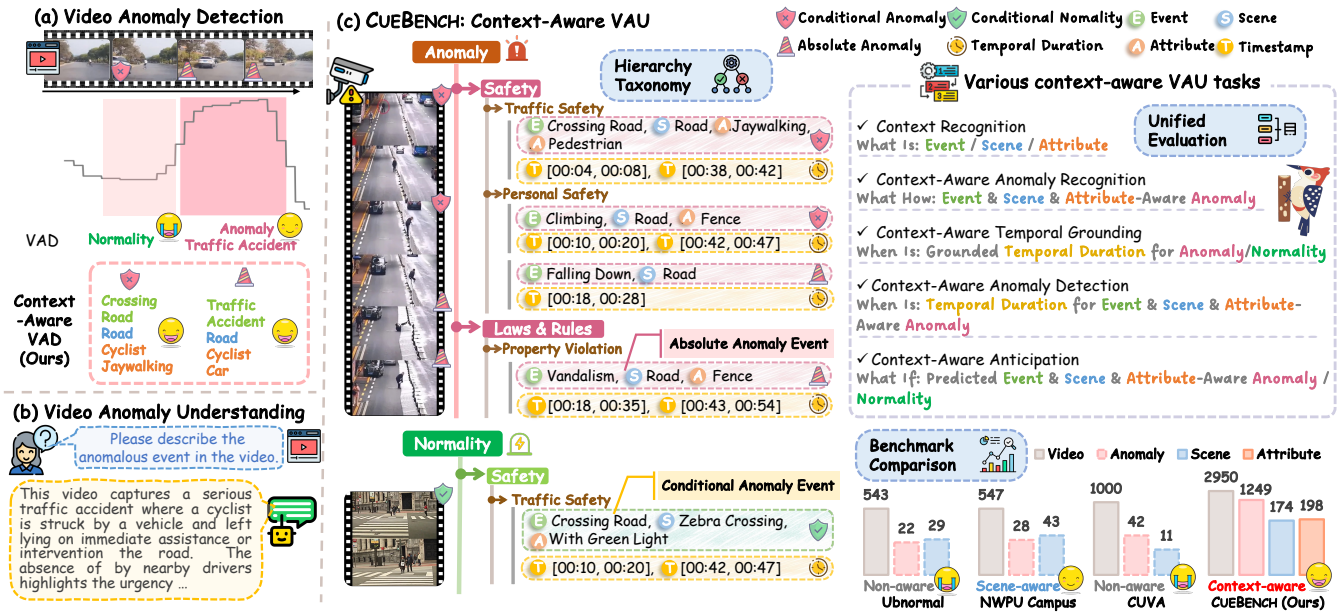

Figure 1: **Comparison of existing benchmarks.** (a) Traditional VAD aims to detect deviations from normal patterns and identify the time window of the occurring anomaly, yet exhibiting insufficient comprehension of subtle anomalies and lacking context-awareness (*e.g.*, *cyclist jaywalking while crossing road*). (b) Current VAU benchmarks primarily emphasize the interpretation of absolutely anomalous events with explainable outputs. (c) Our large-scale CUEBENCH features a diverse collection of **context-aware anomalies and normalities** from real-world scenarios, organized within a comprehensive **hierarchical taxonomy**, and supports **unified evaluation** across five challenging VAU tasks.

patible solutions. Such fragmentation underscores the need for a unified framework benchmarking diverse demands, fostering holistic and integrated real-world VAU.

To satisfy these desiderata, we develop **CUEBENCH**, the first benchmark dedicated to unified, context-aware video anomaly understanding in real-world. Compared with existing benchmarks in Figure 1, CUEBENCH highlights the following distinct characteristics:

- **Context Awareness.** Given the complex context dependencies of real-world anomalies, CUEBENCH is the first to introduce and integrate the concepts of 18 *absolute anomaly events* (*e.g.*, "falling down", "vandalism") and 14 *conditional anomaly events* (*e.g.*, "crossing road", "climbing") *w.r.t.* subtle contextual cues drawn from 174 scenes and 198 attributes. Note that both anomalies and normalities in CUEBENCH are represented as context triplets comprising events, scenes, and attributes. Hence, along with the diversity of anomalies (1249), the normalities (194) are context-dependent and diversified as well, going beyond the rare occurrences and monotonous absolute anomalies in existing benchmarks.
- **Comprehensive Hierarchical Taxonomy.** As anomalies vary widely in types, contexts, and impacts, we comprehensively build a 5-level event-centric hierarchical taxonomy, extending from the fundamental anomaly *vs.* normality to fine-grained triplets. The key insight is that anomaly errors often carry far more severe consequences than simple context misinterpretations. The refined differentiation of violation and inherent severity (*e.g.*, on *safety*, *laws&rules*, *life&health*) in the hierarchy enables trustworthy evaluation and prioritization in real-world.
- **Unified Evaluation Framework.** Through a suite of five test tasks and crafted evaluation metrics, CUEBENCH enables comprehensive gauges of models' capabilities across recognition, detection, grounding, and anticipation. Within the unified task space, we hope that further development of universal architectures and training objectives will continue to advance this field.

Leveraging this dataset, we present **CUE-R1**, a unified generative approach that incorporates supervised and reinforcement fine-tuning with verifiable, task-aligned, and hierarchy-refined rewards tailored to context-aware VAU. Extensive results on CUEBENCH reveal that existing generalized and specialized VLMs, both generative and discriminative, remain unsatisfactory, while CUE-R1 provides new insights for developing a universal solution.

