

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

Joao Alexandre Carneira Pereira^{1,2,4}, Vasco Lopes^{1,2,3}, João C. Neves^{1,3}, David Semedo^{1,4}

¹NOVA LINCS, Lisboa

²DeepNeuronic, Covilhã

³University of Beira Interior, Covilhã

⁴NOVA 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 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 resulting in subjective judgments that are 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) FineW³, 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 with strong visual cues.

Code and Dataset — <https://finevau.github.io>

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 pressing demand for systems that can robustly process this data and identify unusual or suspicious events without human intervention. While early

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

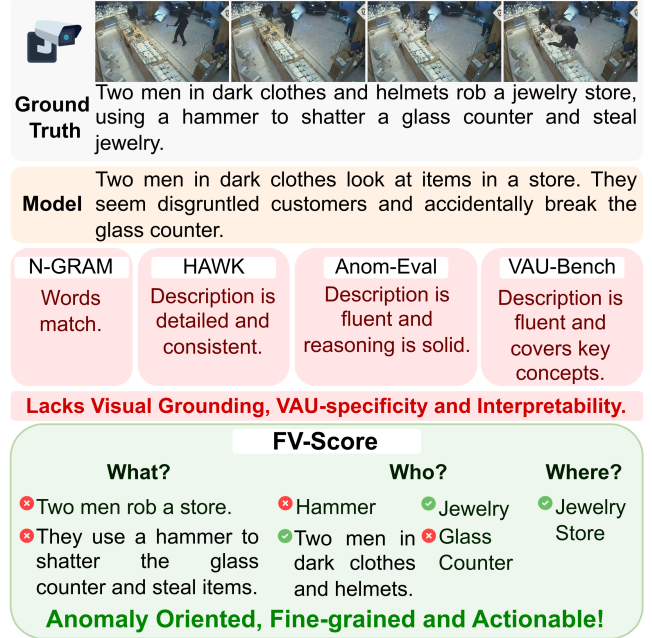


Figure 1: Evaluation in Video Anomaly Understanding. The performance of VAU models is commonly assessed with metrics that a) disregard semantic equivalence; b) focus on language criteria such as fluency and detail; and c) yield non-interpretable, vague scores. Existing metrics fail to signal if the model provides a factually incorrect but coherent description. In contrast, our metric can correctly classify the description as incorrect by decomposing it into the key elements humans rely on to perceive anomalies.

anomaly detection systems have shown proficiency in classifying and localizing predefined sets of anomalies (Sultani, Chen, and Shah 2018; Wu 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,

has inspired significant advancements in VAU, enabling the shift towards more expressive and informative understanding tasks, such as dense video captioning (Yuan et al. 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; Lavie and Agarwal 2007; Vedantam, Lawrence 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. 2024), 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. Conse-

quently, 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
- **FineW³**, 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) by 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. 2025d; Bai et al. 2025; Zhu et al. 2025a; Zhang et al. 2025a) has allowed for rapid progress towards VAU. The pioneering work in UCA (Yuan et al. 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.

Benchmark	Metric	Criteria	Focus
UCA (Yuan et al. 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, FI, In, FC ‡	Language & Anomaly Reasoning
FV-Score (Ours)	LLM	Events, Entities, Location	Human-aligned Anomaly Perception

Table 1: **Comparison with SOTA VAU metrics.** 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: events (**What?**), entities (**Who?**) and location (**Where?**). The following abbreviations are used: § Lexical Overlap; † CI: Contextual Integration; DO: Detail Orientation; CU: Contextual Understanding; TU: Temporal Understanding; ‡ CA: Classification Accuracy; KC: Key Concept Alignment; FI: Linguistic Fluency; In: Informativeness; FC: Factual Consistency.

