SEEING ANOMALIES IN THE DARK: VIDEO ANOMALY UNDERSTANDING WITH LOW-LIGHT CONDITIONS FOR MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video anomaly understanding (VAU) aims to detect and classify anomalous events in videos, providing fine-grained explanations of their nature and context. However, existing works in this direction often assume well-lit scenes, paying limited attention to low-light scenarios where many real-world anomalies occur, *e.g.,* nighttime intrusions and blackout looting. To address this challenge, we introduce **VAULT**, a benchmark for low-light video anomaly understanding across various granularities. Specifically, VAULT contributes: **(1)** broad coverage of poorly illuminated anomalies with four illumination levels, covering 25 anomaly categories across 6 scene types, and **(2)** a unified evaluation suite comprising four tasks, *i.e.*, temporal grounding, anomaly classification, anomaly description, and video question answering (VQA). Building upon VAULT, we propose **EventVAULT**, an event-guided framework that combines RGB appearance cues with event-stream dynamics for low-light VAU. Instead of relying solely on RGB appearance, EventVAULT leverages event streams to capture transient dynamics and address three key challenges: *when* to acquire informative RGB evidence, *where* to focus within the scene, and *how* anomalous events evolve. Extensive experiments demonstrate that the performance of existing MLLMs deteriorates as illumination decreases, while EventVAULT consistently improves performance across all four tasks. Our dataset and findings jointly highlight the need for more robust VAU under extreme visual conditions. Code available at: `https://anonymous.4open.science/r/VAULT`.

1 INTRODUCTION

Detecting, localizing, and analysing anomalous events in complex real-world environments is a long-standing challenge in multimodal video learning. The goal of this task, *i.e.*, video anomaly understanding (VAU), is to provide an actionable interpretation of abnormal events, supporting reliable monitoring and informed decision-making with reduced manual inspection. Specifically, given video observations and task-specific queries, a model is expected to localize anomalies, identify their categories, describe how they unfold, and answer questions about the observed events. Ongoing advancements in VAU have expanded from recognition and temporal localization in conventional VAD (Zhou et al., 2023; Liu et al., 2024; Leng et al., 2024; Wu et al., 2024b) to causal reasoning, motion-aware understanding, and long-term interpretation with MLLMs (Du et al., 2024b; Tang et al., 2024; Zhang et al., 2025b), which collectively advance this paradigm toward a more comprehensive and interpretable understanding of anomalous events.

Despite these advances, RGB-based VAU remains vulnerable to degraded visual evidence in nighttime and poorly lit environments. To better understand this challenge, we analyze how low illumination affects both model performance and visual representations, yielding two findings.

Finding-1 (Robustness). *Low illumination and weak RGB motion evidence jointly challenge VAU.* As shown in Fig. 1 (left), low illumination remains challenging even when RGB motion is highly observable, with a 30.5% F1 gap between the highest and lowest-illumination groups. Even under low illumination, performance remains sensitive to RGB motion observability, with F1 dropping to 37.0% when motion cues are weak. These observations suggest that degraded appearance and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

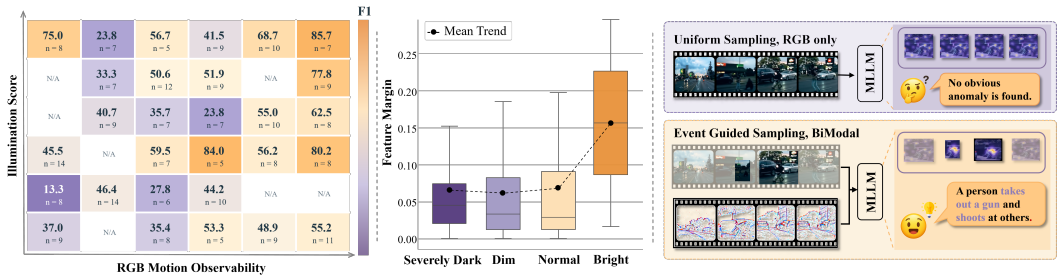

Figure 1: **Left**: F1 scores across groups defined by illumination level and RGB motion observability. Each populated cell reports its score and sample count n; N/A denotes inadequate samples. **Middle**: feature separability under different illumination conditions. Box plots show cosine-distance margins between normal and anomalous frame features. **Right**: an illustrative comparison between uniformly sampled RGB-only observations and event-guided RGB–event inputs.

limited dynamic evidence jointly hinder anomaly understanding, motivating a closer examination of how low illumination affects the visual representations used for semantic reasoning.

Finding-2 (Separability). *Normal and anomalous representations become less distinguishable under low illumination.* Intuitively, anomalies often occupy only a small fraction of frames, making the visual separability between normal and anomalous frames crucial for reliable VAU. To probe this, we group the data into four illumination levels, *i.e.*, `severely dark`, `dim`, `normal`, and `bright`, and extract frame features with Qwen3-VL-8B-Instruct (Bai et al., 2025a) to measure the inter-class cosine distance between normal and anomalous representations. As shown in Fig. 1(middle), bright scenes exhibit a substantially larger feature margin, whereas severely dark, dim, and normal scenes yield smaller and relatively comparable margins. This indicates that degraded illumination compresses the representation gap between normal and anomalous observations, making low-light anomalies harder to distinguish from their surrounding context.

Based on the observations outlined above, we introduce **VAULT**, the first-ever comprehensive low-light video anomaly understanding benchmark crafted for multimodal LLMs. VAULT comprises 1,776 real-world videos totaling 41.37 hours, covering 25 anomaly categories across six scene types and diverse indoor/outdoor settings. Each video sample is categorized into four illumination levels, enabling controlled evaluation across lighting conditions. It further supports four complementary tasks: *temporal grounding, anomaly classification, anomaly description, and video question answering*, allowing systematic evaluation of both semantic and temporal anomaly understanding under varying illumination. Moreover, our preceding observations reveal that low illumination severely degrades RGB representations, whereas event signals remain sensitive to intensity changes and preserve informative motion patterns, as illustrated in Fig. 1(right). This complementarity motivates us to exploit event streams as dynamic cues for low-light VAU. To this end, we propose **EventVAULT**, an event-guided framework that complements RGB appearance cues with event-based dynamic evidence. To identify *when* and *where* anomaly-relevant evidence emerges, Temporal Evidence Allocation (**TEA**) and Spatial Evidence Localization (**SEL**) leverage event activity to allocate RGB observations across time and space. To further capture *how* anomalies evolve, Adaptive Event Evidence Representation (**AEER**) extracts dynamic variations from event streams. The resulting RGB and event evidence is then jointly processed by the MLLM for fine-grained anomaly understanding.

To the best of our knowledge, VAULT is the first benchmark dedicated to evaluating video anomaly understanding under low-light conditions. Extensive experiments on VAULT reveal that existing MLLMs struggle with temporal localization and anomaly classification, and these difficulties intensify as illumination decreases. Moreover, the proposed EventVAULT consistently improves performance across all four tasks, with larger gains under lower illumination—improving temporal grounding mIoU over Qwen3-VL-8B by **7.00** overall and by **7.96** percentage points on the low-light subset.

Our main contributions are threefold: **1)** We introduce **VAULT**, the first benchmark dedicated to low-light video anomaly understanding. It spans 25 anomaly categories, six scene types, and four illumination levels, with four complementary tasks. **2)** We propose **EventVAULT**, a model-agnostic framework that leverages event dynamics for temporal and spatial evidence allocation and complementary dynamic representation, improving low-light anomaly understanding across diverse models. **3)** We perform comprehensive evaluation using a wide spectrum of multimodal LLMs, and the established results provide valuable insights to the research community for future improvement.

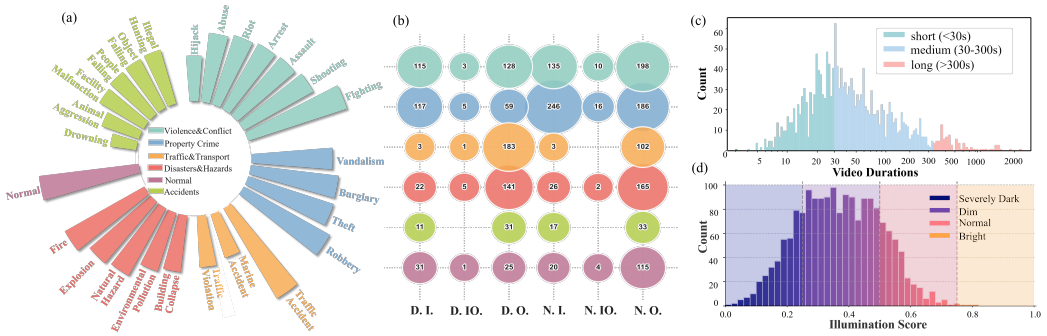

Figure 2: **Statistics of the proposed VAULT dataset.** (a) Broad coverage of diverse anomaly categories. (b) Diverse scene distribution across day/night and indoor/outdoor settings. (c) Large temporal diversity spanning short- to long-duration videos. (d) Strong low-light characteristics with coverage across multiple illumination levels.

2 Related Work

Existing video anomaly research has progressed from conventional detection (Sultani et al., 2018; Wu et al., 2020) toward fine-grained semantic understanding (Du et al., 2024b; Tang et al., 2024; Zhang et al., 2025b; Zhu et al., 2025). However, current VAU benchmarks and methods are largely developed under general visual conditions, while existing low-light studies remain mainly focused on anomaly detection (Han et al., 2026; Mo et al., 2026). Meanwhile, event-based sensing has shown promising complementary value for low-light perception and VAD (Li et al., 2024; Kim & Yoon, 2025; Wu et al., 2026). Our work bridges these directions by introducing a low-light VAU benchmark and exploiting event-derived dynamics for fine-grained RGB–event anomaly understanding. Further discussion of related benchmarks and methods is provided in Appendix B.

3 The VAULT Benchmark

3.1 Task Definition, Data Collection and Annotation

To support comprehensive and fine-grained video anomaly understanding under challenging illumination conditions, we formulate four complementary tasks: *temporal anomaly localization*, *multi-label anomaly classification*, *anomaly description*, and *anomaly VQA*. These tasks jointly evaluate whether a model can determine *when* an anomalous event occurs, recognize *what* type of anomaly is present, describe *how* the event progresses in natural language, and answer fine-grained questions about its visual content.

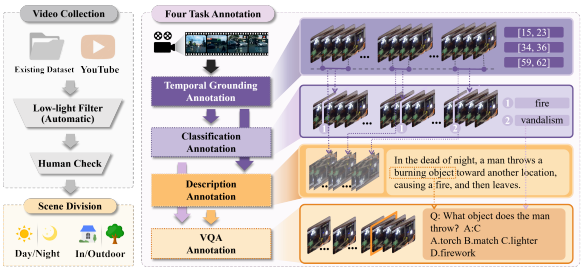

Figure 3: Overview of the VAULT collection and annotation pipeline. The rightmost panel illustrates the definitions of the four tasks in VAULT.

To obtain broad coverage of real-world anomalous events, we collect candidate videos from two complementary sources. First, we reuse videos from existing VAD and VAU datasets (Wu et al., 2020; Sultani et al., 2018; Du et al., 2024a; Zhu et al., 2024; Perez et al., 2019), which provide representative surveillance scenarios and commonly studied anomaly categories. Second, we retrieve additional videos from YouTube to enrich the dataset with more diverse scenes, illumination conditions, and anomaly categories. The collected candidates subsequently undergo automatic low-light screening and human verification before annotation.

