# FineVAU: A Novel Human-Aligned Benchmark for Fine-Grained Video Anomaly Understanding

**João Pereira**[1,2,4], **Vasco Lopes**[1,2,3], **João Neves**[1,3], **David Semedo**[1,4]
[1]NOVA LINCS, Lisboa
[2]DeepNeuronic, Covilhã
[3]University of Beira Interior, Covilhã
[4]NOVA FCT, Lisboa
jaca.pereira@campus.fct.unl.pt, vasco.lopes@deepneuronic.com, jcneves@ubi.pt, df.semedo@fct.unl.pt

## Abstract

Video Anomaly Understanding (VAU) is a novel task focused on describing unusual occurrences in videos. Despite its growing interest, the evaluation of VAU remains an open challenge. Existing benchmarks rely on n-gram-based metrics (e.g., BLEU, ROUGE-L) or LLM-based evaluation. The first fails to capture the rich, free-form, and visually grounded nature of LVLM responses, while the latter focuses on assessing language quality over factual relevance, often leading to subjective judgments misaligned with human perception. In this work, we address this issue by proposing FineVAU, a new benchmark for VAU that shifts the focus towards rich, fine-grained and domain-specific understanding of anomalous videos. We formulate VAU as a three-fold problem, with the goal of comprehensively understanding key descriptive elements of anomalies in video: events (*What*), participating entities (*Who*) and location (*Where*). Our benchmark introduces a) FV-Score, a novel, human-aligned evaluation metric that assesses the presence of critical visual elements in LVLM answers, providing interpretable, fine-grained feedback; and b) Fine$W^3$, a novel, comprehensive dataset curated through a structured and fully automatic procedure that augments existing human annotations with high quality, fine-grained visual information. Human evaluation reveals that our proposed metric has a superior alignment with human perception of anomalies in comparison to current approaches. Detailed experiments on FineVAU unveil critical limitations in LVLM's ability to perceive anomalous events that require spatial and fine-grained temporal understanding, despite strong performance on coarse grain, static information, and events that typically comprise strong visual cues [1].

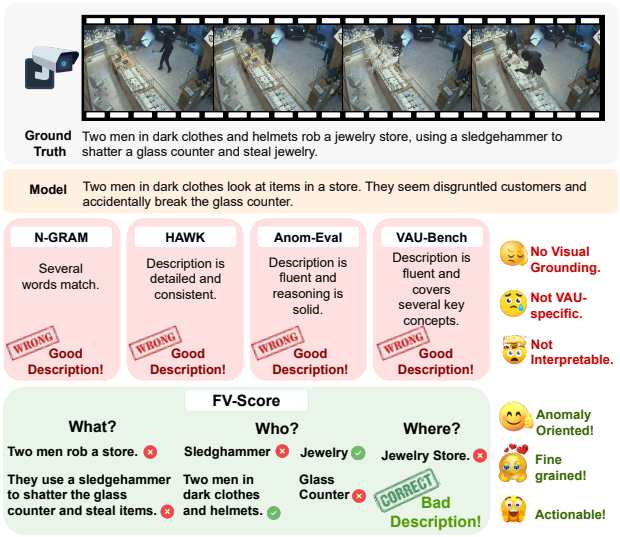


Figure 1: **Evaluation in Video Anomaly Understanding.** The performance of video anomaly understanding models is commonly assessed with metrics that a) disregard semantic equivalence over lexical overlap; b) attribute a large focus over language-specific criteria, such as fluency, consistency and attention to detail; and c) provide non-interpretable scores based on vague and complex scales. In this example, the model answer does not provide a good description of the actual events in the video, but existing metrics will fail to signal this. In contrast, our metric can correctly classify the description as incorrect due to decomposition into the key elements humans rely on to perceive anomalies.

## Introduction

The ability to automatically and robustly detect anomalies in video footage has become increasingly critical across a wide range of applications, from public safety to infrastructure monitoring. As video content continues to grow in scale and diversity, there is a pressing demand for systems that can robustly process this data and identify unusual or suspicious events without human intervention. While early anomaly detection systems have shown proficiency in classifying and localizing predefined sets of anomalies (Sultani, Chen, and Shah 2018; Wu, Su et al. 2025; Ye, Liu, and He 2025; Shao et al. 2025), realistic Video Anomaly Understanding (VAU), which entails a deep understanding of the nuances of abnormal events and the underlying scene, remains an open challenge. The recent emergence of Large Vision-Language Models (LVLMs) (Liu et al. 2023; Bai et al. 2025; Li et al. 2024; Zhang et al. 2025a; Zhu et al. 2025a) and their strong generalization capabilities for diverse vision tasks, have inspired significant advancements



[1]Dataset and source-code will be made public available upon publication.

in VAU, enabling the shift towards more expressive and informative understanding tasks, such as dense video captioning (Yuan, Zhang, and Liu 2024; Zhang et al. 2025b; Tang et al. 2024), video question answering (Liu et al. 2025; Tang et al. 2024; Zhu et al. 2025b), and chain-of-thought (CoT) reasoning (Du et al. 2024b,a; Zhu et al. 2025b).