VAU Evaluation. Current VAU evaluation methods, largely borrowed or adapted from general-purpose scenarios, and suffer from critical limitations. N-gram-based metrics (e.g., BLEU (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)), used in UCA, Holmes-VAU 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 a 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 θ , g being a ground truth element, and G_* the reference set. Then, we define a structural scoring function,

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

where $\text{dim} \in \{\text{what}, \text{who}, \text{where}\}$, 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. 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 (<i>Who</i> , <i>Where</i>)	Ternary (<i>What</i>)
0 $\leftarrow$ Missing / incorrect	0 $\leftarrow$ Missing / incorrect
1 $\leftarrow$ Present, correct	0.5 $\leftarrow$ Partial, minor errors
	1 $\leftarrow$ 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_θ 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 λ_{what} , λ_{who} , λ_{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_{dim} , and the report R for a video, attests semantic membership m_θ for each ground truth element $g \in G$.

Following MovieChat and VideoChatGPT, two well-established video understanding benchmarks, we materialize the membership function m_θ using a frontier LLM model, Gemini-2.5-Flash (Team 2025) 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 *What, Who, Where* (FineW³), 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 the 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 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* 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 provides information regarding the location where events take place, including the physical *environment* (e.g., “jewelry store”), *time of*

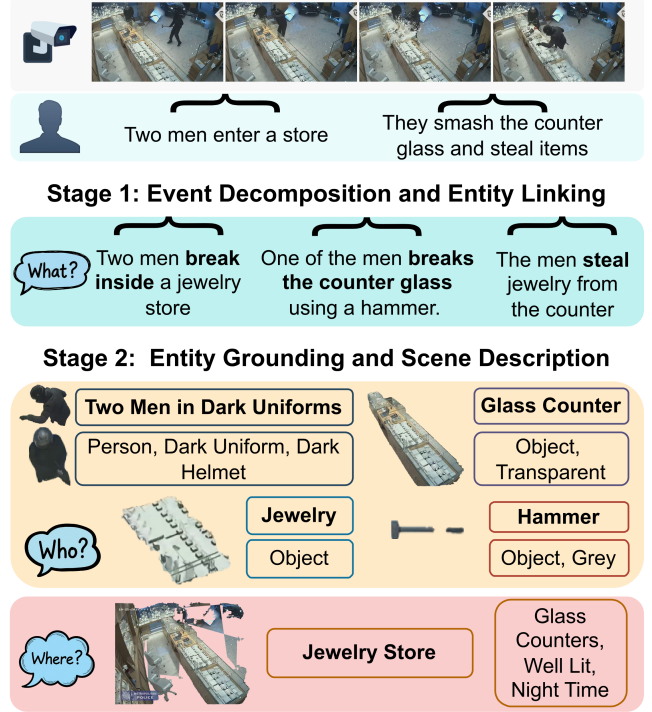


Figure 2: **Our two-stage, fully automated pipeline for scalable annotation of fine-grained VAU data.** We ground our annotation process on high-quality human annotations, and use a LVLM to: 1) augment and refine existing events and identify the entities involved; an 2) augment entity and location information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

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 Figure 2, and leverage assistance from a LVLM. We employ Gemini-2.5-Pro (Team 2025) due to its strong multimodal understanding and long context capability. For each video, we uniformly sample frames at 1 fps and provide them alongside the original UCA annotations to the model.

Stage 1: Event Decomposition and Entity Linking. The first stage of our pipeline augments and refines existing event annotations from UCA (Yuan et al. 2024). An LVLM processes raw, human-generated event descriptions, together with their respective videos, 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

Metric	PCC $\rho \uparrow$	1-R ² $\downarrow$	Kd $\tau \uparrow$	Sp $\tau \uparrow$
<i>N-gram Baselines</i>				
CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)	-0.63	0.60	-0.59	-0.58
BLEU $\dagger$ (Papineni et al. 2002)	0.19	0.96	0.17	0.17
METEOR (Lavie and Agarwal 2007)	0.45	0.80	0.41	0.40
ROUGE-L (Lin 2004)	0.47	0.78	0.43	0.44
<i>LLM Judge Baselines</i>				
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 metrics with human judgment.** $\dagger$ BLEU is the mean score of BLEU- $\{1$ to $4\}$.

event and assign concise identifiers, which are then explicitly linked to the respective event.