To facilitate large-scale annotation and maintain a consistent annotation workflow, we further develop a dedicated annotation pipeline, as illustrated in Fig. 3. Each retained video is reviewed by human annotators and annotated for temporal intervals, anomaly categories, natural-language descriptions, and VQA samples. Detailed video selection criteria and annotation procedures are provided in Appendix A.

Table 1: **Comparison with existing benchmarks.** "VAD" and "VAU" denote video anomaly detection and video anomaly understanding, respectively. "LL" indicates whether low-light conditions are explicitly considered. "Nat." and "Syn." denote naturally captured and synthetically generated low-light data, respectively.

Dataset	Type	Scale		Low-light Design		Annotation / Task Coverage			
		#Videos	Hours	LL	Source	Anomaly Types	Temporal	Descript.	QA/VQA
CUHK Avenue (Lu et al., 2013)	VAD	37	0.5	✗	–	5	Frame	✗	✗
ShanghaiTech (Luo et al., 2017)	VAD	437	3.67	✗	–	11	Frame	✗	✗
UCF-Crime (Sultani et al., 2018)	VAD	1,900	128	✗	–	13	Frame	✗	✗
XD-Violence (Wu et al., 2020)	VAD	4,754	217	✗	–	6	Frame	✗	✗
UBnormal (Acsintoae et al., 2022)	VAD	543	2.2	✗	–	22	Frame	✗	✗
NWPU Campus (Cao et al., 2023)	VAD	547	16.29	✗	–	28	Frame	✗	✗
MSAD (Zhu et al., 2024)	VAD	720	4.14	✗	–	55	Frame	✗	✗
LVAD (Han et al., 2026)	VAD	37 / 437	–	✓	Syn.	Inherited	Frame	✗	✗
EWAD Benchmarks (Wu et al., 2026)	VAD	3,443	148.2	✗	–	13 / 1 / 22	Frame	✗	✗
CUVA (Du et al., 2024b)	VAU	1,000	32.46	✗	–	42	Segment	✓	✓
HIVAU-70K (Zhang et al., 2025b)	VAU	5,443	–	✗	–	–	Hierarchical	✓	✓
VALU (He et al., 2026)	VAU	1,019	–	✗	–	–	5-level	✓	✓
FineVAU (Pereira et al., 2026)	VAU	1,544	–	✗	–	Inherited	Event-level	✓	✗
VAU-Bench (Zhu et al., 2025)	VAU	4,602	169.1	✗	–	19	Time duration	✓	✓
VAULT (Ours)	**VAU**	**1,776**	**41.37**	✓	**Nat.**	**25 + Normal**	**Multi-segment**	✓	✓

3.2 Benchmark Overview

VAULT is designed to support fine-grained video anomaly understanding under diverse illumination conditions. Beyond determining whether an anomaly occurs, it evaluates *what* happens, *when* it happens, and whether detailed event information remains understandable as illumination deteriorates. The benchmark contains 1,776 real-world videos totaling 41.37 hours, covering 25 anomalous event categories and one normal category, and is partitioned into 1,115/356/305 for training, validation and test sets respectively. To support comprehensive evaluation, VAULT provides multi-label anomaly categories, open-ended descriptions, multi-segment temporal annotations, and video-grounded QAs. In total, the dataset contains 1,776 descriptions, 2,159 category labels, 1,928 anomalous temporal intervals, and 5,328 VQA samples. In addition, videos are organized across multiple scene types and four illumination levels, enabling fine-grained analysis of anomaly understanding under varying visual conditions. Detailed dataset statistics are presented in Fig. 2.

3.3 Comparison with Existing Benchmarks

Table 1 compares VAULT with representative VAD and VAU benchmarks in terms of low-light design and annotation/task coverage. Traditional VAD datasets mainly support anomaly recognition and temporal localization, with limited supervision for semantic interpretation. Existing low-light benchmarks further consider degraded illumination, but are primarily detection-oriented and often rely on synthetically generated low-light data. In contrast, recent VAU benchmarks provide richer semantic supervision, yet illumination is rarely treated as an explicit dimension of benchmark construction or evaluation. VAULT bridges these gaps as the first illumination-aware VAU benchmark built from naturally captured data. It supports multiple fine-grained tasks, while its illumination-level annotations enable systematic evaluation of model robustness as visual conditions deteriorate. This design allows VAULT to evaluate not only anomaly understanding itself, but also how reliably such understanding can be maintained under challenging illumination.

4 Methodology

4.1 Overall Pipeline

As illustrated in Fig. 4, given an RGB video as input, EventVAULT first employs an Event Stream Generator to synthesize event stream. The generated events are then processed by TEA and SEL to guide informative RGB frame sampling in the temporal and spatial domains. Subsequently, AEER extracts dynamic cues from the event stream and converts them into interpretable, RGB-compatible event visualizations. Finally, the sampled RGB frames, transformed event images, and textual prompts are jointly fed into a pretrained MLLM, which is fine-tuned with LoRA (Hu et al., 2022)to perform temporal grounding, classification , description and VQA tasks.

4.2 Event-guided Evidence Modeling

4.2.1 Temporal Evidence Allocation

Uniform video sampling treats all temporal locations equally and may overlook short-lived anomalous evidence. To allocate the limited visual budget toward dynamically informative moments, we introduce **TEA** to estimate event activity over the temporal bins.

Given an RGB video $\mathcal{V} = \{I_n\}_{n=0}^{N-1}$ with duration T and its temporally aligned event stream $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{M}$, where (x_i, y_i), t_i, and $p_i \in \{-1, +1\}$ denote the event location, timestamp, and polarity, respectively. Specifically, the video is divided into B temporal bins and the event plane into a coarse spatial grid, producing an event-count tensor

$$A_{b,h,w} = \sum_{i=1}^{M} \mathbb{I}(t_i \in b)\mathbb{I}((x_i, y_i) \in (h, w)). \tag{1}$$

Spatial Aggregation. We first summarize the shared event-count tensor along its spatial dimensions:

$$C_b = \sum_{h,w} A_{b,h,w} \tag{2}$$

The resulting sequence $\{C_b\}_{b=0}^{B-1}$ describes how event activity varies over time.

Motion Persistence Filtering. To attenuate isolated event-count fluctuations while retaining temporally coherent motion responses, we apply logarithmic compression to C_b and smooth the resulting activity over neighboring bins:

$$\begin{aligned} \tilde{a}_b &= \log(1 + C_b), \\ a_b &= 0.25\tilde{a}_{b-1} + 0.50\tilde{a}_b + 0.25\tilde{a}_{b+1}. \end{aligned} \tag{3}$$

This local filtering reduces the prominence of isolated spikes while preserving activity supported by neighboring temporal bins.

Motion Saliency Calibration. To emphasize relatively salient dynamics despite clip-specific event-rate variations, we calibrate the smoothed activity using its temporal percentiles:

$$q_b = \mathrm{clip}\left(\frac{a_b - P_{20}}{P_{90} - P_{20}}, 0, 1\right), \qquad \kappa = \frac{P_{90} - P_{20}}{\max(|P_{90}|, \epsilon)}, \tag{4}$$

where P_{20} and P_{90} denote the 20-th and 90-th percentiles of $\{a_b\}_{b=0}^{B-1}$ while $\epsilon > 0$ ensures numerical stability.

To reduce false activity caused by camera shake and shot transitions, we mildly modulate q_b using the normalized spatial concentration f_b, derived from the fraction of events contained in the most active 10% of spatial cells:

$$s_b = q_b(0.8 + 0.2f_b), \tag{5}$$

4.2.2 Spatial Evidence Localization

Temporal allocation alone cannot resolve spatially small anomalies, particularly under low illumination where the anomalous subject may occupy only a limited portion of the frame. We thus introduce **SEL** to further localize event-responsive spatial regions for each TEA-focused observation.

Motion Persistence Aggregation. To stabilize short-term motion responses while retaining their spatial layout, SEL aggregates local event maps over neighboring temporal bins. For each TEA-focused timestamp τ_k, we locate its bin $b_k = \mathrm{clip}(\lfloor \tau_k B/T \rfloor, 0, B-1)$ and compute

$$\bar{A}_{h,w}^{(k)} = \sum_{b=\max(0,b_k-1)}^{\min(B-1,b_k+1)} A_{b,h,w}. \tag{6}$$

Dominant Motion Selection. To preserve dominant motion regions while suppressing scattered background responses, we localize dynamic evidence by selecting the most active cells that collectively account for a prescribed fraction of the local event mass. Let $H_g \times W_g$ denote the spatial grid size, and sort the counts in $\bar{A}^{(k)}$ as $a_{(1)}^{(k)} \geq \cdots \geq a_{(H_g W_g)}^{(k)}$. For a nonempty map, the selected set is

$$\begin{aligned} K_k^{80} &= \min\left\{ K : \sum_{j=1}^{K} a_{(j)}^{(k)} \geq \mu \sum_{j=1}^{H_g W_g} a_{(j)}^{(k)} \right\}, \\ \Omega_k^{80} &= \{(h_{(j)}, w_{(j)})\}_{j=1}^{K_k^{80}}, \qquad \mu = 0.80. \end{aligned} \tag{7}$$

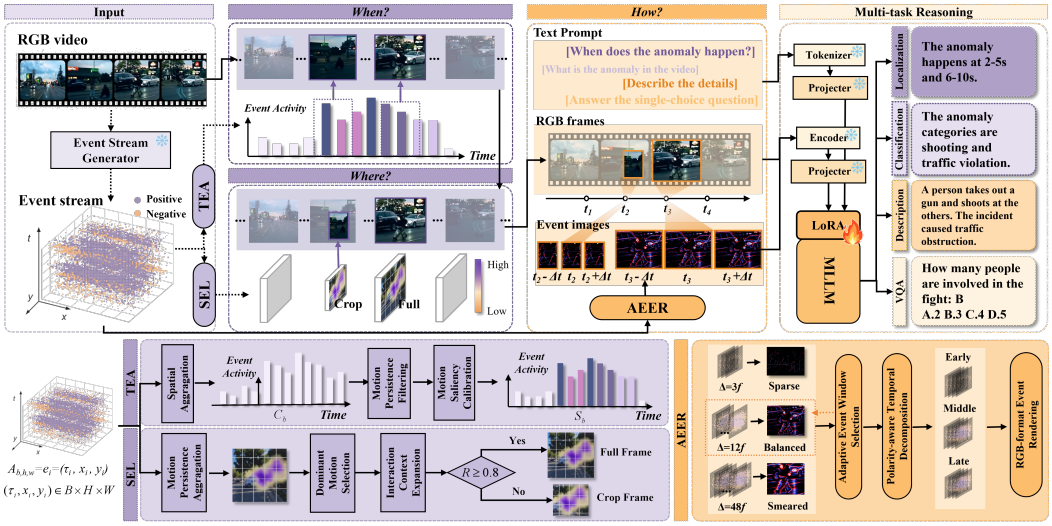

Figure 4: **Overview of EventVAULT.** TEA and SEL determine *when* and *where* informative anomaly evidence should be observed, while AEER captures complementary event dynamics for joint RGB–event reasoning across four VAU tasks.

Here, $(h_{(j)}, w_{(j)})$ is the cell associated with the j-th largest count. This mass-based selection reduces the influence of low-count, scattered responses on the initial localization.