Despite rapid progress, current VAU reference benchmarks largely overlook evaluation, adopting inadequate metrics, weakly correlated with human judgments, hindering the accurate assessment of the true capabilities of proposed models. These can be split into two categories: 1) traditional n-gram based metrics (e.g., BLEU, ROUGE-L) (Papineni et al. 2002; Lin 2004; Banerjee and Lavie 2005; Vedantam, Zitnick, and Parikh 2015), which measure lexical overlap rather than factual accuracy or contextual understanding, and thus are inherently ill suited for free-form outputs of modern LVLMs; and 2) LLM-based metrics (Du et al. 2024b,a; Tang et al. 2024; Zhu et al. 2025b), often directly adopted from general video understanding tasks (Maaz et al. 2023), which focus on textual fluency and overall coherent reasoning capabilities, lacking the necessary granularity to pinpoint VAU-specific aspects, resulting in subjective scores that are misaligned with human perception of anomalies.

To address this pressing gap in VAU evaluation, we propose FineVAU, a novel automatic and highly human correlated benchmark, that drives the focus towards a rich, fine-grained and domain-specific understanding of anomalies in videos, covering key aspects of human anomaly perception. Namely, by identifying the key structural characteristics of a video anomaly, we formulate FineVAU as a three-perspective problem, comprising comprehensive understanding of 1) events (***What?***), 2) entities (***Who?***), and 3) location information (***Where?***) from anomaly videos (see Figure 1). Grasping these perspectives is key to enable effective and coherent model reasoning about the existence of an anomaly in video.

We enable our structured evaluation through a novel dataset, curated with a fully automatic LVLM-assisted pipeline that systematically decomposes and structures existing human-labeled anomaly description annotations into high-quality knowledge, carefully determining the *What, Who, Where* dimensions. Leveraging this anomaly structuring, we propose FV-Score, a novel LLM-based metric that frames evaluation as a multi-part detection problem, with the goal of extracting the *What, Who and Where*, from LVLMs' reasoning and generated responses. In this setting, FV-Score brings three key properties: a) it breaks anomaly video evaluation in individual dimensions, providing finer-grained, structured and explainable signals, b) achieves strong correlation with human annotations, and c) pushes LVLMs to identify anomalies in video by reasoning and grounding responses across these three dimensions.

Our experiments underscore the advantages of FV-Score's nuanced and fine-grained feedback, unveiling that state-of-the-art LVLMs struggle to report and perceive anomalous events that lack strong visual cues, despite a more accurate understanding of static entities and scene elements. Consequently, our benchmark represents a new, challenging frontier for human-aligned VAU.

In summary, our main contributions are as follows:

- **FineVAU** , a novel benchmark for Video Anomaly Understanding (VAU) that emphasizes fine-grained, human-aligned evaluation grounded in the core components of anomaly comprehension: events (*What*), entities (*Who*), and location (*Where*).
- **FV-Score**, an LLM-based metric that performs key element detection on LVLM answers, providing interpretable and actionable feedback that is tightly aligned with human perception
- **Fine**$W^3$, a high-quality dataset that enriches existing high quality anomaly video annotations with *What, Who, Where* information through a systematic and scalable augmentation pipeline leveraging LVLMs.
- Extensive experiments across diverse LVLMs demonstrate the importance of our evaluation, revealing critical blind spots in current models' ability to capture complex and subtle anomalies.

## Related Work

**Video Anomaly Detection.** Early works primarily focus on Video Anomaly Detection (VAD), framing it as a video-level classification problem (e.g., Shoplifting, Robbery) or localizing abnormal frames (Sultani, Chen, and Shah 2018; Wu et al. 2020). Despite providing coarse-grain anomaly signals, these methods offer a broad high level understanding of the video, limiting actionable insights into the nature or context of the anomaly. We address the more challenging task of VAU, requiring fine-grained understanding about the core elements of anomaly videos.

**Video Anomaly Understanding.** More recently, the advent of Large Vision-Language Models (LVLMs) (Zhang et al. 2024b; Bai et al. 2025; Zhu et al. 2025a; Zhang et al. 2025a) has allowed for rapid progress towards VAU. The pioneer work in UCA (Yuan, Zhang, and Liu 2024) introduces dense human-labeled captions to describe the events of videos in the popular UCF-Crime (Sultani, Chen, and Shah 2018) dataset. HAWK (Tang et al. 2024) proposes sets of synthetically generated video descriptions and question-answer pairs. Similarly, Holmes-VAU (Zhang et al. 2025b) proposes anomaly video descriptions at clip, event and video-level. More recently, ECVA (Du et al. 2024a) (originally CUVA (Du et al. 2024b)) introduces manually annotated anomaly reasoning data, covering the cause, description and result of abnormal events. VAU-Bench (Zhu et al. 2025b) proposes synthetic Chain-of-Thought (CoT) reasoning for anomaly, explaining events by analyzing causal factors, temporal dynamics, and contextual cues. These works lack rich object and scene dimension information, which are tightly coupled with the nature of the anomaly, and rely on synthetic annotations containing hallucinations. We leverage high-quality human-labeled annotations and augment them with rich, verifiable information covering three fundamental anomaly understanding dimensions.