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 60 videos randomly sampled from the UCF-Crime dataset (Sultani, Chen, and Shah 2018). We recruit 8 human experts with computer

λ_{what}	λ_{who}	λ_{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** reveal that a strong weight for entity components achieves a higher correlation with human judgment, indicating that humans value an accurate perception of involved entities.

vision or multimodal evaluation background and assign each video to 3 different experts, resulting in 180 total ranking judgments. For each video, experts rank three reports of varying quality generated by Gemini 2.5-flash (Team 2025): 1) a high-quality report covering all critical information; 2) a medium-quality report omitting some key details; and 3) a low-quality report failing to describe the anomalies. To mitigate bias, reports are presented in randomized order without disclosing their quality level. The study achieves a Pairwise Percentage Agreement score of 68%, indicating substantial inter-annotator agreement, particularly given the subjective nature of the task.

We measure the correlation between state-of-the-art VAU evaluation metrics and human rankings by employing four standard agreement measures (Dong et al. 2024). We use Pearson Correlation Coefficient (PCC ρ) 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 τ) and Spearman’s Rho (Sp τ), 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 achieves 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 ρ - 0.47, Kd τ 0.43). Critically, BLEU (Papineni et al. 2002) and CIDEr (Vedantam, Lawrence 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.

Ablations on λ_{what} , λ_{who} and λ_{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 experiments demonstrate that humans highly value reports that correctly identify the main entities involved and accurately

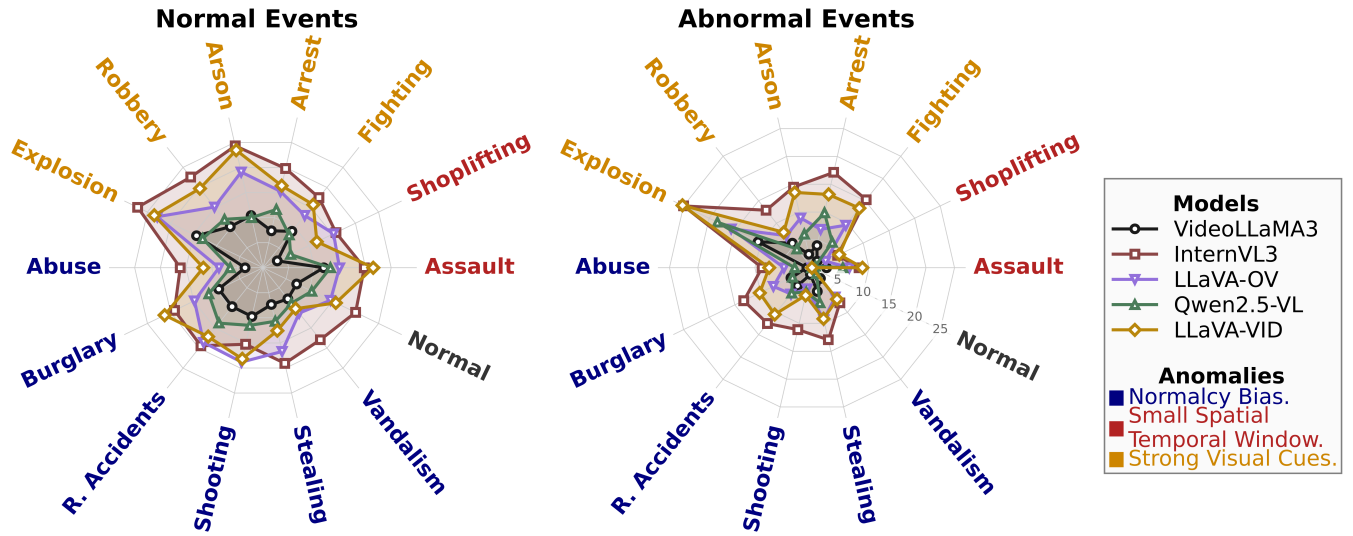


Figure 6: **LVLm’s performance breakdown at event dimension.** Performance per high level anomaly category, summarizing the nature of the events depicted in the video (e.g., *Fighting* contains events depicting the conflict escalation from peaceful coexistence to a physical altercation). More details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

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. 2025d) are more capable of understanding events in comparison to large context and corpus alternatives (Qwen2.5-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 FineW³ dataset, we conduct an extensive evaluation of five LVLms and unveil 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.