Interaction Context Expansion. Cropping tightly around the selected cells may exclude nearby subjects or objects needed to interpret an interaction. We construct their enclosing normalized box $R_k = (l_k, u_k, r_k, d_k)$ and expand it to retain surrounding context. If the expanded box R_k occupies less than 80% of the frame, the frame is cropped to R_k; otherwise, the full frame is retained.

4.2.3 Adaptive Event Evidence Representation

TEA and SEL use event activity to organize RGB evidence, but the underlying event stream itself contains additional information about the short-term evolution of local motion changes. To expose this information to an RGB-pretrained MLLM, we propose **AEER**, which converts asynchronous event streams into aligned image-form visual evidence.

Adaptive Event Window Selection. To avoid introducing redundant event evidence and concentrate the visual budget on the most informative dynamics, AEER ranks the materialized RGB observations using their TEA evidence scores and retains at most three representative anchors $(\hat{t}_k, \hat{R}_k)$, where $\hat{R}_k$ is the aligned SEL region or full frame. Since event activity can vary substantially across different moments, a fixed temporal window may either contain insufficient evidence or introduce excessive irrelevant responses. We therefore adapt the event history to the local activity for each anchor. Let f denote the RGB frame rate and $\mathcal{D} = \{3, 6, 12, 24, 48\}$ the candidate window lengths in frame periods. Each candidate contains events preceding $\hat{t}_k$ within its aligned region. We choose the shortest candidate with sufficient density and nonempty early, middle, and late stages, whose counts $N_{k,s}(d)$ are defined below:

$$\begin{aligned} d_k^* &= \min\left\{d \in \mathcal{D} : \gamma_k(d) \geq \eta \ \wedge \ N_{k,s}(d) > 0, \ \forall s \in \{1,2,3\}\right\}, \\ \Delta_k^* &= d_k^*/f, \qquad \eta = 1.0. \end{aligned} \tag{8}$$

Polarity-aware Temporal Decomposition. To retain short-term temporal evolution without collapsing the entire window into a single event map, each candidate interval is divided into three equal-duration stages. Positive and negative events are accumulated separately in ROI-local coordinates and result in $H_k^{(d)}(s, p, y, x) \in \mathbb{R}^{3\times 2\times H_k^R \times W_k^R}$, where $p \in \{-1, +1\}$ denotes polarity.

RGB-format Event Rendering. To enable the MLLM to directly exploit motion-related information from event signals without introducing an additional visual encoder, we render the decomposed event maps as image-form inputs using a shared contrast scale. Each anchor yields three chronologically ordered images $\{I_k^{E,\text{early}}, I_k^{E,\text{middle}}, I_k^{E,\text{late}}\}$, so at most three anchors produce nine event images.

Table 2: **Comparison with state-of-the-art models on VAULT.** ***Zero-shot:*** We evaluate two major categories of models. The upper and lower panels correspond to open-source video-based models and VAU-specific models, respectively. ***SFT:*** We fine-tune Qwen3-VL-8B-Instruct and Qwen3.5-9B on VAULT. † denotes LoRA-based SFT, while ‡ denotes MoE-based SFT. Higher (↑) is better for all metrics. The best and second-best results for each metric are highlighted in **bold** and underlined, respectively.

	Model	Local.(%)	Classification(%)			Description					VQA(%)
		mIoU↑	Precis.↑	Recall↑	F1↑	S_{GPT}↑	BLEU-4↑	METEOR↑	CIDEr↑	ROUGE-L↑	ACC↑
Zero-shot	Qwen3.5-9B	46.31	46.24	63.61	51.11	28.72	2.39	18.38	7.87	19.54	54.10
	Qwen3-VL-8B-Instruct	20.51	43.33	55.68	46.15	30.47	2.48	**20.87**	6.22	19.66	68.31
	Qwen3-VL-2B-Instruct	15.42	40.76	53.39	44.08	19.42	2.19	17.07	3.78	16.23	48.31
	Qwen2.5-VL-7B-Instruct	11.24	47.43	50.11	46.87	4.61	0.40	5.09	0.72	9.54	64.15
	Qwen2-VL-7B-Instruct	10.82	51.70	48.58	48.07	17.80	1.60	12.77	4.81	16.80	67.10
	MiniCPM-V 4.5 (8.7B)	20.50	46.89	51.20	46.97	17.87	1.10	11.84	3.82	14.82	71.15
	VideoLLaMA3-7B (8B)	25.24	50.49	50.00	48.57	19.32	1.24	12.29	4.88	16.03	70.82
	VideoLLaMA2-7B (8B)	11.72	32.50	55.25	37.79	14.80	2.00	14.72	5.74	19.00	45.46
	Video-ChatGPT (7B)	10.95	6.91	28.69	8.21	7.83	1.89	15.23	2.99	18.86	1.42
	VTimeLLM (7B)	10.82	25.23	56.50	30.38	8.04	1.40	12.00	2.64	17.64	1.53
	HAWK (7B)	10.16	0.98	0.98	0.98	13.52	0.86	15.35	0.26	12.48	2.30
	Holmes-VAU (2B)	5.34	10.39	10.39	10.39	18.40	1.43	12.71	3.44	15.69	44.37
	CUVA	10.82	6.39	5.36	5.63	3.35	1.14	14.59	1.50	13.63	12.68
	VAU-R1	11.39	51.15	47.81	48.13	21.09	1.12	10.92	2.66	14.54	71.48
	Cue-R1	11.87	33.44	34.97	33.01	6.90	2.19	16.85	3.58	18.06	49.18
	Vad-R1	15.06	41.15	50.98	43.33	20.05	2.42	19.74	5.37	19.74	67.43
SFT	Qwen3-VL-8B-Instruct†	48.54	47.54	60.38	50.72	29.28	3.18	19.81	13.25	20.99	73.55
	+ EventVAULT (*Ours*)	55.54	57.16	50.27	52.30	**31.10**	3.24	19.76	11.45	21.20	78.25
	Qwen3-VL-8B-Instruct‡	14.33	56.34	49.13	51.26	27.93	2.85	18.61	10.83	20.92	74.97
	+ EventVAULT (*Ours*)	52.16	57.65	50.27	52.46	30.04	3.54	20.14	11.04	21.57	76.61
	Qwen3.5-9B†	58.89	62.84	55.68	57.84	30.35	2.22	17.35	7.40	19.15	75.19
	+ EventVAULT (*Ours*)	**62.92**	**70.82**	**63.83**	**65.79**	28.82	**3.55**	20.57	**14.67**	**21.78**	**79.13**

4.3 Joint RGB–Event Instruction Tuning

To jointly optimize the model for anomaly localization, category recognition, semantic description, and question answering, we perform instruction tuning across all four tasks defined in VAULT using complementary RGB and event evidence. For each training instance, let $\mathcal{X}$ denote the complete multimodal instruction consisting of the selected RGB observations, aligned event evidence, and task query, and let $\mathcal{Y} = \{y_t\}_{t=1}^{L}$ denote the target response. Training follows the standard autoregressive objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{L} \log P_\theta \left(y_t \mid \mathcal{X}, y_{<t} \right), \tag{9}$$

where θ denotes the trainable LoRA parameters.

5 Experiment

5.1 Experimental Setup

Benchmark Models. We evaluate models on VAULT under two settings: zero-shot and supervised fine-tuning (SFT). **Zero-shot:** we consider two model families: *(a) general video understanding models*, including the Qwen/Qwen-VL series (Qwen Team, 2026; Bai et al., 2025b; Wang et al., 2024a), MiniCPM-V 4.5 (Yu et al., 2025), VideoLLaMA2/3 (Cheng et al., 2024; Zhang et al., 2025a), Video-ChatGPT (Maaz et al., 2024), and VTimeLLM (Huang et al., 2024) and *(b) VAU-specific models*, including HAWK (Tang et al., 2024), Holmes-VAU (Zhang et al., 2025c), CUVA (Du et al., 2024b), VAU-R1 (Zhu et al., 2025), Cue-R1 (Yu et al., 2026) and Vad-R1 (Huang et al., 2025). **SFT:** we fine-tune Qwen3-VL-8B-Instruct using LoRA and MoE (Shazeer et al., 2017) LoRA, comparing each baseline with its EventVAULT-enhanced counterpart. We employ a comprehensive set of metrics to evaluate the four tasks, with detailed evaluation protocols provided in Appendix C.

Implementation Details. We generate event streams using v2e (Hu et al., 2021) and fine-tune Qwen3-VL-8B-Instruct using LoRA and MoE-LoRA. We utilize the AdamW (Loshchilov & Hutter, 2019) optimizer with a warm-up ratio of 0.1, linearly increasing the learning rate to 2×10^{-5}, followed by cosine decay. We set the effective batch size to 32. The LoRA rank r and scaling parameter α are set to 32 and 64, respectively. All experiments are conducted on a server equipped with eight NVIDIA RTX A6000 GPUs, each with 48GB of memory.

Table 3: **Ablation study for the designed components of EventVAULT on VAULT.** The RGB-only model is used as the baseline. The best result in each column is highlighted in bold.

Method	TEA	SEL	AEER	Localization(%)	Classification(%)			Description					VQA(%)
				mIoU↑	Precision↑	Recall↑	F1↑	S_{GPT} ↑	BLEU-4↑	METEOR↑	CIDEr↑	ROUGE-L↑	Accuracy↑
① RGB only (Baseline)	✗	✗	✗	48.54	47.54	**60.38**	50.72	29.28	3.18	19.81	13.25	20.99	73.55
② + TEA	✓	✗	✗	54.98	56.61	50.66	52.20	31.15	**3.38**	19.84	12.37	21.35	74.00
③ + SEL	✗	✓	✗	51.03	55.96	50.60	52.04	29.44	3.06	19.57	10.46	20.80	74.32
④ + TEA + SEL	✓	✓	✗	52.83	55.19	47.81	50.00	30.50	3.22	19.14	12.18	20.99	76.28
⑤ + AEER	✗	✗	✓	48.98	55.71	50.77	51.85	**31.65**	3.31	**19.96**	**14.10**	**21.48**	73.99
⑥ **+ TEA + SEL + AEER** (*Ours*)	✓	✓	✓	**55.54**	**57.16**	50.27	**52.30**	31.10	3.24	19.76	11.45	21.20	**78.25**

5.2 Comparison with State of the Arts

Table 2 presents zero-shot and supervised fine-tuning (SFT) results on extensive models. From the results in the table, we have the following key observations.

1) Zero-shot results demonstrate the challenges posed by VAULT. General video understanding models provide useful semantic capabilities but struggle with precise anomaly identification. Their Stronger VQA results suggest some understanding of event details, but anomaly recognition and localization remain challenging under poor illumination (*e.g.,* MiniCPM-V 4.5 achieves 71.15% VQA acc., but only 20.50% temporal mIoU). VAU-specific models do not show a consistent advantage on low-light samples despite their specialization, especially on temporal grounding and description tasks(*e.g.,* mIoU of VAU-R1 is only 11.39%, S_{GPT} of CUVA and Cue-R1 are far lower than that of others), suggesting limited transfer of existing anomaly-understanding capabilities to VAULT.