2 Benchmark: CUEBENCH

2.1 Data Statistics

We compare CUEBENCH with existing VAU benchmarks in Table 1. Generally, our CUEBENCH comprises 2,950 newly collected videos sourced from multiple domains on YouTube totaling 54.5 hours of footage. Each video ranges from 10s to 5min in length, with rich annotations of contexts and anomaly labels. The labeled context-aware segments span approximately 62% of the total duration.

Benchmark	Domain	Length	#Video	#Abs.	#Con.	#Norm.	Dependency	Task Setup				
								R	G	D	A	Q
Traditional Video Anomaly Detection Datasets												
Subway Entrance (Adam et al. 2008)	Pedestrian	1.5h	1	5	*NA*	1	Deviation	✗	✗	✓	✗	✗
UCSD Ped1 (Wang and Miao 2010)	Pedestrian	0.1h	5	5	*NA*	1	Deviation	✗	✗	✓	✗	✗
CUHK Avenue (Lu, Shi, and Jia 2013)	Pedestrian	0.5h	5	5	*NA*	1	Deviation	✗	✗	✓	✗	✗
ShanghaiTech (Luo, Liu, and Gao 2017)	Pedestrian	-	13	11	*NA*	1	Deviation	✗	✗	✓	✗	✗
UCF-Crime (Sultani, Chen, and Shah 2018)	Crime	128h	1900	13	*NA*	1	Event	✗	✗	✓	✗	✗
Street Scene (Ramachandra and Jones 2020)	Traffic	3.7h	81	17	*NA*	1	Deviation	✗	✗	✓	✗	✗
XD-Violence (Wu et al. 2020)	Violence	217h	4754	6	*NA*	1	Event	✗	✗	✓	✗	✗
Ubnormal (Acsintoae et al. 2022)	Pedestrian	2.2h	543	22	*NA*	1	Deviation	✗	✗	✓	✗	✗
NWPU Campus (Cao et al. 2023)	Pedestrian	16h	547	28	4	1	Event, Scene	✗	✗	✓	✓	✗
MSAD (Zhu et al. 2024)	Multiple	-	720	55	*NA*	1	Event	✗	✗	✓	✗	✗
Video Anomaly Understanding Datasets												
CUVA (Du et al. 2024b)	Multiple	32.5h	1000	42	*NA*	*NA*	Event	✓	✗	✓	✗	✓
HAWK (Tang et al. 2024)	Mixture	142.5h	8000	-	*NA*	*NA*	Event	✗	✗	✗	✗	✓
HIVAU-70k (Zhang et al. 2025c)	Mixture	-	5443	19	*NA*	1	Event	✗	✗	✓	✗	✓
CUEBENCH (Ours)	Multiple	54.5h	2950	18 → 840	14 → 409	14 → 194	Event, Scene, Attribute	✓	✓	✓	✓	✓

Table 1: We review existing VAD and VAU benchmarks and highlight key characteristics of CUEBENCH. "Mixture" denotes the combination of existing public datasets. Different from others, CUEBENCH is the first large-scale benchmark for context-aware VAU. Due to anomaly dependencies of contexts from absolute and conditional anomaly events with different scenes and attributes, #Anomaly and #Normality are highly diversified, progressing from event categories (L-4) to (→) context triplets (L-5) in hierarchy taxonomy. It is designed to evaluate various tasks including anomaly recognition (R), temporal grounding (G), anomaly detection (D) and context anticipation (A), all of which can be approached in a unified VQA manner (Q).

Context Indispensability. Unlike existing VAU benchmarks, CUEBENCH is explicitly designed to be context-indispensable, encompassing 174 scenes and 198 attributes besides 32 event categories *w.r.t.* 18 *absolute anomaly events* and 14 *conditional anomaly events*. Notably, it features 1,443 distinct context triplets, each representing a combination of an event with various scenes and attributes, linked to either an anomaly or normality. This yields **840 absolute anomalies** (*e.g.*, ⟨`vandalism, road, fence`⟩), *i.e.*, triplets involving *absolute anomaly events* that remain anomalous across various scenes and attributes, **409 conditional anomalies** (*e.g.*, ⟨`crossing road, road, pedestrian jaywalking`⟩), and **194 conditional normalities** (*e.g.*, ⟨`crossing road, zebra crossing, green light`⟩), *i.e.*, triplets containing *conditional anomaly events* whose abnormal/normal states hinge on context cues. Such rich contextual grounding ensures that understanding anomalies in CUEBENCH requires nuanced reasoning beyond superficial event recognition.

Event-centric Hierarchy Taxonomy. CUEBENCH incorporates a comprehensive five-level event-centric hierarchy taxonomy, where each leaf node in Level 5 (L-5) represents a distinct context triplet *w.r.t.* a normality or anomaly. At the top, L-1 distinguishes two fundamental states: Anomaly *vs.* Normality. This branches into three L-2 domains and further into nine L-3 effects underscoring both the shared and distinct characteristics across various real-world anomalies. Note that L-4 comprises 34 event nodes, as conditional anomaly events *i.e.*, "throwing rubbish" and "smoking", exhibit two distinct anomaly effects depending on contexts.