**VAU Evaluation.** Current VAU evaluation methods, largely borrowed or adapted from general-purpose scenarios, and suffer from critical limitations. N-gram-based met-

| Benchmark | Metric | Criteria | Focus |
|---|---|---|---|
| UCA (Yuan, Zhang, and Liu 2024) | N-Gram | Lexical Overlap | Language |
| HIVAU (Zhang et al. 2025b) | N-Gram | Lexical Overlap | Language |
| Hawk (Tang et al. 2024) | Both | Lexical§, Detail, Reasonability, Consistency | Language & Anomaly Description |
| SurveillanceVQA (Liu et al. 2025) | LLM | CI, DO, CU, TU, † | Language |
| AnomEval (Du et al. 2024b) | LLM | Basic Reasoning, Consistency, Hallucination | Language & Anomaly Causality |
| VAU-EVAL | LLM | CA, KC, Fl, In, FC ‡ | Language & Anomaly Reasoning |
| **FV-Score (Ours)** | **LLM** | **Events, Entities, Location** | **Human-aligned Anomaly Perception** |

Table 1: **Comparison with evaluation metrics from SOTA VAU benchmarks.** Unlike current metrics, which focus on lexical overlap (n-gram based) or on textual fluency and reasoning capabilities (LLM-based), our metric assesses fine-grained understanding of anomaly-specific elements according to human perception: events (***What?***), entities (***Who?***) and location (***Where?***). The following abbreviations are used in the table: § Lexical Overlap; † CI: Contextual Integration; DO: Detail Orientation; CU: Contextual Understanding; TU: Temporal Understanding; ‡ CA: Classification Accuracy; KC: Key Concept Alignment; Fl: Linguistic Fluency; In: Informativeness; FC: Factual Consistency.

rics (e.g., BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), CIDEr (Vedantam, Zitnick, and Parikh 2015)), used in UCA, HolmesVAU and HAWK, measure direct lexical overlap between a reference and a predicted caption, thus failing to accurately capture the inherent intricacies of anomalies in free-form responses and reasoning traces, and penalizing factually correct but lexically divergent answers. Addressing these, LLM-based judges have been proposed. AnomEVAL (Du et al. 2024a), adopted in the ECVA benchmark, focuses on assessing causal reasoning by evaluating a model's ability to understand the cause and result of anomalies. However, the scores provided lack fine-grained grounding to anomaly characteristics. SurveillanceVQA-589K (SurvVQA) (Liu et al. 2025) employs a multi-dimensional evaluation protocol to assess multiple criteria, but relies on subjective judgments of correctness across broad categories, rather than a verifiable detection of specific visual elements crucial for VAU. Finally, VAU-Eval (Zhu et al. 2025b), introduced in VAU-Bench, assesses a model's reasoning capabilities against structured question-answering and rationale annotations. Its focus on language-specific aspects and holistic criteria are insufficiently granular to determine accurate perception of anomaly-specific information. We address this by proposing FineVAU-Judge, covering fine-grained anomaly dimensions: *What, Who, Where*.

## Fine-Grained Video Anomaly Understanding

With FineVAU, we formulate VAU as the goal of comprehensively understanding the key structure of anomaly videos according to human perception of anomalies, grounded in an hierarchy composed by three main structural dimensions: events (*What*), involved entities and their attributes (*Who*) and location (*Where*).

### Problem Formulation

Let $V = \{f_1, f_2, \dots, f_T\}$ be an untrimmed video, represented as a sequence of $T$ frames. The goal of a VAU model $M$ is to generate a natural language report of the video $R = M(V)$. We define a structured ground truth, $G$, for each video $V$. This ground truth is a set of fundamental anomaly elements, partitioned according to the three hierarchical dimensions:

1. **What (Events):** $G_{\text{what}} = \{e_1, e_2, \dots, e_{N_e}\}$, a set of $N_e$ textual descriptions capturing the key actions (e.g., "sets fire"), interactions (e.g., "fighting") and isolated state changes (e.g., "explosion") occurring in the video.
2. **Who (Entities):** $G_{\text{who}} = \{w_1, w_2, \dots, w_{N_w}\}$, a set of $N_w$ concise textual descriptions of the involved actors or objects, including their salient visual attributes (e.g., clothing, color, age group).
3. **Where (Location):** $G_{\text{where}} = \{l_1, l_2, \dots, l_{N_l}\}$, a set of $N_l$ attributes detailing the scene where the events unfold.

The complete ground truth of a video is the union of these sets: $G = G_{\text{what}} \cup G_{\text{who}} \cup G_{\text{where}}$. The quality of a model's report $R$ is measured by its coverage of the critical elements of the ground truth $G$. To measure this coverage, we replace set hard membership in $G$ by a semantic-aware membership function $m_\theta(g, G_*)$, with parameters $\theta$, $g$ being a ground-truth element, and $G_*$ the reference set. Then, we define a structural scoring function,

$$\mathcal{J}_{\text{dim}}(R) = \sum_{g_i \in G_{\text{dim}}} m_\theta(g_i, G_{\text{dim}}), \tag{1}$$

where $dim \in \{what, who, where\}$, which scores the presence and correct mention of each ground truth element $g_i \in G_{\text{dim}}$ in the generated report $R$, and calculates the total score. The membership function $m_\theta$, returns a positive score when semantic membership is attested, or 0 otherwise. Given the central role of the *what* dimension (what action/situation is taking place), we consider two membership degrees.