2) The consistent performance gains after SFT validate the value of the annotations provided by VAULT. As shown in the corresponding baseline rows of Table 2, all configurations improve anomaly classification and VQA over the zero-shot model. For Qwen3-VL-8B-Instruct, classification F1 increases from 46.15% to 50.72% and 51.26%, while VQA accuracy rises from 68.31% to 73.55% and 74.97%, respectively. For Qwen3.5, a substantial improvement of 21.09% can be observed on the VQA task. These gains demonstrate the value of VAULT annotations for adapting general-purpose models to low-light anomaly understanding.

3) Results of SFT with EventVAULT support the effectiveness of our EventVAULT. As observed, our method improves temporal mIoU by 7.00% and 37.83% under LoRA and MoE LoRA with Qwen3-VL, respectively and by 4.03% under LoRA with Qwen3.5, with evident increases on other tasks. The improvements on all configurations prove that EventVault can be used as a plug-and-play module to improve different models under low-light conditions.

5.3 Ablation Study

In this section, we conduct ablation studies to evaluate the impact of different design choices on the performance of EventVAULT. We perform experiments on the proposed VAULT benchmark, reporting the results across four tasks.

Effect of Various Components in EventVAULT. The proposed method consists of three key components: Temporal Evidence Allocation (TEA), Spatial Evidence Localization (SEL), and Adaptive Event Evidence Representation (AEER). We progressively integrate each module into the model. As observed in Table 3, adding TEA or SEL individually improves the RGB-only baseline. Combining the two modules brings further gains (VQA: 76.28%), suggesting their complementary contributions. Incorporating AEER further improves performance (mIoU: $52.83\% \rightarrow 55.54\%$; VQA: $76.28\% \rightarrow 78.25\%$), despite its mixed effects when used alone. Generally, the final model yields the best results among the six models.

Effect of RGB-frame Sampling in TEA. Brief anomalies may be overlooked by global frame sampling, motivating TEA to use event activity to guide the allocation of temporal observations. To evaluate this design, we compare global-only sampling, activity-region-only sampling, and TEA in Table 4. As observed, TEA outperforms both two counterparts, *i.e.*,

Table 4: **Ablation study of different RGB frame sampling strategies in TEA.**

Method	mIoU↑	F1↑	S_{GPT} ↑	Acc.↑
Global-only	50.59	52.48	33.14	75.52%
Activity-region-only	48.95	52.79	28.21	74.97%
TEA (*Ours*)	55.54	52.30	31.10	78.25%

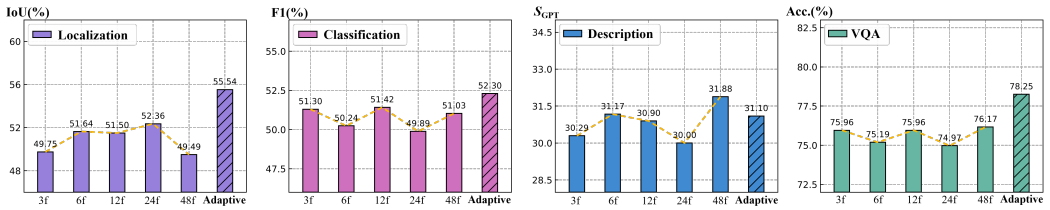

Figure 5: **Ablation study for the temporal window strategy in AEER.** Fixed event windows of 3, 6, 12, 24, and 48 source-frame periods are compared with the proposed adaptive window selection strategy on four tasks. The four panels report classification F1, temporal grounding mIoU, description S_{GPT}, and VQA accuracy, respectively. Hatched bars denote the adaptive configuration.

global-only and activity-region-only sampling, achieving an mIoU of 55.54% and a VQA accuracy of 78.25%. These results support its effectiveness in selecting informative temporal observations for anomaly understanding.

Effect of Spatial Observation in SEL. SEL uses local event activity to focus on spatially small anomalies while expanding the selected regions to preserve interaction context. Here, we compare full-frame observation, tight ROI cropping, and adaptive ROI (*ours*) in Table 5. We find that Adaptive ROI improves over both alternatives, *e.g.,* mIoU: $50.43\%/52.54\% \rightarrow 55.54\%$; Acc.: $74.97\%/74.54\% \rightarrow 78.25\%$, for full frame/tight ROI versus adaptive ROI. These gains suggest that focusing on relevant regions while retaining surrounding context benefits spatial evidence localization.

Table 5: **Ablation study of different spatial observation strategies in SEL.**

Method	mIoU↑	F1↑	S_{GPT} ↑	Acc.↑
Full frame	50.43	52.86	30.83	74.97%
Tight ROI	52.54	52.46	30.75	74.54%
Adaptive ROI (*Ours*)	55.54	52.3	31.10	78.25%

Effect of Temporal Window Strategy in AEER. To evaluate the effectiveness of adaptive temporal aggregation, we ablate the window selection strategy in AEER. We compare the proposed adaptive strategy with fixed windows of [3, 6, 12, 24, 48] source-frame periods across four VAU tasks. As shown in Fig. 5, the optimal fixed window varies across tasks, and longer windows do not consistently improve performance. In contrast, the proposed adaptive selection outperforms the best fixed configuration for each task, improving F1, mIoU and VQA accuracy by $+1.71\%$, $+6.07\%$ and $+2.73\%$, respectively. The results in the figure reveal that adjusting the aggregation duration to varying event dynamics supports more effective anomaly understanding.

6 Conclusion

In this work, we investigate fine-grained video anomaly understanding under challenging illumination conditions. To fill the lack of dedicated benchmarks for this problem, we introduce **VAULT**, a comprehensive low-light VAU benchmark that integrates diverse real-world illumination conditions with multi-dimensional annotations and four complementary understanding tasks. Experiments on VAULT reveal that current video MLLMs become increasingly unreliable in semantic understanding as illumination and RGB motion observability decrease, highlighting insufficient visual evidence as a critical bottleneck for VAU. To alleviate this issue, we further propose **EventVAULT**, a plug-in method which introduces event-derived dynamics to complement degraded RGB observations. Consistent improvements across all four tasks, particularly under challenging illumination, demonstrate the value of complementary dynamic evidence for robust anomaly understanding. Overall, VAULT provides a systematic testbed for studying illumination-aware VAU, while EventVAULT establishes an effective direction for improving anomaly understanding when conventional RGB evidence becomes unreliable. Given the prevalence and practical importance of low-light anomalies, together with the substantial challenges revealed in this work, we believe robust anomaly understanding under adverse illumination will remain an important direction for future VAU Research.

AI USE STATEMENT

In this work, we used generative AI tools to assist with and polish manuscript writing, primarily for improving clarity, fluency, and presentation. We also used generative AI tools for retrieval and discovery, including identifying potentially relevant related work. All suggested references were manually checked against the original sources before inclusion, and all AI-assisted text was reviewed and revised by the authors. No other use of generative AI requiring disclosure is applicable to this work. We take responsibility for the final content of this work, including all text, claims, citations, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made substantial efforts to ensure the reproducibility of this work. The construction of VAULT, including video collection, low-light screening, task definitions, annotation protocols, and dataset statistics, is described in the main paper and Appendix. The complete design of Event-VAULT, including TEA, SEL, AEER, and joint RGB–event instruction tuning, is provided in the methodology section together with the corresponding implementation details and hyperparameter settings. We further report the experimental setup, evaluation metrics, and task-specific evaluation protocols in the main paper and Appendix. The dataset, annotations, and source code required to reproduce the reported results are available at `https://anonymous.4open.science/r/VAULT`.

REFERENCES

Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. UBnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20143–20153, 2022. URL `https://openaccess.thecvf.com/content/CVPR2022/html/Acsintoae_UBnormal_New_Benchmark_for_Supervised_Open-Set_Video_Anomaly_Detection_CVPR_2022_paper.html`.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL `https://arxiv.org/abs/2511.21631`.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL `https://arxiv.org/abs/2502.13923`.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL `https://aclanthology.org/W05-0909/`.

Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. 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 (CVPR)*, pp. 20392–20401, 2023. doi: 10.1109/CVPR52729.2023.01953. URL `https://doi.org/10.1109/CVPR52729.2023.01953`.

Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18319–18329, 2024a. URL `https://openaccess.thecvf.com/content/CVPR2024/html/Chen_Prompt-Enhanced_Multiple_Instance_Learning_for_Weakly_Supervised_Video_Anomaly_Detection_CVPR_2024_paper.html`.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly

detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 387–395, 2023. doi: 10.1609/aaai.v37i1.25112. URL https://doi.org/10.1609/aaai.v37i1.25112.

Yingxian Chen, Jiahui Liu, Ruidi Fan, Yanwei Li, Chirui Chang, Shizhen Zhao, Wilton W. T. Fok, Xiaojuan Qi, and Yik-Chung Wu. Aligning effective tokens with video anomaly in large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22695–22706, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_Aligning_Effective_Tokens_with_Video_Anomaly_in_Large_Language_Models_ICCV_2025_paper.html.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b. URL https://arxiv.org/abs/2412.05271.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Hang Du, Guoshun Nan, Jiawen Qian, Wangchenhui Wu, Wendi Deng, Hanqing Mu, Zhenyan Chen, Pengxuan Mao, Xiaofeng Tao, and Jun Liu. Exploring what why and how: A multifaceted benchmark for causation understanding of video anomaly. *arXiv preprint arXiv:2412.07183*, 2024a. URL https://arxiv.org/abs/2412.07183.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, Dajiu Huang, Jing Feng, Linli Chen, Can Zhang, Xuhuan Li, Hao Zhang, Jianhang Chen, Qimei Cui, and Xiaofeng Tao. 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*, pp. 18793–18803, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Du_Uncovering_What_Why_and_How_A_Comprehensive_Benchmark_for_Causation_CVPR_2024_paper.html.

Linxuan Han, Fan Zhou, Baoquan Zhao, Hao Zhou, and Yu Liu. LVAD: A realistic data synthesis strategy and coarse-to-fine framework for low-light video anomaly detection. *Pattern Recognition*, 180:114222, 2026. doi: 10.1016/j.patcog.2026.114222. URL https://doi.org/10.1016/j.patcog.2026.114222.

Yixiao He, Menghao Zhang, Haifeng Sun, Jing Wang, Kangheng Lin, Jinghan Wang, Chenye Xu, Pengfei Ren, Qi Qi, and Jingyu Wang. VALU: A benchmark for video anomaly temporal localization and understanding at multiple semantic levels. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 1252–1296, 2026. URL https://aclanthology.org/2026.acl-long.56/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1312–1321, 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/EventVision/html/Hu_v2e_From_Video_Frames_to_Realistic_DVS_Events_CVPRW_2021_paper.html.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VTimeLLM_Empower_LLM_to_Grasp_Video_Moments_CVPR_2024_paper.html.

Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3952. URL https://doi.org/10.52202/085713-3952.

Taewoo Kim and Kuk-Jin Yoon. Event-guided unified framework for low-light video enhancement, frame interpolation, and deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8524–8534, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kim_Event-guided_Unified_Framework_for_Low-light_Video_Enhancement_Frame_Interpolation_and_ICCV_2025_paper.html.

Jiaxu Leng, Zhanjie Wu, Mingpi Tan, Yiran Liu, Ji Gan, Haosheng Chen, and Xinbo Gao. Beyond Euclidean: Dual-space representation learning for weakly supervised video violence detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0552. URL https://doi.org/10.52202/079017-0552.

Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3250–3259, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_Event-assisted_Low-Light_Video_Object_Segmentation_CVPR_2024_paper.html.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Yiran Liu, Zhanjie Wu, Mengjingcheng Mo, Ji Gan, Jiaxu Leng, and Xinbo Gao. Dual space embedding learning for weakly supervised audio-visual violence detection. In *IEEE International Conference on Multimedia and Expo*, 2024. doi: 10.1109/ICME57554.2024.10687755. URL https://doi.org/10.1109/ICME57554.2024.10687755.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2720–2727, 2013. doi: 10.1109/ICCV.2013.338. URL https://doi.org/10.1109/ICCV.2013.338.

Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 341–349, 2017. doi: 10.1109/ICCV.2017.45. URL https://doi.org/10.1109/ICCV.2017.45.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.679. URL https://doi.org/10.18653/v1/2024.acl-long.679.

Mengjingcheng Mo, Jiankang Zheng, Jiaxu Leng, and Xinbo Gao. Retrieval-guided contextual inference for training-free video anomaly detection in low-light scenarios. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 930–939, 2026. doi: 10.1145/3805622.3810823. URL https://doi.org/10.1145/3805622.3810823.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Joao Alexandre Cardeira Pereira, Vasco Lopes, João C. Neves, and David Semedo. Finevau: A novel human-aligned benchmark for fine-grained video anomaly understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8403–8411, 2026. doi: 10.1609/aaai.v40i10.37790. URL https://doi.org/10.1609/aaai.v40i10.37790.

Mauricio Perez, Alex C. Kot, and Anderson Rocha. Detection of real-world fights in surveillance videos. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2662–2666, 2019. doi: 10.1109/ICASSP.2019.8683676. URL https://doi.org/10.1109/ICASSP.2019.8683676.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.html.

Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4435. URL https://doi.org/10.52202/079017-4435.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. URL https://arxiv.org/abs/2409.12191.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024b. URL https://arxiv.org/abs/2411.10442.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pp. 322–339, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/7476_ECCV_2020_paper.php.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18297–18307, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Open-Vocabulary_Video_Anomaly_Detection_CVPR_2024_paper.html.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. VadCLIP: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024b. doi: 10.1609/aaai.v38i6.28423. URL https://doi.org/10.1609/aaai.v38i6.28423.

Peng Wu, Yuting Yan, Guansong Pang, Yujia Sun, Qingsen Yan, Peng Wang, and Yanning Zhang. Towards video anomaly detection from event streams: A baseline and benchmark datasets. In *European Conference on Computer Vision*, pp. 92–109, 2026. URL https://arxiv.org/abs/2603.24991.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8679–8688, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ye_VERA_Explainable_Video_Anomaly_Detection_via_Verbalized_Learning_of_Vision-Language_CVPR_2025_paper.html.

Tianyu Yu, Zefan Wang, Chongyi Wang, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. URL https://arxiv.org/abs/2509.18154.

Yating Yu, Congqi Cao, Zhaoying Wang, Weihua Meng, Jie Li, Yuxin Li, Zihao Wei, Zhongpei Shen, and Jiajun Zhang. Cuebench: Advancing unified understanding of context-aware video anomalies in real-world. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14): 12187–12195, 2026. doi: 10.1609/aaai.v40i14.38209. URL https://doi.org/10.1609/aaai.v40i14.38209.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zanella_Harnessing_Large_Language_Models_for_Training-free_Video_Anomaly_Detection_CVPR_2024_paper.html.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-VAU: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13843–13853, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Holmes-VAU_Towards_Long-term_Video_Anomaly_Understanding_at_Any_Granularity_CVPR_2025_paper.html.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13843–13853, 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Holmes-VAU_Towards_Long-term_Video_Anomaly_Understanding_at_Any_Granularity_CVPR_2025_paper.html.

Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3769–3777, 2023. doi: 10.1609/aaai.v37i3.25489. URL https://doi.org/10.1609/aaai.v37i3.25489.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2856. URL https://doi.org/10.52202/079017-2856.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*, 2025. URL https://arxiv.org/abs/2505.23504.

A Data Collection and Annotation Workflow

A.1 Data Collection Protocol

We collect raw videos from two sources: existing datasets and YouTube. First, we reuse candidate videos from several established VAD and VAU datasets, including UCF-Crime (Sultani et al., 2018), MSAD (Zhu et al., 2024), XD-Violence (Wu et al., 2020), CCTV-Fights (Perez et al., 2019), and ECVA (Du et al., 2024a) , which collectively cover diverse surveillance scenarios and a wide range of well-established anomaly types. Second, we retrieve additional real-world videos from YouTube to enrich the dataset with scenarios, anomaly categories, and illumination conditions. To ensure that VAULT reflects realistic and meaningful low-light anomaly understanding, we follow the criteria below when selecting videos from existing datasets and online:

- **Real-world anomalies.** Only videos depicting real-world anomalous events are retained. Animations, game footage, unrealistic cinematic effects, and news reports that do not directly present the anomaly are excluded.
- **Challenging low-light conditions.** The videos should contain challenging illumination conditions. Low light is not limited to globally dark scenes, but also includes extremely dark regions, spatially uneven illumination, strong shadows, backlighting, underexposure, and large contrast between bright and dark regions.
- **Natural illumination degradation.** Low-light characteristics should naturally arise from the captured environment, lighting conditions, or camera exposure, rather than being synthetically generated through post-processing.
- **Directly observable anomaly evidence.** The anomalous event should be directly observable from the video itself, rather than inferred solely from captions, narration, audio, or external contextual information.
- **Sufficient temporal context.** The video should preserve sufficient temporal context to understand and localize the occurrence and evolution of anomalous events, rather than containing only severely truncated event fragments.
- **Human interpretability.** Despite challenging illumination, the anomalous event should remain reliably interpretable by human annotators. Videos dominated by near-complete darkness, severe occlusion, extreme blur, or visual corruption are excluded.
- **Limited editing and acquisition artifacts.** Videos dominated by frequent shot transitions, severe camera shake, flashing effects, or other artifacts that introduce substantial activity unrelated to the anomalous event are excluded.

A.2 Annotation Pipeline

Automatic low-light screening. After collecting the candidate videos, we first perform automatic low-light screening while computing an illumination score for each video. Manually examining a large collection of candidates is both costly and difficult to standardize; therefore, we employ an automatic procedure to identify videos with low apparent brightness before human verification. Specifically, RGB frames are converted into the HSV color space, and the value channel V is used to characterize image brightness. For a video $\mathcal{V}$ consisting of T sampled frames, we define its normalized apparent brightness score as

$$B(\mathcal{V}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} V_t(i,j), \tag{10}$$

where $V_t(i,j)$ denotes the normalized HSV-V value of pixel (i,j) in the t-th sampled frame. A lower score indicates a darker appearance in the camera imaging domain. Note that this score measures the *observed apparent brightness* of a video rather than the physical illumination intensity of the underlying environment. The computed illumination score serves two purposes: it enables efficient large-scale low-light screening and provides a quantitative measure for subsequent illumination-aware analysis in VAULT. As shown in Fig. 2, more than 85% of the collected videos fall within the low-apparent-brightness range, confirming low-light conditions as a dominant characteristic of the benchmark.

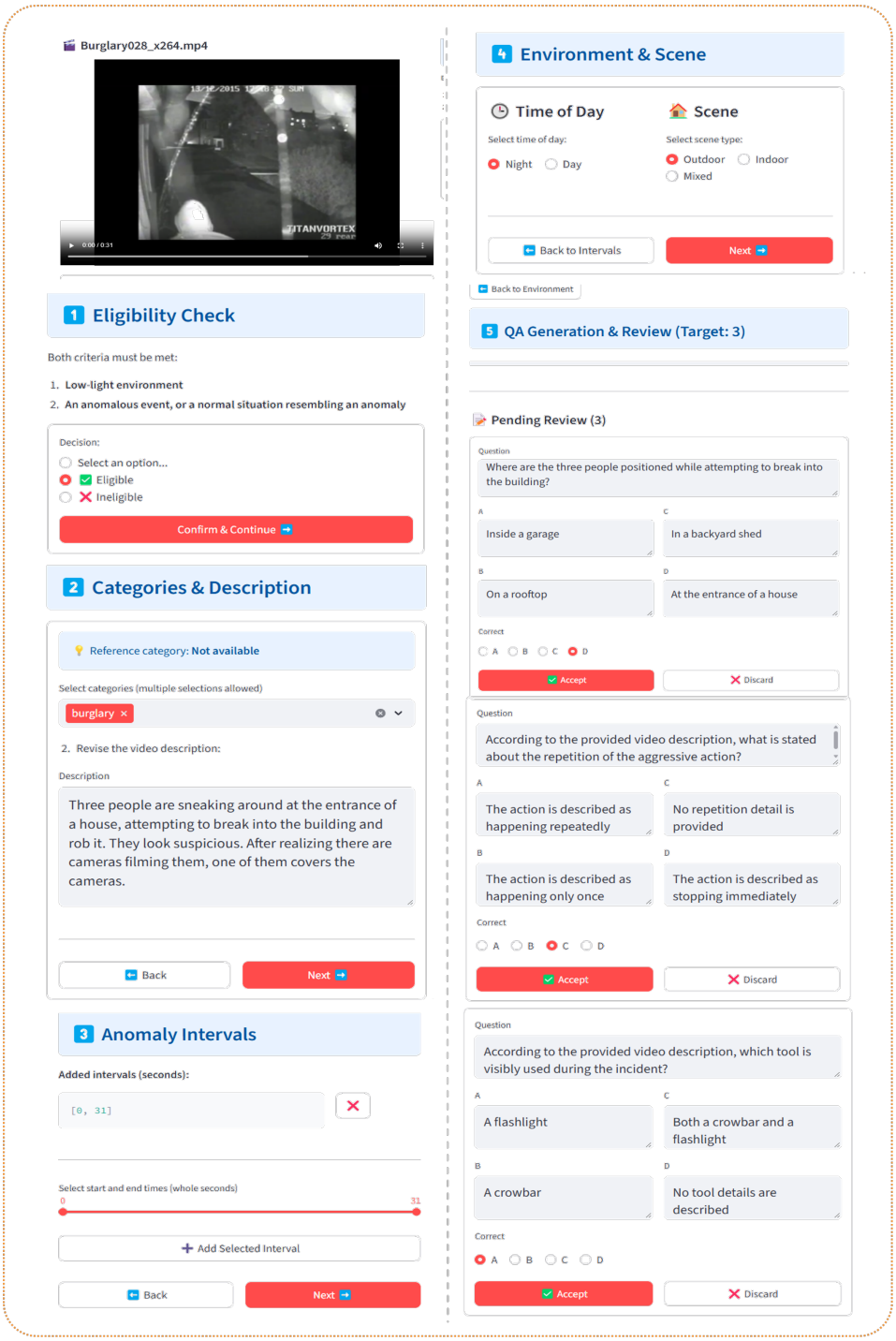

Figure 6: Overview of the VAULT annotation interface.

After automatic low-light screening, the retained videos enter the human annotation stage. Annotators first exclude videos that violate the selection criteria described in Appendix A.1, and then perform fine-grained annotation on the remaining samples. The overall annotation pipeline is illustrated in Fig. 3, and the annotation interface is shown in Fig. 6.