Training & Testing Settings. To ensure an open-world setting, we divide the dataset into two sets: the test set comprising 1,222 videos covering all 1,443 distinct context triplets, while the training set with the remaining 1,728 videos containing only 440 context triplets. Notably, the test set features higher density of context triplets than the training set (1.68 *vs.* 1.21 triplets per video), enabling a more challenging and realistic evaluation.

2.2 Task Definition

To comprehensively evaluate models' ability of VAU, we define a unified suite of five tasks built around the concept of context-aware reasoning. Each task targets a distinct yet complementary aspect of semantic, temporal and causal anomaly understanding, encouraging holistic perception and interpretation of video events under real-world complexity.

What Is. (1) *Context Recognition*: Identify the specific contextual elements (*i.e.*, events, scenes, or attributes) present in the video. This task serves as a fundamental perceptual evaluation of models' context-aware capabilities.

What How. (2) *Context-Aware **Anomaly** Recognition.* Identify the specific context triplets occurring in the video, and determine the existence of accurate absolute and conditional anomalies accordingly. We introduce two paradigms for this task: **(a)** Automatically distinguish all anomalies from normalities with their corresponding contexts in a *top-down* manner. **(b)** Extract the context triplets (whether anomalous or not), then assign anomaly scores to each group based on their semantics in a *bottom-up* manner.

When Is. (3) *Context-Aware Temporal Grounding.* Ground target moments *i.e.*, one or more continuous intervals from untrimmed videos according to the queries based on context

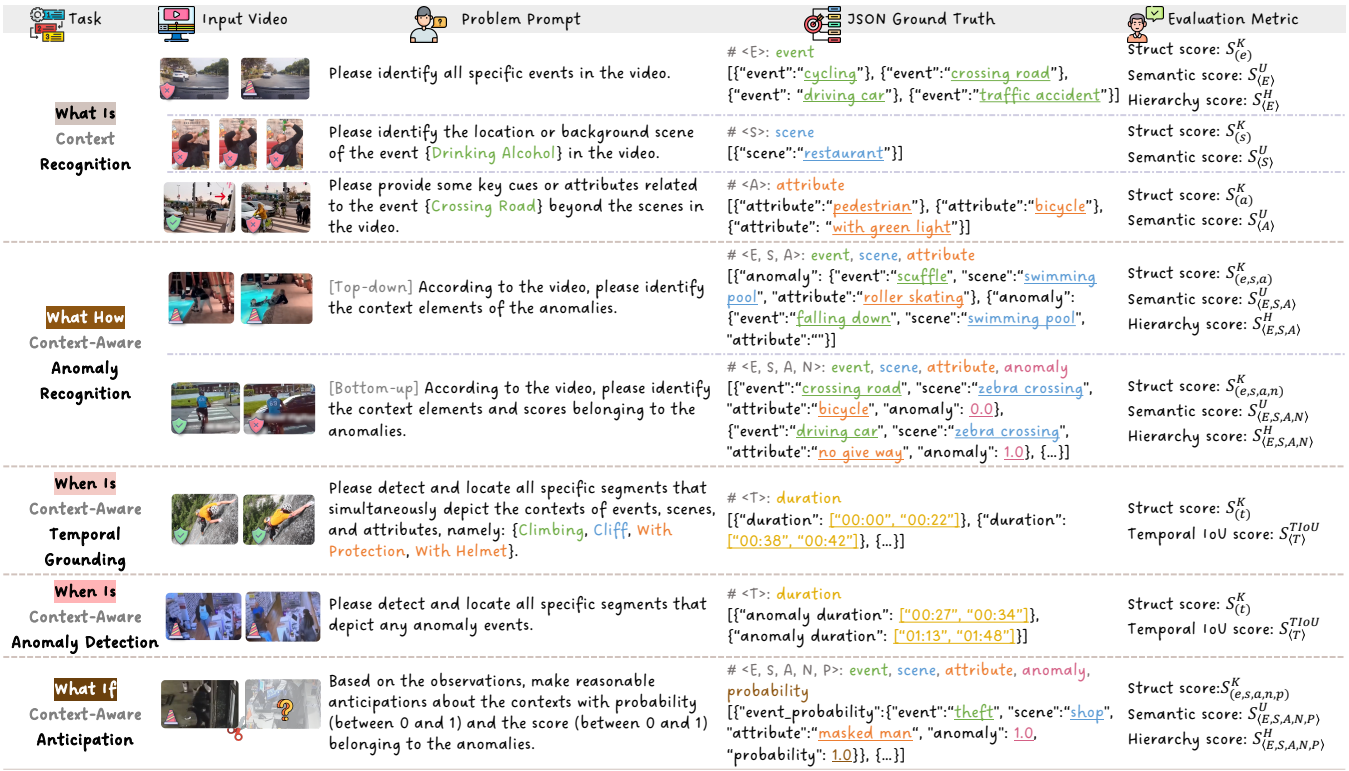

Figure 2: **Evaluation framework with task examples of CUEBENCH.** Our benchmark advances the evaluation of five challenging context-aware VAU tasks in a unified generative manner, by prompting the generative VLMs with videos and task-related problems. The VLMs are required to respond accordingly in a JSON-style format rather than free-texts. This enables accurate evaluation of various tasks for generative VLMs by checking the answers with ground-truths.

triplets that suggest an anomaly or a normality. **(4)** *Context-Aware **Anomaly** Detection.* Automatically detect and localize all temporal clips that show any anomalies by ascertaining the contexts underlying the occurrences.