Unlike previous works, which encompass complex scales and broad criteria, we argue for straightforward scoring instructions, simplifying the task for an LLM judge and increasing interpretability. Thus, we use a binary scale for *Who* and *Where*, and a ternary scale for the *What* dimensions:

| Binary (*Who*, *Where*) | Ternary (*What*) |
|---|---|
| 0 ← Missing / incorrect | 0 ← Missing / incorrect |
| 1 ← Present, correct | 1/2 ← Partial, minor errors |
| | 1 ← Accurate, complete |

The ternary scale provides flexibility to deal with scenarios where the answer partially covers the elements of the ground truth (e.g., $R$ contains *"Two people are having an heated discussion."* and $g$ is *"Two men are fighting."*). The membership function $m_\theta$ strictly follows the scale in the aforementioned table.

The overall quality of the report $R$ is then quantified by a scoring function $\mathcal{S}(R)$, which separately aggregates scores across the three dimensions of ground truth elements:

$$\mathcal{S}(R) = \lambda_{\text{what}} \cdot \mathcal{J}_{\text{what}}(R) + \lambda_{\text{who}} \cdot \mathcal{J}_{\text{who}}(R) + \lambda_{\text{where}} \cdot \mathcal{J}_{\text{where}}(R) \quad (2)$$

where $\lambda_{\text{what}}, \lambda_{\text{who}}, \lambda_{\text{where}}$ are weights controlling the relative importance of each dimension. FineVAU defines VAU as the goal of generating a natural language report $R$, for a given video $V$, that maximizes the score $\mathcal{S}(R)$ with respect to its ground truth $G$.

### FV-Score and FineVAU-Judge

We define $\mathcal{S}(R)$ as **FV-Score** and $\mathcal{J}$ as **FineVAU-Judge**, a LLM judge which given an a dimension ground truth set $G_{\text{dim}}$, and the report $R$ for a video, attests semantic membership $m_\theta$ for each ground truth element $g \in G$.

Following MovieChat and VideoChatGPT, two well-established video understanding benchmarks, we materialize the membership function $m_\theta$ using a frontier LLM model (Gemini) that allows judgments using a triplet of $\{R, G, P\}$, where $P$ is a highly detailed prompt that instructs the model to provide a structured output with the membership scores for each ground truth element $g \in G$, according to a LVLM model output report $R$, and the membership scale defined in the previous section. We detail $P$ in the supplementary material.

## *What, Who, Where* Dataset

We build the ground truth $G$ of our FineVAU benchmark by curating ***W**hat, **W**ho, **W**here* (Fine$W^3$), a novel dataset containing fine-grained and structured information from anomaly videos.

### Structured Annotation Scheme

The core novelty of our dataset is a comprehensive and structured annotation scheme that maps human perception of anomaly videos, according to the three dimensions defined in FineVAU.

At the *What* dimension, the dataset captures key actions, interactions and isolated state changes in the video as a sequence of chain of discrete, atomic or highly correlated events. The *Who* dimension, provides detailed information regarding the entities involved in the events. Each entity is assigned a unique *identifier* that concisely describes it (e.g., "assailant"). Concurrently, event descriptions (e.g., "The man holds a gun...") are explicitly linked to the unique identifiers of all the entities involved. Additionally, each entity contains information about its *category* (e.g., "person", "vehicle"), and a rich set of observable, category-specific *visual attributes*: *clothing*, approximate *age group*, *gender*

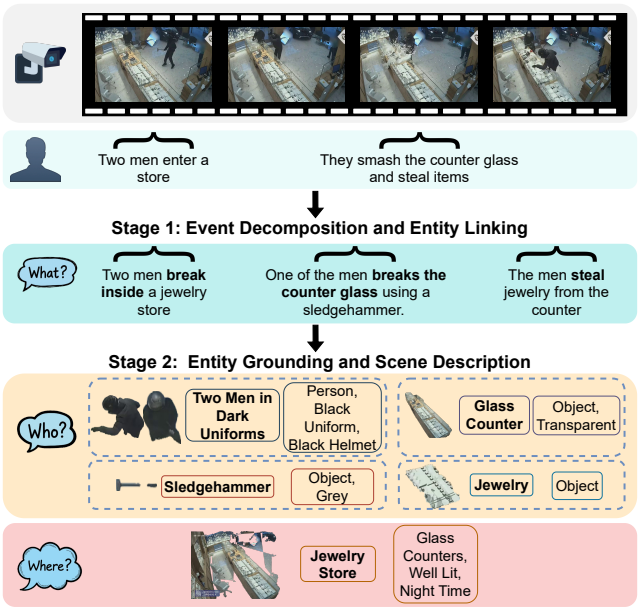


Figure 2: **Our two-stage, fully automated pipeline for scalable annotation of fine-grained VAU data.** We ground our annotation process on existing, high-quality human annotations of anomaly video events, and use a LVLM to: 1) augment and refine existing events and identify the entities involved; an 2) augment entity and location information with descriptive information regarding their physical attributes.

and *distinguishing feature* (e.g., "a large beard") for a person; *size*, *color* and *brand* for a vehicle; and at least *color* and *size* for a general object. Finally, at the (*Where*) dimension, the dataset depicts information regarding the location where events take place, including the physical *environment* (e.g., "jewelry store"), *time of day* (e.g., night time), *lighting conditions*, *crowd density* and a *salient feature* that uniquely identifies the scene (e.g., "a large painting on the wall").