Temporal grounding annotation. Annotators are required to watch the entire video and record the temporal extent of every anomalous event. Since a video may contain multiple anomalous

events, the annotation consists of multiple start–end timestamp pairs at second-level precision. If no anomaly is present, the temporal label is recorded as $(-1, -1)$.

Anomaly classification annotation. Annotators select the corresponding anomaly categories from the predefined category set maintained by the annotation system. If an appropriate category is not available, they may propose a new category. Such proposals are reviewed by the annotation-system administrator, and approved category updates are subsequently announced to all annotators to maintain label consistency. The final annotation is either *normal* or a set of anomaly labels drawn from the category set $\mathcal{C}$.

Anomaly description annotation. To reduce the annotation burden, we employ GPT-4o (OpenAI, 2024) as an annotation assistant. Based on the anomaly categories selected in the previous stage together with sampled video frames, GPT-4o generates an initial natural-language description, which is then returned to the annotation interface for human review and revision. Annotators are responsible for correcting inaccurate or incomplete content before submission. For videos without anomalous events, annotators are instead asked to describe the normal activities occurring in the video.

Anomaly VQA annotation. VQA annotation follows a similar human-in-the-loop procedure. Given the annotated anomaly categories and descriptions, GPT-4o proposes three video-grounded multiple-choice questions together with candidate options and answers. Annotators then review and revise these proposals. They are encouraged to prioritize challenging questions that require visual evidence under low-light conditions rather than questions that can be answered from generic context alone. If a proposed VQA sample is considered low-quality or uninformative, the annotator may request regeneration from GPT-4o or manually create the question, options, and answer. Videos without anomalies are also annotated with VQA samples centered on normal events and scene content.

B Related Work

Video Anomaly Benchmarks. Early video anomaly benchmarks primarily support anomaly detection and temporal localization. UCF-Crime (Sultani et al., 2018) provides long untrimmed surveillance videos covering diverse real-world crimes, while XD-Violence (Wu et al., 2020) extends anomaly detection to larger-scale and more diverse violent events. With the shift from detection to semantic understanding, recent benchmarks provide richer supervision beyond binary anomaly labels. CUVA (Du et al., 2024b) introduces causation-oriented annotations for understanding what happens, why it happens, and how an anomalous event evolves. HIVAU-70K (Zhang et al., 2025b) further supports hierarchical anomaly understanding across multiple temporal granularities, while VALU (He et al., 2026) emphasizes fine-grained temporal-semantic annotations for video anomaly understanding. In parallel, LVAD (Han et al., 2026) begins to study illumination degradation by constructing low-light variants of conventional VAD datasets. However, existing low-light benchmarks remain detection-oriented, whereas current VAU benchmarks rarely treat illumination as an explicit dimension of data construction and evaluation. VAULT addresses this gap by combining real-world low-light conditions with classification, temporal grounding, description, and VQA in a unified VAU benchmark.

Video Anomaly Detection. Conventional VAD focuses on identifying anomalous events and locating their temporal occurrence. Under weak supervision, recent methods improve anomaly discrimination through different forms of representation modeling. MGFN (Chen et al., 2023) introduces magnitude-aware feature learning to enhance the separation between normal and anomalous snippets, while UR-DMU (Zhou et al., 2023) models normal and abnormal patterns with dual memory units. DSEL (Liu et al., 2024) and DSRL (Leng et al., 2024) explore more expressive geometric spaces for modeling complex anomaly relations. With the emergence of large-scale vision-language pretraining, VadCLIP (Wu et al., 2024b), PE-MIL (Chen et al., 2024a), and OVVAD (Wu et al., 2024a) exploit semantic priors from pretrained vision-language models to improve anomaly recognition, temporal localization, and open-vocabulary generalization. LAVAD (Zanella et al., 2024) further introduces LLM-based reasoning into training-free anomaly inference. Despite substantial

progress, these methods mainly produce anomaly scores, temporal boundaries, or category predictions, and therefore provide limited support for interpreting the detailed semantics of anomalous events.

Video Anomaly Understanding with MLLMs. The development of multimodal large language models has recently shifted anomaly analysis toward fine-grained video anomaly understanding. General-purpose video MLLMs, including the Qwen-VL family (Qwen Team, 2026; Bai et al., 2025a;b; Wang et al., 2024a), MiniCPM-V (Yu et al., 2025), VideoLLaMA2/3 (Cheng et al., 2024; Zhang et al., 2025a), Video-ChatGPT (Maaz et al., 2024), and VTimeLLM (Huang et al., 2024), provide strong foundations for video perception, temporal reasoning, and instruction following, but are not specifically optimized for anomalous events. Recent anomaly-oriented models further extend VLMs and MLLMs toward anomaly-specific reasoning. VERA (Ye et al., 2025) introduces verbalized learning with learnable guiding questions, enabling frozen VLMs to jointly detect anomalies and generate interpretable explanations without parameter updates. CUVA (Du et al., 2024b) focuses on causation-oriented anomaly reasoning, while HAWK (Tang et al., 2024) incorporates motion-aware information for open-world anomaly understanding. Holmes-VAU (Zhang et al., 2025b) addresses long-term and multi-granularity anomaly understanding through hierarchical supervision and anomaly-focused temporal sampling. VA-GPT (Chen et al., 2025) explicitly addresses the spatial and temporal sparsity of abnormal events by selecting spatially effective tokens and generating temporal effective tokens for anomaly summarization and localization. VAU-R1 (Zhu et al., 2025) further explores reasoning-enhanced anomaly understanding with MLLMs. These studies highlight the importance of identifying sparse and informative anomaly evidence in addition to improving high-level semantic reasoning.

Low-Light Anomaly Analysis. Low illumination introduces a distinct challenge because the visual evidence required by both VAD and VAU can become weak or ambiguous before high-level reasoning is performed. LVAD (Han et al., 2026) studies this problem through illumination-aware low-light anomaly detection, while ReCI (Mo et al., 2026) improves low-light VAD using contextual information under degraded RGB observations. These studies establish the importance of illumination robustness for anomaly analysis, but their objectives remain primarily anomaly scoring and temporal localization. Fine-grained tasks such as anomaly description, multi-label recognition, and visual question answering under varying illumination remain substantially less explored.

Our work lies at the intersection of these directions. Rather than using events solely for reconstruction, enhancement, or anomaly scoring, EventVAULT generates event stream from RGB videos and exploits event-derived dynamics in two complementary ways: TEA and SEL use event activity to determine *when* and *where* informative RGB observations should be retained, while AEER provides aligned event dynamics as complementary visual evidence for understanding *how* anomalous events evolve. Together with VAULT, this enables systematic study of fine-grained RGB–event anomaly understanding under diverse illumination conditions.

C Evaluation Metrics

Our evaluation suite covers four major tasks: (a) **Temporal Grounding**: We report mean global temporal intersection over union (mIoU), measuring the overlap between the unions of predicted and ground-truth anomalous intervals. (b) **Anomaly Classification**: Precision, recall, and F1 assess the correctness and completeness of multi-label predictions. (c) **Anomaly Description**: BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015), and S_{GPT}, which evaluate Semantic Correctness, Detail Coverage, Event Causality, Anomaly Saliency, and Hallucination Penalty; (d) **VQA**: We adopt permutation-robust accuracy, counting a question as correct only when the model answers correctly under all four cyclic permutations of its candidate choices. All other metrics follow their standard definitions except for S_{GPT}. The detailed scoring criteria for S_{GPT} are provided below.

C.1 Scoring Criteria for S_{GPT}

Given a reference description R and a model-generated prediction P, we employ GPT-4o-mini as the evaluator to independently assess the prediction from five complementary dimensions: **Semantic**

Correctness, **Detail Coverage**, **Event Causality**, **Anomaly Saliency**, and **Hallucination Penalty**. The first four dimensions are scored from 0 to 10, while Hallucination Penalty adopts a non-positive score ranging from -10 to 0. The evaluator is explicitly instructed to focus on factual consistency rather than linguistic style, such that concise descriptions are not penalized as long as they correctly capture the underlying anomalous event.

Semantic Correctness. Semantic Correctness evaluates whether the prediction and the reference describe the same core event. The evaluator focuses exclusively on factual alignment and ignores differences in wording, tone, sentence structure, or descriptive richness. Therefore, a concise prediction can still receive a high score as long as the main entities, actions, and event semantics are consistent with the reference.

Prompt for Semantic Correctness

```
You are a strict factual evaluator for video anomaly descriptions.

Compare the predicted description with the reference description.

## Instructions:

- Focus EXCLUSIVELY on factual alignment. Do they describe the same
core event?

- DO NOT penalize the prediction for being concise, robotic, or
lacking descriptive adjectives. A dry, straightforward description
that captures the facts is worth 10/10.

- Ignore differences in tone, style, or emotional vocabulary.

Return only a Python dictionary string in this format:

{'score': X}

where X is an integer between 0 and 10.
```

Figure 7: Prompt used for evaluating semantic correctness in S_{GPT}.

Detail Coverage. Detail Coverage measures how completely the prediction preserves the essential objective information contained in the reference. In particular, the evaluation considers three primary components: *who* is involved, *what* abnormal action occurs, and *where* the event takes place. The metric evaluates factual coverage rather than linguistic richness. For example, descriptions with different levels of lexical sophistication are considered equivalent if they convey the same entities and actions.

Prompt for Detail Coverage

```
You are an evaluator for detail coverage in video anomaly
descriptions.

## Instructions:

- Check if the prediction covers the essential objective elements
present in the reference: Who (entities involved), What (the exact
abnormal action), and Where (the environment).

- Only score based on the presence of these factual elements.

- DO NOT evaluate the richness of the language. "A man hits a dog"
is equal to "A male individual aggressively strikes a canine" if the
core entities and actions match.

Return only a Python dictionary string in this format:

{'score': X}

where X is an integer between 0 and 10.
```

Figure 8: Prompt used for evaluating detail coverage in S_{GPT}.

Event Causality. Event Causality assesses whether the generated description correctly captures the temporal and causal progression of the anomalous event. The evaluator examines whether the prediction preserves the sequence of relevant event stages, such as an initial state, an abnormal trigger, and its subsequent consequence or reaction. The evaluation emphasizes the correctness of the event sequence rather than sentence complexity. A simple chronological description is therefore sufficient if it accurately represents the event progression.

Prompt for Event Causality

```
You are an evaluator for event causality and temporal flow in video
anomaly descriptions.

## Instructions:

- Does the prediction correctly reflect the sequence of events (e.g.,
Normal state -> Abnormal trigger -> Consequence/Reaction)?

- Focus on the accuracy of the action sequence, not the complexity
of the sentence structure. A simple chronological list of events is
completely acceptable.

Return only a Python dictionary string in this format:

{'score': X}

where X is an integer between 0 and 10.
```

Figure 9: Prompt used for evaluating event causality in S_{GPT}.

Anomaly Saliency. Anomaly Saliency evaluates whether the prediction successfully identifies the primary anomalous event while avoiding distraction from irrelevant background activities. It additionally examines whether the generated description correctly captures the specific anomaly category represented in the reference. The evaluator does not reward dramatic or exaggerated expressions; objective descriptions are preferred as long as they accurately characterize the underlying anomaly.