What If. (5) *Context-Aware Anticipation.* Infer the subsequent normalities or anomalies by reasoning the context triplets, based on the observed video clips.

2.3 Evaluation Framework

Figure 2 presents the evaluation of five challenging context-aware VAU tasks in a unified generative manner.

Problem Formulation. Given a video input $\mathcal{V}$ along with the problem $\mathcal{T}_p$ and format prompt $\mathcal{T}_f$ *w.r.t.* task $\mathcal{T}$, we prompt generative VLMs to output the answer lists ($\mathcal{O} = [o_1, \ldots, o_r]$) in a JSON format. According to $\mathcal{T}_p$ and $\mathcal{T}_f$, the model π can generate different task-specific outputs as:

$$\{\mathcal{O}, \mathcal{R}\} \text{ or } \{\mathcal{O}\} = \pi(\mathcal{V}, \mathcal{T}_p, \mathcal{T}_f^K, \mathcal{T}_f^V), \tag{1}$$

where $\mathcal{R}$ represents the response of the reasoning process, $\mathcal{T}_f^K$ and $\mathcal{T}_f^V$ specify the required task-specific key names and value types respectively, *e.g.*, for bottom-up context-aware anomaly recognition, $\mathcal{T}_f^K$ = (event, scene, attribute, anomaly), $\mathcal{T}_f^V = \langle E, S, A, N \rangle$. Each element $o_i = \{o_i^K : o_i^V\}$ in $\mathcal{O}$ denotes a key-value pair, and the key bag and value content are formulated as $\mathcal{O}^K = \{o_i^K\}_{i=1}^r$ and $\mathcal{O}^V = \{o_i^V\}_{i=1}^r$, respectively. This enables us to accurately probe various tasks by checking the VLM's output $\mathcal{O}$ with ground-truths $\mathcal{G} = [\{g_j^K : g_j^V\}_{j=1}^t]$ via a tailored evaluation metric suite to capture both structure alignment and task-related content quality, avoiding the bias scoring of LLMs.

Evaluation Metrics. To assess structure alignment, we design a structure-based F1 score which calculates binary matching between the output and ground-truth key bags $\mathcal{O}^K$ and $\mathcal{G}^K$ in the key space K:

$$S^K = \frac{2\left|\mathcal{O}^K \cap \mathcal{G}^K\right|}{2\left|\mathcal{O}^K \cap \mathcal{G}^K\right| + \left|\mathcal{O}^K \setminus \mathcal{G}^K\right| + \left|\mathcal{G}^K \setminus \mathcal{O}^K\right|}, \tag{2}$$

For content quality evaluation of "`What`" tasks, we first compute semantic embeddings (Devlin et al. 2019) from value content of both output ($\mathcal{O}^V$) and ground-truth ($\mathcal{G}^V$), denoted as $\mathcal{O}^U$ and $\mathcal{G}^U$ in the Euclidean space U. We then construct a semantic matching matrix $\mathcal{M} \in \mathbb{R}^{r \times t}$, where each element $m_{i,j}$ is a binary variable indicating whether $o_i^U \in \mathcal{O}^U$ and $g_j^U \in \mathcal{G}^U$ are matched using Hungarian algorithm (Kuhn 1955) based on cosine similarity. Thus, the semantic score S^U is defined as:

$$S^U = \frac{1}{r \cdot t} \sum_i \sum_j m_{i,j} \cdot \cos(o_i^U, g_j^U). \tag{3}$$

Given that semantic similarity can be overly lenient to hallucinated answers and often fails to reflect task alignment

Method	Event		Scene		Attr.		Anomaly (TD)			Anomaly (BU)			Ground		Detection		Anticipation		
	K	U	K	U	K	U	K	U	H	K	U	H	K	T	K	T	K	U	H
Commercial VLMs																			
Gemini-1.5-flash	60.1	38.4	84.8	59.9	29.4	19.6	45.1	39.3	3.1	57.3	38.6	3.4	54.4	20.7	51.7	21.7	61.3	18.9	**1.7**
Qwen-VL-Plus	39.2	14.4	63.9	36.9	24.4	4.8	31.1	22.9	1.3	27.8	13.8	0.4	61.1	17.4	33.0	7.5	48.2	2.1	0.0
Open-source VLMs																			
Qwen2.5-VL-3B	58.5	35.5	67.4	41.7	55.2	38.3	53.8	33.8	1.5	62.7	30.1	2.1	44.1	17.7	63.4	23.2	67.0	3.9	0.4
Qwen2.5-VL-7B	44.4	19.7	67.6	41.4	58.5	37.7	16.8	10.1	0.7	26.7	16.3	1.8	46.7	17.1	27.3	6.7	80.1	6.7	0.0
InternVideo-2.5	21.9	12.6	18.8	13.4	9.7	6.1	1.1	1.1	0.1	29.7	16.2	1.2	18.0	1.7	7.9	1.0	0.0	0.0	0.0
Video-ChatGPT	22.2	12.8	17.7	16.4	11.4	5.2	1.5	2.0	0.1	25.8	14.3	1.1	19.0	1.8	7.4	0.9	0.0	0.0	0.0
Video-LLaVA	29.3	13.5	26.9	17.2	13.1	9.0	17.5	13.8	0.3	23.2	15.2	1.2	23.0	3.6	7.1	1.2	0.0	0.0	0.0
Open-source R1 VLMs																			
Open-R1-Video	52.8	30.5	69.1	49.3	48.1	32.1	17.8	13.9	0.8	51.2	21.0	1.8	32.9	6.2	4.7	0.9	68.3	4.0	0.0
Video-R1	25.2	9.5	14.0	1.8	47.7	25.2	52.4	35.3	1.2	27.2	6.9	0.2	38.4	23.0	71.8	19.2	27.6	0.0	0.0
Video-Chat-R1	64.9	33.1	86.1	58.0	67.3	45.2	22.9	14.2	0.5	46.9	25.3	1.6	61.8	20.4	35.9	9.3	81.1	11.6	0.0
CUE-R1	**83.7**	**73.2**	**96.7**	**82.3**	**81.3**	**68.1**	**71.6**	**67.7**	**7.7**	**81.7**	**61.3**	**13.6**	**83.8**	**35.9**	**82.4**	**35.2**	**80.7**	**43.7**	0.6