### Annotation Pipeline

Our dataset requires a high level of granularity in information that is not present in current VAU datasets. Therefore, we develop a fully automated and scalable pipeline to enrich and augment existing, human-labeled VAU data with high quality, fine-grained and structured anomaly-oriented information. We split the annotation process into two stages, as depicted in Fig. 2, and leverage assistance from a LVLM.

**Stage 1: Event Decomposition and Entity Linking.** The first stage of our pipeline augments and refines existing event annotations from UCA (Yuan, Zhang, and Liu 2024). An LVLM processes raw, human-generated event descriptions to 1) decompose complex sentences into a chain of fine-grained, causally linked atomic events; 2) complement annotations by identifying unmentioned events or objects that are clearly observable in the video; and 3) identify all participating entities for each event and assign concise identifiers, which are then explicitly linked to the respective event.

| Metric | PCC $\rho \uparrow$ | 1-R$^2$ $\downarrow$ | Kd $\tau \uparrow$ | Sp $\tau \uparrow$ |
|---|---|---|---|---|
| *N-gram Baselines* | | | | |
| CIDEr | -0.63 | 0.60 | -0.59 | -0.58 |
| BLEU† | 0.19 | 0.96 | 0.17 | 0.17 |
| METEOR | 0.45 | 0.80 | 0.41 | 0.40 |
| ROUGE-L | 0.47 | 0.78 | 0.43 | 0.44 |
| *LLM Judge Baselines* | | | | |
| AnomEVAL (Du et al. 2024a) | 0.42 | 0.82 | 0.39 | 0.37 |
| VAU-EVAL (Zhu et al. 2025b) | 0.53 | 0.72 | 0.49 | 0.47 |
| **FV-Score** | **0.61** | **0.63** | **0.56** | **0.56** |

Table 2: **Correlation of VAU evaluation metrics with human judgment.** † BLEU is the mean score of BLEU-{1 to 4}.

**Stage 2: Entity Grounding and Scene Description.** This stage builds on the output of the first stage. We leverage the LVLM to 1) augment linked entities with rich and fine-grained information regarding their observable physical attributes; and 2) identify and describe the physical properties of the location where the events take place. Further information regarding the prompts for the LVLM and resulting annotations can be seen in the supplementary material.

## Dataset Statistics

Our dataset contains a total of 1544 videos. These videos contain a total of 17813 events, from which 13393 are normal and 4420 are abnormal. These events are associated with a total of 59392 entities, which in turn reference a total of 74593 individual attributes. Finally, there are 7669 annotated location attributes. Figure 3 shows the distribution of the number of annotations per video, for all annotation dimensions. At the event dimension, we distinguish abnormal and normal events. As expected, there is a much larger number of normal events per video, since abnormal events usually occur infrequently and in small temporal windows. Figure 4 plots the word cloud of event annotations, showing a clear trend towards anomaly and movement related topics. Figure 5 shows the distribution of the duration of videos in our dataset, originally sourced from CCTV footage, often containing several minutes or even hours of video. There is also a broad range of possible locations, including public streets, highways, shopping centers or private households. Depending on the location and the time of day, there may be large amounts of entities performing a wide variety of different actions (e.g., people walking in a crowded street, customers shopping in a convenience store with several items). The combination of these factors, coupled with our fine-grained approach to VAU, results in a densely annotated, extremely challenging benchmark that enables the true assessment of the capability of LVLMs to fully grasp the complexity of anomalies, establishing a new frontier for VAU.

# Assessing FV-Score's Human Correlation

We conduct a comprehensive human evaluation study to validate the alignment of FV-Score with human perception of anomaly report quality. Our study utilizes a set of 60 videos randomly sampled from the UCF-Crime dataset (Sultani, Chen, and Shah 2018). For each video, 8 human experts rank three video reports of varying quality generated by Gemini 2.5-flash (Gemini et al. 2024), simulating answers from different LVLM: a *high-quality* report covering all critical information of the video; a *medium-quality* report that omits some key information; and a *low-quality* report that fails to describe the anomalies in the video. To mitigate bias, we present the reports in a randomized order without disclosing their quality level, and ensure each video is evaluated by three different experts. The study achieves a Pairwise Percentage Agreement score of 68%, indicating substantial inter-annotator agreement, particularly given the subjective nature of the task. To measure the correlation between state-of-the-art VAU evaluation metrics and human rankings, we employ four standard agreement measures (Dong et al. 2024). We use Pearson Correlation Coefficient (PCC $\rho$) to assess the linear relationship between metric scores and human scores, and $1 - R^2$ to measure the unexplained variance, where lower values signify a better fit. To evaluate ordinal association, we use both Kendall's Tau (Kd $\tau$) and Spearman's Rho (Sp $\tau$), which measure the agreement in the rankings produced by the metric and by human judges. Results are available in Table 2. Our study does not consider the metric used in SurveillanceVQA (Liu et al. 2025) nor HAWK (Tang et al. 2024), as they are strictly defined for QA tasks.