Prompt for Anomaly Saliency

```
You are an evaluator for anomaly saliency in video anomaly
descriptions.

## Instructions:

- Does the prediction correctly identify the core abnormal event
without being distracted by irrelevant background actions?

- Did it capture the specific anomaly category accurately?

- DO NOT penalize if the prediction lacks dramatic or exaggerating
words (e.g., "catastrophic", "horrific"). Objective terms (e.g.,
"car crash", "fire starts") are preferred.

Return only a Python dictionary string in this format:

{'score': X}

where X is an integer between 0 and 10.
```

Figure 10: Prompt used for evaluating anomaly saliency in S_{GPT}.

Hallucination Penalty. In addition to evaluating descriptive quality, we explicitly assess unsupported information introduced by the model. Hallucination Penalty measures factual content appearing in the prediction that is not supported by the reference description. Importantly, missing information is not considered a hallucination, since incompleteness is separately captured by Detail Coverage.

The evaluator mainly considers the following three types of hallucination:

- **Over-specification:** introducing unsupported attributes such as colors, exact times, or precise locations, which can be particularly unreliable under low-light conditions;
- **Fabricated entities:** mentioning people, objects, vehicles, or other entities that are absent from the reference;
- **Invented causality:** introducing unsupported explanations or causes for an anomalous event when the reference only describes the observed action.

Unlike the previous metrics, Hallucination Penalty follows a negative scoring scheme:

$$S_{\text{hall}} \in [-10, 0], \tag{11}$$

where 0 denotes a fully faithful description without unsupported information.

Prompt for Hallucination

```
You are a strict fact-checker evaluating hallucinations in video
descriptions.

## Instructions:

- Compare the prediction to the reference. Penalize the prediction
ONLY for ADDING information that is NOT supported by the reference.

- Hallucinations include:

  1. Over-specification: Guessing specific colors, exact times, or
specific locations not mentioned in the reference (common in low-
visibility or dark scenes).

  2. Fabricated Entities: Mentioning people, objects, or vehicles
that do not exist in the reference.

  3. Invented Causality: Guessing the "reason" for an anomaly when
the reference only states the action.

- Missing information is NOT a hallucination (that is handled by
detail coverage). Only penalize for fabricated additions.
```

Figure 11: Prompt used for evaluating hallucination in S_{GPT}.

Evaluation Protocol. For each video, the generated anomaly description is independently evaluated under the five criteria described above. The reference description and the prediction are provided to the evaluator using dimension-specific evaluation instructions. We use GPT-4o-mini as the evaluator and set the temperature to zero to reduce randomness during evaluation.

Considering the different importance of each evaluation dimension for anomaly description, we assign different weights to the corresponding scores:

$$S_{\text{base}} = 2.5 \times S_{\text{sem}} + 2 \times S_{\text{detail}} + 2 \times S_{\text{cause}} + 3.5 \times S_{\text{sal}}. \tag{12}$$

In particular, hallucination differs from the other four dimensions, as it is a negative evaluation criterion. We therefore treat it separately and rescale it into a multiplicative coefficient applied to the positive-dimension score:

$$w_{\text{hall}} = 1 + \frac{S_{\text{hall}}}{20}. \tag{13}$$

The final S_{GPT} is computed as:

$$S_{\text{GPT}} = S_{\text{base}} \times w_{\text{hall}} \tag{14}$$

This formulation jointly rewards semantic fidelity, information coverage, temporal reasoning, and anomaly awareness while explicitly penalizing unsupported content.

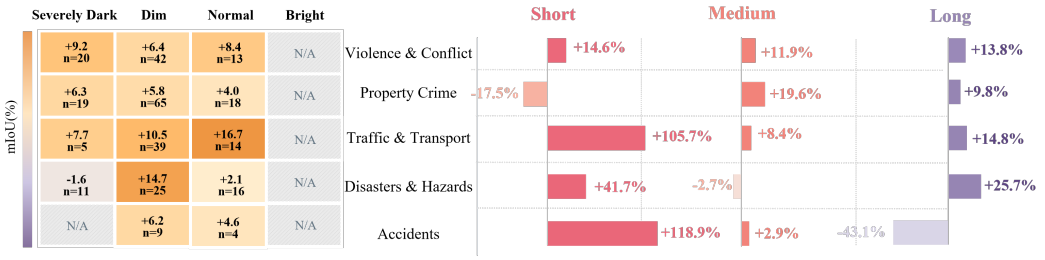

Figure 12: **Fine-grained analysis of the mIoU gains brought by EventVAULT across different anomaly groups under varying illumination conditions and video durations.** Left: absolute mIoU improvements across illumination levels, where each cell also reports the number of samples; N/A denotes inadequate samples. Right: relative mIoU improvements for short-, medium-, and long-duration videos.

At the dataset level, the score of each evaluation dimension is obtained by averaging over all samples with valid evaluation results:

$$\bar{S}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} S_{i,d}, \tag{15}$$

where d denotes an evaluation dimension, N_d is the number of samples with a valid score for dimension d, and $S_{i,d}$ denotes the corresponding score of the i-th sample.

Compared with a single lexical similarity metric, this multidimensional evaluation provides a more interpretable assessment of generated anomaly descriptions. It allows us to distinguish whether model errors mainly arise from incorrect event recognition, missing critical details, inaccurate event progression, insufficient anomaly focus, or hallucinated content.

D Fine-grained Analysis across Illumination and Video Duration

To further understand where the proposed method brings the largest benefits, we analyze the performance gains of EventVAULT over the RGB-only baseline across different anomaly groups under varying illumination conditions and video durations, as shown in Fig. 12.

Illumination-aware analysis. The improvement brought by EventVAULT is generally more pronounced under challenging illumination, indicating that event-derived dynamic evidence becomes particularly valuable when RGB appearance cues are degraded. This trend, however, is not uniform across anomaly groups. Some groups exhibit consistently positive gains across multiple illumination levels, while others show stronger improvements only under specific lighting conditions. For example, *Traffic & Transport* benefits consistently from EventVAULT across all available illumination levels, with mIoU gains of $+7.7\%$, $+10.5\%$, and $+16.7\%$ points under severely dark, dim, and normal conditions, respectively. This suggests that event-derived motion cues provide useful complementary information for traffic-related anomalies across a broad range of illumination conditions, rather than only under extreme darkness. In contrast, *Disasters & Hazards* exhibits a more illumination-sensitive pattern: EventVAULT causes a slight degradation of -1.6 points under severely dark conditions, but yields a substantial $+14.7\%$ point improvement under dim illumination. This variation indicates that the usefulness of event cues depends not only on illumination severity, but also on the motion characteristics and visual structure of individual anomaly categories. Overall, these results suggest that EventVAULT does not simply provide a uniform performance offset; instead, it selectively compensates for categories whose discriminative evidence becomes less reliable as illumination deteriorates.

Duration-aware analysis. A similar category-dependent pattern appears when the videos are grouped by duration. EventVAULT generally provides positive improvements across short-, medium-, and long-duration videos, suggesting that the proposed event-guided representation remains beneficial under different temporal scales. Notably, several anomaly groups obtain particularly large gains on short videos, including relative improvements exceeding 100%, indicating

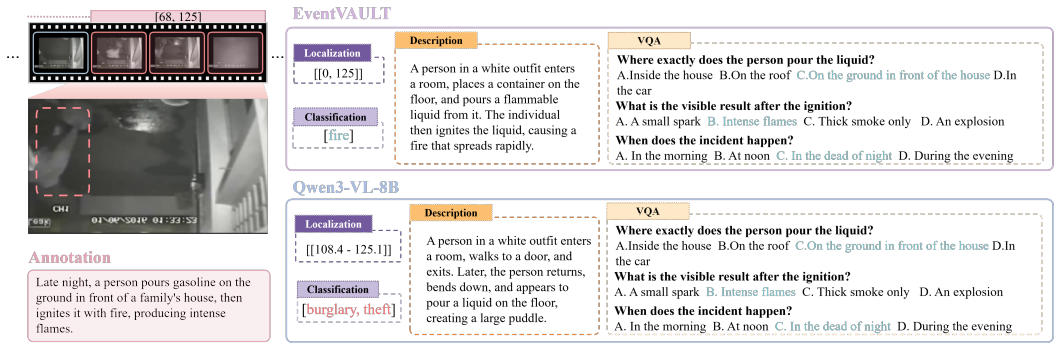

Figure 13: **Capturing anomaly-defining dynamics in nighttime arson.** The RGB-only baseline recognizes the preceding pouring action but fails to incorporate the subsequent ignition into its anomaly interpretation, whereas EventVAULT captures the transition to fire and predicts the corresponding anomaly category.

that event-guided temporal evidence allocation is effective for capturing brief and easily missed anomalous dynamics. For medium-length videos, the gains are generally more moderate, suggesting that sufficient RGB observations already provide stronger temporal coverage in this regime. Long videos show a different pattern: most anomaly groups still benefit from EventVAULT, with relative improvements ranging approximately from $+9.8\%$ to $+25.7\%$, whereas one group exhibits a noticeable degradation. This suggests that event guidance can help preserve sparse anomaly evidence over extended temporal contexts, while its effectiveness may decrease when event activity is dominated by prolonged or background motion unrelated to the anomaly.

Taken together, these results suggest that EventVAULT is particularly effective when anomalous evidence is visually degraded or temporally sparse, while the category-dependent variations further indicate that its benefit is primarily driven by the motion characteristics of individual anomaly types rather than by illumination or duration alone.

E Qualitative Study

We further provide qualitative comparisons to illustrate *how* EventVAULT improves anomaly understanding beyond the RGB-only baseline. Rather than treating the visualizations as standalone case analyses, we use them to examine whether the improvements observed in the quantitative results are reflected in the model outputs. Specifically, we focus on three representative behaviors: (1) capturing anomaly-defining dynamics under low illumination, (2) distinguishing fine-grained anomalous semantics when temporal evidence is similar, and (3) integrating temporally extended or multi-stage anomalous events. The examples are evaluated across localization, classification, description, and VQA using the same ground-truth annotations for both models.

Capturing anomaly-defining dynamics. Fig. 13 shows a nighttime arson case in which the critical evidence is not a single static appearance cue but the progression from pouring a flammable liquid to ignition. The RGB-only baseline recognizes the pouring action, yet its description terminates before the ignition and its classification shifts to unrelated categories (*burglary* and *theft*). In contrast, EventVAULT preserves the causal event progression in its description and identifies *fire*, increasing the classification F1 from 0 to 66.67%. This example provides an intuitive explanation for the benefit of exploiting event-sensitive temporal evidence: the additional dynamic cues help the model retain the short but anomaly-defining transition that is difficult to recover from low-light RGB appearance alone. The predicted interval is still over-extended, indicating that improved event recognition does not necessarily imply perfectly precise temporal boundaries.