Table 2: **Unified Evaluation on CUEBENCH.** We comprehensively gauge 11 VLMs, including 10 state-of-the-art VLMs and our CUE-R1 in the unified evaluation framework. "K", "U", "H" and "T" refer to the structure, semantic, hierarchy and temporal IoU scores, respectively. "TD" and "BU" denote top-down and bottom-up anomaly recognition, respectively.

accurately in anomaly understanding, we propose a novel hierarchy score. This metric leverages the event-centric hierarchy taxonomy H to better assess human-aligned performance for event-related tasks. Unlike the semantic score, we retrieve the most likely leaf nodes of o_i^U within H as its proxy (anomaly or normality) $\hat{o}_i^H$, based on their semantic similarities, and g_j^V can be reflected to g_j^H directly. After that, the hierarchy distance $d_{i,j}^H$ of each paired proxy and ground truth $(\hat{o}_i^H, g_j^H)$ is computed and then normalized as the final hierarchy score:

$$S^H = \frac{1}{r \cdot t} \sum_i \sum_j m_{i,j} \left(1 - \frac{d_{i,j}^H}{d_{\max}^H}\right) \cdot \mathbb{I}\left(d_{i,j}^H \leq \tau \cdot d_{\max}^H\right), \tag{4}$$

where $d_{\max}^H$ is the maximum depth of H and τ is the threshold for valid hierarchy alignment. For content quality evaluation of "`When`" tasks, we adopt the temporal IoU metrics as the temporal score S^{TIoU}.

3 Method: CUE-R1

To facilitate the comprehensive integration of context-aware capability *w.r.t.* various tasks into the training process, we develop CUE-R1 in a unified generative pipeline, based on reinforcement learning (RL) with GRPO algorithm (Shao et al. 2024). Following the rule-based reward paradigm of Open-R1 (Guo et al. 2025), our RL setup requires reward signals that are both reliable and precise. To ensure this, the training data is centered around tasks with clearly verifiable outputs, structured in a JSON format. This enables accurate reward computation using simple rules as mentioned in Section 2.3, thereby promoting stable and effective RFT (Liu et al. 2025; Shen et al. 2025). Our rule-based accuracy reward seamlessly aligns the policy model π_θ with task-specific evaluation preferences, enhancing the model's context-aware anomaly understanding capabilities. It serves as a verification function that checks for ideal matches between output and ground-truth answers as:

$$R_{\text{acc}} = R^K + \begin{cases} R^{\text{TIoU}}, & \text{if } \mathcal{T}_f^V = \langle T \rangle, \\ \lambda R^U + (1-\lambda) R^H, & \text{if } \mathcal{T}_f^V = \langle E, \cdot \rangle, \\ R^U, & \text{otherwise.} \end{cases} \tag{5}$$

Here, the struct reward R^K, semantic reward R^U and temporal reward R^{TIoU} are derived from S^K, S^U and S^{TIoU}, respectively, and λ controls the balance between semantic and hierarchy rewards for event-related tasks. To provide smoother hierarchy-refined guidance, we modify the hierarchy score S^H by discarding the thresholding term and redefine the hierarchy reward as:

$$R^H = \frac{1}{r \cdot t} \sum_i \sum_j m_{i,j} \cdot \left(1 - \frac{d_{i,j}^H}{d_{\max}^H}\right). \tag{6}$$

The overall reward used in CUE-R1 is composed of a format reward and a accuracy reward:

$$R = R_{\text{format}} + R_{\text{acc}}, \tag{7}$$

where $R_{\text{format}} = 1$ if the response contains both <think> and <answer> HTML tags, otherwise $R_{\text{format}} = 0$.

Given video and prompt inputs, the policy model π_θ generates a group of responses containing both reasoning processes and final answers. Each response is passed through the overall verifiable reward function (R) *w.r.t.* different context-aware VAU tasks to compute the reward. The advantage of each response (A_i) is then evaluated and used to update π_θ, along with the KL-regularization from the reference model for the training stability:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\{\mathcal{O}_i\}_{i=1}^N \sim \pi_{\theta_{\text{old}}}(\mathcal{O}|q)} \frac{1}{N} \sum_{i=1}^N \Big(\\ \min\left(s \cdot A_i,\ \text{clip}(s, 1-\epsilon, 1+\epsilon) \cdot A_i\right) \\ - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \Big), \end{aligned} \tag{8}$$

where $s = \frac{\pi_\theta(\mathcal{O}_i|q)}{\pi_{\theta_{\text{old}}}(\mathcal{O}_i|q)}$, ϵ and β are the hyperparameters.