Evaluation reveals that FV-Score achieves a superior alignment with human judgment over all baselines. FV-Score records the highest scores across all correlation measures (with exception to CIDEr in unexplained variance), achieving a Pearson correlation of **0.61** and a Kendall's Tau of **0.56**. This performance marks a clear improvement over the strongest n-gram baseline, ROUGE-L (Lin 2004) (PCC $\rho$ - 0.47, Kd $\tau$ 0.43). Critically, BLEU (Papineni et al. 2002) and CIDEr (Vedantam, Zitnick, and Parikh 2015) show systematic disagreement with humans. Surprisingly, LLM-based metrics perform similarly to their n-gram counterparts, despite utilizing strong GPT-based judges. These findings support the hypothesis that the structured and fine-grained evaluation provided by our metric is better aligned with human perception of anomalies.

| $\lambda_{\text{what}}$ | $\lambda_{\text{who}}$ | $\lambda_{\text{where}}$ | PCC $\rho \uparrow$ | 1-R$^2$ $\downarrow$ | Kd $\tau \uparrow$ | Sp $\tau \uparrow$ |
|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 1.0 | 0.51 | 0.74 | 0.46 | 0.47 |
| 1.0 | 1.0 | 2.0 | 0.47 | 0.77 | 0.42 | 0.42 |
| 2.0 | 1.0 | 1.0 | 0.56 | 0.69 | 0.50 | 0.50 |
| 1.0 | 2.0 | 1.0 | **0.61** | **0.63** | **0.56** | **0.56** |

Table 3: **Ablations on FV-Score weights.** Our ablations reveal that a strong weight for entity components achieves a higher correlation with human judgment. This indicates that human anomaly description preference is highly correlated with accurate perception of involved entities.

**Ablations on $\lambda_{\text{what}}$, $\lambda_{\text{who}}$ and $\lambda_{\text{where}}$.** We conduct ablations on the weights used for the scores of different dimensions in FV-Score (see Table 3). While current metrics often focus mostly on event understanding (*What*), our exper-

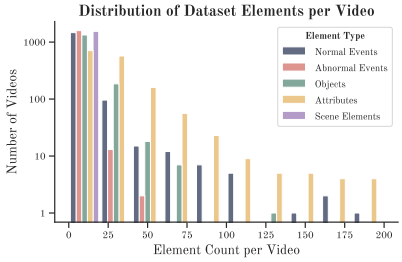


Figure 3: **Annotation count per dimension.** We split events into normal and abnormal, and separate counts of entities and their physical attributes. The large number of annotations stems from the granularity of our annotation schema.

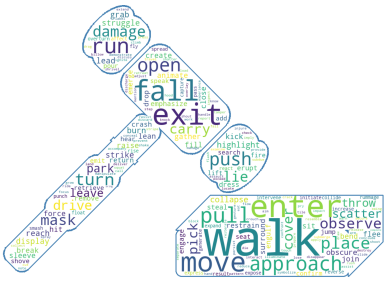


Figure 4: **Word Cloud[2] for event annotations in Fine$W^3$.** Events often describe motion (e.g., "run", "walk"), interaction ("exit", "approach") and abnormality ("fall", "damage", "struggle", "grab", "kick")

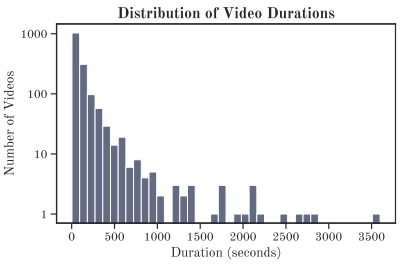


Figure 5: **Histogram of Video Duration.** Our dataset contains challenging long videos with up to 1h duration. This is expected given the source of the videos of our dataset, composed of CCTV footage.

| | Overall Performance | | | | | Location | | | | | Entity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Location** | **Event** | **Entity** | **Attribute** | **All** | **Lighting** | **Env** | **Crowd** | **Time** | **Salient** | **Person** | **Vehicle** | **Others** |
| VideoLLaMA3 | 40.3 | 6.5 | 24.3 | 10.2 | 19.3 | 44.1 | 64.7 | 30.0 | 35.4 | 27.4 | 20.8 | 22.2 | 27.5 |
| LLaVA-OV | 58.3 | 13.0 | 41.1 | 19.9 | 32.2 | 65.1 | 80.1 | 42.2 | 60.0 | 44.1 | 38.5 | 37.6 | 44.0 |
| Qwen2.5-VL | 70.8 | 9.1 | 38.3 | 20.3 | 32.9 | 80.2 | 83.6 | **68.0** | **80.7** | 41.8 | 29.6 | 37.9 | 44.5 |
| LLaVA-VID | 65.7 | 14.4 | 44.0 | 21.0 | 35.0 | 65.1 | **87.0** | 56.8 | 69.0 | 50.8 | 42.2 | 38.0 | 47.4 |
| InternVL3 | **71.8** | **18.0** | **51.2** | **25.5** | **40.5** | **80.4** | 86.6 | 59.3 | 79.7 | **53.1** | **54.0** | **44.8** | **51.5** |
| Mean | 61.3 | 12.2 | 39.8 | 19.4 | 32.0 | 67.0 | 80.4 | 51.3 | 65.0 | 43.4 | 37.0 | 36.1 | 43.0 |

Table 4: **Results of SOTA LVLMs on FineVAU** reveal clear difficulties of LVLMs in our benchmark.

iments demonstrate that humans highly value reports that correctly identify the main entities involved and accurately describe them (*Who*). This finding further supports the validity of our fine-grained and structured approach to VAU.