Improving fine-grained anomaly semantics. The car-intrusion example in Fig. 14 isolates a different source of improvement. EventVAULT and the RGB-only baseline obtain nearly identical temporal localization (mIoU 88.20% versus 87.50%), so the performance difference cannot be attributed to better boundary prediction in this case. Instead, EventVAULT predicts only the reference

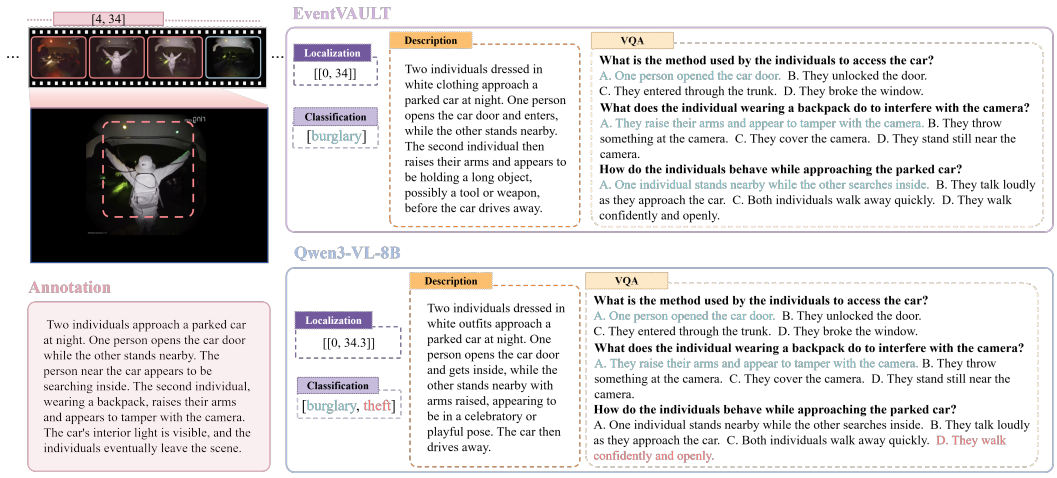

Figure 14: **Fine-grained semantic understanding of nighttime car intrusion.** Although the two models produce similar temporal localization, EventVAULT gives a more precise anomaly category and correctly resolves the interaction between the two individuals in VQA.

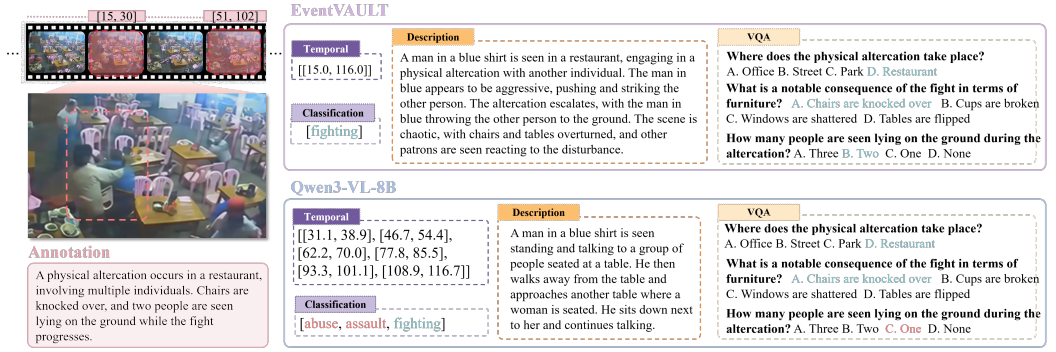

Figure 15: **Integrating temporally extended fighting events.** EventVAULT produces a more coherent localization of the overall altercation, predicts the correct anomaly category, and preserves key interaction details in VQA, whereas the RGB-only baseline produces fragmented temporal predictions and weaker semantic understanding.

category *burglary*, while the baseline additionally introduces *theft*, resulting in classification F1 scores of 100.00% and 66.67%, respectively. A similar difference appears in VQA: both models recognize obvious actions such as opening the vehicle and raising the arms, but only EventVAULT correctly identifies the coordinated behavior in which one person remains nearby while the other searches inside the car. This comparison suggests that the proposed event-aware representation is particularly useful for refining *what* is happening after the relevant temporal evidence has already been located, rather than merely improving localization.

Integrating extended and multi-stage anomalies. Fig. 15 illustrates a more temporally complex case containing two annotated fighting intervals. The RGB-only baseline produces six short and fragmented predictions (mIoU 31.1%), whereas EventVAULT returns a single interval covering the dominant span of the altercation and reaches 64.4% mIoU. The improvement also extends beyond localization. EventVAULT predicts the reference category *fighting* alone (classification F1 100.00%), while the baseline introduces additional *abuse* and *assault* labels (F1 50.00%). Moreover, EventVAULT correctly answers questions concerning the scene, overturned chairs, and the number of people lying on the ground, whereas the baseline misses the latter detail. Together, these outputs indicate that the proposed representation helps maintain coherent anomaly evidence across a longer temporal span instead of reducing the event to isolated short-term observations.

Table 6: **Comparison on VAU-Bench under the w/o thinking setting.** QA accuracy and temporal grounding metrics are reported in percentages, while each VAU-Eval dimension is scored out of 10. Published results are from VAU-R1 Zhu et al. (2025). The best and second-best results for each metric within each dataset are highlighted with darker and lighter backgrounds, respectively.

Dataset	Model	QA(%)	VAU-Eval						Temporal Grounding%			
		Acc.	CLS	KM	FLU	INF	FAC	Total	mIoU	R@0.3	R@0.5	R@0.7
MSAD	InternVL2.5-2B	76.67	6.84	6.23	8.55	6.64	6.64	34.90	–	–	–	–
	Qwen2.5-VL-7B	84.58	6.75	**6.41**	9.27	**7.74**	**6.92**	**37.08**	45.90	70.83	45.83	21.67
	InternVL2.5-8B-MPO	82.50	6.83	6.33	8.32	6.37	6.86	34.72	–	–	–	–
	Qwen2-VL-2B	77.08	5.94	5.43	8.77	6.29	5.90	32.25	0.00	0.00	0.00	0.00
	+*SFT*	82.92	6.04	5.43	8.89	6.55	5.93	32.84	–	–	–	–
	+*RFT*	82.92	6.05	5.49	8.89	6.50	6.05	32.98	–	–	–	–
	Qwen2.5-VL-3B	85.83	5.77	5.24	9.02	6.74	5.70	32.47	21.27	30.00	10.83	4.17
	+*SFT*	86.25	2.89	2.22	4.89	3.52	2.44	15.96	30.65	47.50	30.00	9.17
	+*RFT*	**88.33**	5.97	5.49	9.05	6.84	6.03	33.38	35.77	53.33	34.17	15.83
	EventVAULT (*Ours*)	87.92	**7.68**	5.88	**9.34**	5.77	5.18	33.85	**60.02**	**80.83**	**60.83**	**47.50**
ECVA	Qwen2-VL-2B	–	–	–	–	–	–	–	0.00	0.00	0.00	0.00
	Qwen2.5-VL-7B	–	–	–	–	–	–	–	19.85	25.87	15.17	9.70
	Qwen2.5-VL-3B	–	–	–	–	–	–	–	14.21	17.16	6.47	3.23
	+*SFT*	–	–	–	–	–	–	–	**45.30**	**66.67**	**49.75**	**24.13**
	+*RFT*	–	–	–	–	–	–	–	35.09	49.00	28.86	19.40
	EventVAULT (*Ours*)	**93.84**	**4.74**	**3.78**	**8.83**	**3.91**	**3.03**	**24.30**	35.38	45.66	38.13	23.74
UCF-Crime	InternVL2.5-2B	84.86	4.40	3.08	8.09	5.69	3.47	24.74	–	–	–	–
	Qwen2.5-VL-7B	92.03	4.80	**3.73**	**8.95**	**7.05**	**4.25**	**28.78**	**22.72**	**33.87**	**16.13**	**8.06**
	InternVL2.5-8B-MPO	89.64	3.79	3.20	8.23	5.77	3.48	24.47	–	–	–	–
	Qwen2-VL-2B	87.25	3.47	2.48	7.75	4.49	2.82	21.02	2.74	4.84	0.00	0.00
	+*SFT*	83.67	3.61	2.26	7.30	4.79	2.70	20.66	–	–	–	–
	+*RFT*	88.45	4.04	2.75	7.72	4.89	3.11	22.52	–	–	–	–
	Qwen2.5-VL-3B	91.63	4.31	2.88	8.70	5.95	3.27	25.10	10.91	15.32	6.45	3.23
	+*SFT*	90.84	1.80	1.01	4.15	2.82	1.11	10.89	4.98	3.23	0.81	0.00
	+*RFT*	92.03	4.42	2.98	8.71	5.98	3.39	25.49	16.80	23.39	8.06	4.03
	EventVAULT (*Ours*)	**97.21**	**5.90**	2.78	8.89	3.67	2.47	23.71	21.28	28.91	10.16	3.91

F Generalization Analysis of EventVAULT

To examine whether EventVAULT overfits to the low-light characteristics of VAULT, we further evaluate the trained model on three external anomaly benchmarks, including MSAD, ECVA, and UCF-Crime, without any dataset-specific fine-tuning. Unlike VAULT, these datasets are not designed specifically for low-light video understanding and are dominated by general illumination conditions. They therefore provide a complementary testbed for evaluating whether the proposed RGB–event modeling can generalize beyond the low-light regime.

To ensure a strictly disjoint cross-dataset evaluation, all videos used for generalization testing on MSAD, ECVA, and UCF-Crime are excluded from the VAULT training set, and no source-level video overlap exists between the VAULT training samples and the external evaluation sets. Following the evaluation protocol of VAU-Bench, we compare against InternVL2.5-2B (Chen et al., 2024b), InternVL2.5-8B-MPO (Chen et al., 2024b; Wang et al., 2024b), Qwen2-VL-2B (Wang et al., 2024a), Qwen2.5-VL-3B (Bai et al., 2025b), and Qwen2.5-VL-7B (Bai et al., 2025b), together with their reported SFT/RFT variants where available.

As shown in Table 6, EventVAULT maintains strong performance after being directly transferred from VAULT to datasets with substantially different scene distributions and predominantly normal illumination conditions. The most notable improvement appears on MSAD, where EventVAULT reaches 60.02% mIoU, clearly exceeding the strongest reported baseline of 45.90%. The advantage becomes even larger under stricter localization criteria, with R@0.7 increasing from 21.67% to 47.50%. This suggests that the temporal evidence learned by EventVAULT is not tied to low-light degradation itself, but remains useful for identifying precise anomaly intervals under more general visual conditions.

The same trend can also be observed in semantic understanding. On UCF-Crime, EventVAULT achieves a QA accuracy of 97.21%, improving over the best reported baseline of 92.03% by 5.18

percentage points, while maintaining competitive temporal grounding performance with 21.28% mIoU. On ECVA, the model similarly achieves 93.84% QA accuracy and 35.38% mIoU without any adaptation to the target dataset. These results indicate that the representations learned on VAULT transfer not only across illumination conditions, but also across different anomaly categories and scene configurations.

Overall, the cross-dataset results suggest that EventVAULT does not merely learn a low-light-specific compensation strategy. Instead, its event-guided temporal allocation and dynamic representation provide complementary motion cues that remain effective when RGB appearance becomes more reliable under normal illumination. This generalization supports the view that the proposed framework captures transferable anomaly dynamics rather than exploiting dataset-specific low-light characteristics.

G Limitations and Future Work

This work establishes a dedicated benchmark and an event-guided framework for studying fine-grained VAU under low-light conditions, while also revealing the substantial impact of illumination degradation on current models. Despite these findings, VAULT is still limited by the practical difficulty of collecting the same real-world anomalous event under multiple controlled illumination conditions. Consequently, samples across different illumination levels may also differ in scene content and event characteristics, making it difficult to isolate illumination as the only varying factor. Future work may therefore explore paired or controlled multi-illumination data collection to enable more rigorous analysis of illumination-induced degradation. It could also investigate more effective multimodal sensing and reasoning strategies to recover anomaly evidence that becomes unreliable or unobservable in conventional RGB videos.