Task	Metric	Method	Result (%)
Event Recognition	Top-1 / Top-5 Hierarchy Score	CLIP	35.13 / 73.51
		Open-VCLIP	34.84 / 71.72
		FROSTER	35.52 / 76.34
		Open-MeDe	37.03 / 76.42
		CUE-R1	**57.26 / 84.20**
Temporal Grounding	TIoU	UniVTG	17.65
		TimeChat	19.21
		UniTime	21.43
		CUE-R1	**35.94**
Anomaly Recognition	Top-1 / Top-5 Hierarchy Score	CLIP	10.72 / 29.89
		Open-MeDe	12.01 / 32.83
		VadCLIP	21.21 / 42.33
		Holmes-VAU	29.72 / 53.12
		CUE-R1	**32.18 / 69.71**
Anomaly Detection	TIoU	CLIP	13.28
		VadCLIP	17.91
		Holmes-VAU	29.38
		CUE-R1	**35.17**

Table 3: **Separate Evaluation on CUEBENCH.** We assess various specialized VLMs on four video understanding tasks following standard practices.

4 Experiment

4.1 Implementation Details

We apply CUE-R1 to the Qwen2.5-VL-3B model (Bai et al. 2025), performing one epoch of supervised fine-tuning (SFT) followed by another epoch of reinforcement fine-tuning (RFT) on the CUEBENCH training set, using a learning rate of $1.0e^{-6}$. To ensure training efficiency, we cap the number of video frames at 64, with each frame processed at a resolution of $128 \times 28 \times 28$. For inference, we boost the frame resolution to $256 \times 28 \times 28$ and increase the number of frames to 128 to improve performance.

4.2 Unified Evaluation on Generative VLMs

Table 2 presents a comprehensive quantitative evaluation of 10 state-of-the-art generative VLMs and our proposed CUE-R1 on CUEBENCH, including 2 proprietary VLMs (Gemini-1.5-Flash (Team et al. 2024), Qwen-VL-Plus (Bai et al. 2025)) and 8 popular open-source models, under the proposed unified evaluation framework. From the results, we can summarize the observations: **1) CUE-R1 *vs.* Others.** Our CUE-R1 delivers a significant performance advantage across nearly all metrics of five distinct tasks, outperforming both commercial and open-source baselines. This demonstrates its effectiveness as a universal solution with strong structural alignment, high-quality semantic content and superior temporal comprehension. Notably, in complex reasoning tasks like context-aware anomaly recognition, CUE-R1 achieves semantic/hierarchy scores (%) of $67.7/7.7$ and $61.3/13.6$ in top-down and bottom-up manners respectively, highlighting its enhanced human-aligned reasoning capabilities within event hierarchies. **2) Proprietary *vs.* Open-source VLMs.** Compared with existing open-source VLMs, the proprietary Gemini-1.5-Flash exhibits impressive context-aware reasoning capabilities, while Qwen-VL-Plus shows marginal performance in both structural alignment and context awareness. **3) R1 *vs.* Others.** Note that Qwen2.5-VL-3B/7B (Bai et al. 2025) both achieve more promising performance across various evaluations than previous R1s. Despite Video-R1 (Feng et al. 2025) falling short in most cases, other R1-style models *i.e.*, Open-R1-Video (Wang and Peng 2025) and Video-Chat-R1 (Li et al. 2025b) achieve better performance than other open-source baselines like InternVideo-2.5 (Wang et al. 2025), Video-ChatGPT (Maaz et al. 2023) and Video-LLaVA (Lin et al. 2023a), highlighting their strong video reasoning capabilities. However, from the results, there remains considerable room for a satisfied unified solution in context-aware VAU.

4.3 Separate Evaluation on Specialized VLMs

We further conduct separate task-specific evaluations on CUEBENCH following popular protocols (See Appendix) to assess various specialized VLMs on four core context-aware VAU tasks, as shown in Table 3. Specifically, in event recognition, Open-MeDe (Yu et al. 2025c) achieves the strongest generalization among discriminative VLMs (Radford et al. 2021; Weng et al. 2023; Huang et al. 2024b) designed for open-vocabulary action recognition. In temporal grounding, UniTime (Li et al. 2025c) stands out as the top performer among prior methods (Lin et al. 2023b; Huang et al. 2024a; Ren et al. 2024), benefiting from elaborative training across videos of diverse contexts. For anomaly recognition that requires context-aware capabilities, our evaluation shows that VAU methods *i.e.*, VadCLIP (Wu et al. 2024b) and Holmes-VAU (Zhang et al. 2025c) significantly outperform general action recognition approaches. Holmes-VAU records strong performance for both anomaly recognition and detection, indicating its superior anomaly understanding capabilities. Overall, CUE-R1 outperforms both discriminative and generative specialized VLMs across tasks. Despite the strengths of task-specific models, their performance still manifests clear limitations, particularly in semantic alignment and anomaly reasoning, underscoring the advantage of our unified and context-aware generative approach.