## Experiments and Results

### Experimental Setup

We now leverage FineVAU to evaluate five state-of-the-art, open-source and mid-sized LVLMs, namely Qwen2.5-VL-7B (Bai et al. 2025), InternVL3-9B (Zhu et al. 2025a), VideoLLaMA3-7B (Zhang et al. 2025a), LLaVA-Video-7B (Zhang et al. 2024b) and LLaVA-OneVision-7B (Li et al. 2024). To facilitate reproducibility, we adopt the lmms-eval (Zhang et al. 2024a) platform, widely adopted in multimodal understanding benchmarks. All our experiments are done in a zero-shot setting, using the original model weights, and a model temperature of 0. All prompts used are provided in the supplementary material.

### Results

Table 4 presents the performance of five LVLMs on FineVAU, revealing several important takeaways.

**LVLMs are stronger at perceiving static and coarse grain information.** LVLMs show significantly better performance at reporting location information, with a mean accuracy of 61.3%. We hypothesize that the strong image understanding pretraining of LVLMs (which comprise an image understanding vision encoder) results in a strong capability at grounding static and coarse grain information. This hypothesis is further sustained by their performance at Entity dimension. Despite a low Entity Mean performance of 39.8%, it still surpasses Event and Attribute dimensions. A closer look at performance for individual location attributes evidences once again a similar pattern, since models excel at identifying the physical environment of the videos, and are also capable of perceiving lighting conditions, time of day, and crowd density. At the entity dimension, comprehensive identification of vehicles and people is surprisingly tougher for LVLMs in comparison to other categories. This is likely due to the fact that people and vehicles are the most common objects and are usually more predominant, making them more challenging to report in detail, in contrast with other single unit (e.g., buildings, weapons) and low frequency (e.g., ATMs, animals) categories.

**LVLMs struggle with spatial and temporal fine-grained understanding.** Unlike for static and coarse grain information, LVLMs struggle significantly on fine-grained understanding, both on the spatial and temporal axis. This is sustained by the low accuracy on individual object attributes, contrasting with coarse grain object identification. We argue that this difficulty is grounded on the training bias of

[2]The gavel in the image is a reference to our FineVAU-Judge.

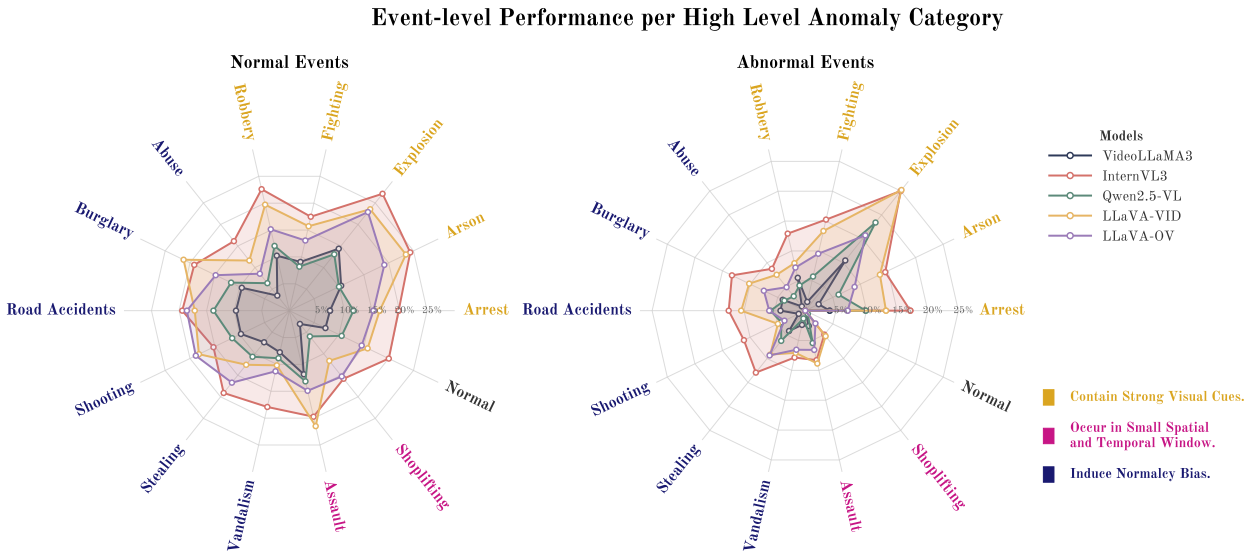


Figure 6: **LVLM's performance breakdown at event dimension.** Performance per high level anomaly category, summarizing the nature of the events depicted in the video (e.g., *Fighting* contains events depicting the conflict escalation from peaceful coexistence to a physical altercation). Mode details on expected events per category are provided in the supplementary material.