4.4 Ablation Study

To assess the contributions of SFT and RFT strategies, we conduct an ablation study by performing two variants on Qwen2.5-VL-3B model. The results in Table 4 clearly demonstrate the effectiveness of both strategies in our training pipeline. Compared to the baseline, SFT yields substantial gains especially in semantic scores, demonstrating its effectiveness in enhancing alignment with structured answers and improving content consistency. While RFT alone brings more gains over SFT on struct scores, the hierarchy scores improve only marginally or stagnate, suggesting that reward signals based solely on task performance could be insufficient to capture fine-grained semantic relations or hierarchical distinctions. By combining SFT and RFT sequentially, CUE-R1 achieves the best overall performance, serving as a robust and context-aware generative VLM for comprehensive VAU. The large improvement in hierarchy scores, especially for complex tasks like anomaly recognition, vali-

Method	Event		Scene		Attr.		Anomaly (TD)			Anomaly (BU)			Ground		Detection		Anticipation		
	K	U	K	U	K	U	K	U	H	K	U	H	K	T	K	T	K	U	H
Baseline	58.5	35.5	67.4	41.4	55.4	38.3	53.8	33.8	1.5	62.7	30.1	2.1	44.1	17.7	63.4	23.2	67.0	3.9	0.4
+SFT	82.4	73.0	95.9	81.5	78.9	65.6	66.3	62.5	7.1	80.9	60.8	8.1	55.7	34.6	51.8	**39.0**	80.6	39.1	0.0
+RFT	79.6	64.7	96.6	82.3	80.8	65.0	72.0	67.0	3.1	80.3	53.8	2.8	83.5	27.5	**83.0**	34.9	80.0	35.5	0.6
Ours	**83.7** ↑25.2	**73.2** ↑37.7	**96.7** ↑29.3	**82.3** ↑40.9	**81.3** ↑25.9	**68.1** ↑29.8	**71.6** ↑17.8	**67.7** ↑33.9	**7.7** ↑6.2	**81.7** ↑19.0	**61.3** ↑31.2	**13.6** ↑11.5	**83.8** ↑39.7	**35.9** ↑18.2	82.4 ↑19.0	35.2 ↑12.0	**80.7** ↑13.7	**43.7** ↑39.8	**0.6** ↑0.2

Table 4: Ablation of three fine-tuning configurations based on Qwen2.5-VL-3B (Baseline). We maintain the same cycle length of two epochs for different training settings to ensure a fair comparison. Improvements over baseline are highlighted in green.

Figure 3: **Case Study.** Comparisons with Qwen2.5-VL-3B and CUE-R1 on context-aware anomaly recognition and detection.

dates the benefit of incorporating human-aligned hierarchical feedback in RFT. The comparison highlights that SFT provides strong structural and semantic grounding, while RFT complements it by refining task alignment.

4.5 Case Study

Figure 3 presents qualitative and quantitative comparisons between Qwen2.5-VL-3B and CUE-R1 on two representative VAU tasks under the unified evaluation paradigm. (a) For anomaly recognition, Qwen2.5-VL-3B fails to recognize the severity and specificity of the anomalies. It correctly identifies the scene and mentions *jumping over lanes*, yet misrepresenting dangerous maneuvers as generic traffic behavior. CUE-R1, in contrast, identifies two well-grounded events: *car driving* and *car crash*, associating them with meaningful attributes like *double jump* and *accident*. This reflects an accurate contextual and semantic interpretation of the anomaly. It scores higher on hierarchy metrics, reflecting better alignment within the event hierarchy. (b) For anomaly detection, Qwen2.5-VL-3B offers surface-level reasoning process: "*a peaceful protest or demonstration*", lacking the anomaly relevant details (e.g., vandalism) and struggles with hallucination and poor anomaly localization. Conversely, CUE-R1 delivers contextually grounded, semantically rich, and temporally precise predictions, highlighting its superior performance for anomaly understanding.

5 Conclusion

This paper presents CUEBENCH, the first large-scale benchmark for evaluating the context-aware video anomaly understanding capabilities of VLMs in a unified framework. We establish a comprehensive event-centric hierarchical taxonomy with absolute and conditional anomaly events and diverse context-aware anomalies and normalities. Our extensive evaluation highlights significant performance gaps remaining among existing state-of-the-art VLMs. Building upon this, we propose CUE-R1, an R1-style method that outperforms the leading VLMs by a notable margin on CUEBENCH. This work not only provides a solid foundation for developing unified generative VLMs, but also serves as a challenging benchmark for VAU in real-world.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62376217, 62576279, 62301434), Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC001), and the Joint Research Project for Meteorological Capacity Improvement (Grants 24NLTSZ003). Finally, we would like to express our sincere appreciation to Yujing Li, Shuo Qin, Tong Lu, Yudong Chen, Mingxuan Li, Jianfeng Wu, Qirui Wang and other contributors for their invaluable contributions to the data collection and annotation. Their dedication and meticulous efforts have been essential to the success of this work.

References

Acsintoae, A.; Florescu, A.; Georgescu, M.-I.; Mare, T.; Sumedrea, P.; Ionescu, R. T.; Khan, F. S.; and Shah, M. 2022. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20143–20153.

Adam, A.; Rivlin, E.; Shimshoni, I.; and Reinitz, D. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3): 555–560.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cai, R.; Zhang, H.; Liu, W.; Gao, S.; and Hao, Z. 2021. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 938–946.

Cao, C.; Lu, Y.; Wang, P.; and Zhang, Y. 2023. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20392–20401.

Cao, C.; Lu, Y.; and Zhang, Y. 2024. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Transactions on Image Processing*, 33: 1810–1825.

Cao, C.; Zhang, H.; Lu, Y.; Wang, P.; and Zhang, Y. 2025. Scene-Dependent Prediction in Latent Space for Video Anomaly Detection and Anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 224–239.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Du, H.; Nan, G.; Qian, J.; Wu, W.; Deng, W.; Mu, H.; Chen, Z.; Mao, P.; Tao, X.; and Liu, J. 2024a. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *arXiv preprint arXiv:2412.07183*.

Du, H.; Zhang, S.; Xie, B.; Nan, G.; Zhang, J.; Xu, J.; Liu, H.; Leng, S.; Liu, J.; Fan, H.; et al. 2024b. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18793–18803.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, C.; Wang, B.; Wen, J.; Liu, C.; Wang, W.; Shen, L.; and Cao, X. 2025. Vad-R1: Towards Video Anomaly Reasoning via Perception-to-Cognition Chain-of-Thought. *arXiv preprint arXiv:2505.19877*.

Huang, D.-A.; Liao, S.; Radhakrishnan, S.; Yin, H.; Molchanov, P.; Yu, Z.; and Kautz, J. 2024a. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, 202–218. Springer.

Huang, X.; Zhou, H.; Yao, K.; and Han, K. 2024b. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. *arXiv preprint arXiv:2402.03241*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Li, F.; Liu, W.; Chen, J.; Zhang, R.; Wang, Y.; Zhong, X.; and Wang, Z. 2025a. Anomize: Better Open Vocabulary Video Anomaly Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29203–29212.

Li, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Li, Z.; Di, S.; Zhai, Z.; Huang, W.; Wang, Y.; and Xie, W. 2025c. Universal Video Temporal Grounding with Generative Multimodal Large Language Models. *arXiv preprint arXiv:2506.18883*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023b. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2794–2804.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, 2720–2727.

Luo, W.; Liu, W.; and Gao, S. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, 341–349.

Ma, J.; Wang, J.; Luo, J.; Yu, P.; and Zhou, G. 2025. Sherlock: Towards Multi-scene Video Abnormal Event Extraction and Localization via a Global-local Spatial-sensitive LLM. In *Proceedings of the ACM on Web Conference 2025*, 4004–4013.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

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

Ramachandra, B.; and Jones, M. 2020. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2569–2578.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Ristea, N.-C.; Croitoru, F.-A.; Ionescu, R. T.; Popescu, M.; Khan, F. S.; Shah, M.; et al. 2024. Self-distilled masked auto-encoders are efficient video anomaly detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15984–15995.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; et al. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488.

Tang, J.; Lu, H.; Wu, R.; Xu, X.; Ma, K.; Fang, C.; Guo, B.; Lu, J.; Chen, Q.; and Chen, Y. 2024. Hawk: Learning to understand open-world video anomalies. *Advances in Neural Information Processing Systems*, 37: 139751–139785.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, S.; and Miao, Z. 2010. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, 1220–1223. IEEE.

Wang, X.; and Peng, P. 2025. Open-R1-Video. https://github.com/Wang-Xiaodong1899/Open-R1-Video.

Wang, Y.; Li, X.; Yan, Z.; He, Y.; Yu, J.; Zeng, X.; Wang, C.; Ma, C.; Huang, H.; Gao, J.; et al. 2025. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Weng, Z.; Yang, X.; Li, A.; Wu, Z.; and Jiang, Y.-G. 2023. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International conference on machine learning*, 36978–36989. PMLR.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, 322–339. Springer.

Wu, P.; Zhou, X.; Pang, G.; Sun, Y.; Liu, J.; Wang, P.; and Zhang, Y. 2024a. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18297–18307.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2024b. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6074–6082.

Xu, J.; Lo, S.-Y.; Safaei, B.; Patel, V. M.; and Dwivedi, I. 2025. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20370–20382.

Yan, C.; Zhang, S.; Liu, Y.; Pang, G.; and Wang, W. 2023. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5527–5537.

Ye, M.; Liu, W.; and He, P. 2025. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8679–8688.

Yu, E.; Lin, K.; Zhao, L.; Wei, Y.; Zhu, Z.; Wei, H.; Sun, J.; Ge, Z.; Zhang, X.; Wang, J.; et al. 2025a. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*.

Yu, Y.; Cao, C.; Zhang, Y.; Lv, Q.; Min, L.; and Zhang, Y. 2025b. Building a multi-modal spatiotemporal expert for zero-shot action recognition with clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9689–9697.

Yu, Y.; Cao, C.; Zhang, Y.; and Zhang, Y. 2025c. Learning to Generalize without Bias for Open-Vocabulary Action Recognition. *arXiv preprint arXiv:2502.20158*.

Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; and Ricci, E. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, H.; Cao, C.; Lv, Q.; Min, L.; and Zhang, Y. 2025b. Autoregressive Denoising Score Matching is a Good Video Anomaly Detector. *arXiv preprint arXiv:2506.23282*.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Huang, X.; Gao, C.; Zhang, S.; Yu, L.; and Sang, N. 2025c. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13843–13853.

Zhou, S.; Shen, W.; Zeng, D.; Fang, M.; Wei, Y.; and Zhang, Z. 2016. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication*, 47: 358–368.

Zhu, L.; Wang, L.; Raj, A.; Gedeon, T.; and Chen, C. 2024. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems*, 37: 89943–89977.

Zhu, Y.; Bao, W.; and Yu, Q. 2022. Towards open set video anomaly detection. In *European Conference on Computer Vision*, 395–412. Springer.