LVLMs, which is largely composed of general videos with high resolution, high quality and less clutter, in contrast with the low resolution and low quality of available anomaly videos. Nonetheless, the major struggle of LVLMs lies on identifying all events in anomaly videos, with models exhibiting a mean accuracy of merely 12.2%. Figure 6 compares the accuracy of LVLMs at the event dimension, according to the high level video category, for both normal (leftmost plot) and abnormal (rightmost plot) events. Noticeably, results once again corroborate our hypothesis that LVLMs perform better in events when strong visual cues are available: *Explosions* and *Arson*, commonly accompanied by flashes of bright light, fire and debris; *Arrests*, which usually involve police officers in characteristic uniforms and vehicles; and *Fights*, which frequently cause high commotion for surrounding individuals. However, anomalies that occur in smaller spatial and temporal windows and require a higher understanding of visual elements and behaviors, are much more challenging. This is the case with *Shoplifting*, which requires understanding sudden behaviors such as placing small items in a bag and leaving without paying.

**LVLMs are biased towards normalcy.** Another noticeable pattern observable in Figure 6 is that despite globally achieving low performance, LVLMs are still more capable at understanding *Normal* events, even in videos that contain abnormalities. We argue that LVLMs are biased towards normalcy, and therefore frequently conflate abnormal events for normal ones (e.g., a fight is depicted as a conversation). This high degree of hallucination is not seen inversely, since LVLMs less frequently hallucinate abnormal events in normal situations, as evidenced by their superior performance in the latter. We hypothesize that the low performance of LVLMs in normal events emerges instead from their inability to recall the high number of events annotated in our dataset. Examples of common LVLM hallucinations and failures to recall granular information can be seen in the supplementary material.

**InternVL3 achieves the top performance across all dimensions.** Noticeably, despite lower context sizes, smaller pretraining corpus and lower scene performance, LLaVA-OneVision (Li et al. 2024) and LLaVA-VID (Zhang et al. 2024b) are more capable of understanding events in comparison to large context and corpus alternatives (Qwen-2.5-VL and VideoLLaMA3). This gap is an additional proof of the critical disconnect between understanding the static context of a video and the anomalous events within it.

## Conclusions and Future Directions

In this work, we introduce FineVAU, a novel benchmark that addresses the critical gap in Video Anomaly Understanding (VAU) evaluation by shifting the focus to a fine-grained and structured assessment of LVLM comprehension across events (*What*), entities (*Who*), and location (*Where*), which are key aspects in human perception of anomalies. Through our proposed FV-Score, supported by an LLM-based FineVAU-Judge, and the Fine$W^3$ dataset, we conduct an extensive evaluation of five LVLMs, unveiling a critical weakness: **While these models are capable of perceiving static scenes and entities, they fundamentally fail to comprehend the fine-grained attribute details and subtle events that occur in small spatial and temporal windows, often hallucinating normalcy**. This crucial finding, made possible by our anomaly structuring and FineVAU, evidences clear next steps towards developing targeted training to mitigate hallucinations and induce detailed, factual understanding, using our structured data. Pairing such data with rigorous benchmarks such as ours is essential for training

and validating the next generation of models capable of truly robust video anomaly understanding.

## References


Bai, S.; et al. 2025. Qwen2.5-vl technical report. *arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Dong, H.; et al. 2024. Benchmarking and improving detail image caption. *arXiv:2405.19092*.

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

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

Gemini, T.; et al. 2024. Gemini: A family of highly capable multimodal models, 2024. *arXiv:2312.11805*.

Li, B.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Liu, B.; Qiao, P.; Ma, M.; Zhang, X.; Tang, Y.; Xu, P.; Liu, K.; and Yuan, T. 2025. SurveillanceVQA-589K: A Benchmark for Comprehensive Surveillance Video-Language Understanding with Large Models. *arXiv preprint arXiv:2505.12589*.

Liu, H.; et al. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Maaz, M.; et al. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the annual meeting of the Association for Computational Linguistics*.

Shao, Y.; He, H.; Li, S.; Chen, S.; Long, X.; Zeng, F.; Fan, Y.; Zhang, M.; Yan, Z.; Ma, A.; et al. 2025. Eventvad: Training-free event-aware video anomaly detection. *arXiv preprint arXiv:2504.13092*.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*.

Wu, P.; Su; et al. 2025. VarCMP: Adapting Cross-Modal Pre-Training Models for Video Anomaly Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ye, M.; Liu, W.; and He, P. 2025. 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*.

Yuan, T.; Zhang, X.; and Liu, K. 2024. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, B.; et al. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv:2501.13106*.

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

Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; and Liu, Z. 2024a. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models.

Zhang, Y.; et al. 2024b. Video instruction tuning with synthetic data. *arXiv:2410.02713*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zhu, L.; Chen, Q.; Shen, X.; and Cun, X. 2025b. VAU-R1: Advancing Video Anomaly Understanding via Reinforcement Fine-Tuning. *arXiv preprint arXiv:2505.23504*.


## Acknowledgments


This work is supported by NOVA LINCS ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020) with the financial support of FCT.IP; and Fundação para a Ciência e Tecnologia ref. 2023.03647.BDANA