Acknowledgments

This work is supported by NOVA LINCS (UID/04516/2025) with the financial support of FCT/IP; and Fundação para a Ciência e Tecnologia ref. 2023.03647.BDANA.

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.
- Dong, H.; Li, J.; Wu, B.; Wang, J.; Zhang, Y.; and Guo, H. 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:2412.07183.
- Du, H.; Zhang, S.; Xie, B.; Nan, G.; Zhang, J.; Xu, J.; Liu, H.; Leng, S.; Liu, J.; Fan, H.; Huang, D.; Feng, J.; Chen, L.; Zhang, C.; Li, X.; Zhang, H.; Chen, J.; Cui, Q.; and Tao, X. 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 (CVPR)*.
- Lavie, A.; and Agarwal, A. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 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:2505.12589.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.
- Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024. 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)*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th 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.; Wang, X.; Tang, H.; Wang, Y.; and Li, S. 2025. EventVAD: Training-Free Event-Aware Video Anomaly Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*.
- Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tang, J.; Lu, H.; Wu, R.; Xu, X.; Ma, K.; Fang, C.; Guo, B.; Lu, J.; Chen, Q.; and Chen, Y.-C. 2024. HAWK: Learning to Understand Open-World Video Anomalies. In *Advances in Neural Information Processing Systems*.
- Team, G. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDER: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; and Yang, Z. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*.
- Wu, P.; Su, W.; He, X.; Wang, P.; and Zhang, Y. 2025. VarCMP: adapting cross-modal pre-training models for video anomaly retrieval. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in 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 (CVPR)*.
- Yuan, T.; Zhang, X.; Liu, K.; Liu, B.; Chen, C.; Jin, J.; and Jiao, Z. 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 (CVPR)*.
- Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; Jin, P.; Zhang, W.; Wang, F.; Bing, L.; and Zhao, D. 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 (CVPR)*.
- Zhang, K.; Li, B.; Zhang, P.; Pu, F.; Cahyono, J. A.; Hu, K.; Liu, S.; Zhang, Y.; Yang, J.; Li, C.; and Liu, Z. 2025c. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*.
- Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2025d. LLaVA-Video: 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.; Gao, Z.; Cui, E.; Wang, X.; Cao, Y.; Liu, Y.; Wei, X.; Zhang, H.; Wang, H.; Xu,

W.; Li, H.; Wang, J.; Deng, N.; Li, S.; He, Y.; Jiang, T.; Luo, J.; Wang, Y.; He, C.; Shi, B.; Zhang, X.; Shao, W.; He, J.; Xiong, Y.; Qu, W.; Sun, P.; Jiao, P.; Lv, H.; Wu, L.; Zhang, K.; Deng, H.; Ge, J.; Chen, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025a. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.

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

In this supplementary material, we provide additional details regarding the prompts utilized in FineVAU (Figures 7, 8, 9 and 11); the annotation structure (Figure 10); the type of abnormal events that are expected from each high level video category; and several qualitative examples of LVLM answers that demonstrate success and failure cases (Figures 12, 13 and 14), supporting our result discussion. Additionally, we externally append the annotations and the videos addressed in this document, in .json and .mp4 files, respectively.

Common Events per Video Category

Abuse

Contains events that involve bad, cruel or violent behavior against children, old people, animals, and women. Also includes police brutality.

Burglary

Contains events that show people (thieves) entering into a building or house with the intention to commit theft.

Robbery

Contains events showing thieves taking money unlawfully by force or threat of force. Often involves using a weapon to impose fear.

Stealing

Contains events showing people taking property or money without permission, usually on the public street. Typical examples would be stealing a vehicle.

Shooting

Contains events showing act of shooting someone with a gun.

Shoplifting

Contains events showing people stealing items from a shop while posing as a shopper.

Assault

Contains events showing a sudden or violent physical attack on someone. Unlike Fighting, the person who is assaulted does not fight back and the attack is sudden and unexpected.

Fighting

Contains events displaying two or more people involved in physical altercations, where they attack each-other.

Arson

Contains events showing people deliberately setting fire to property, such as vehicles or homes.

Explosion

Contains events showing destructive event of something unintentionally exploding, such as a vehicle or a factory.

Arrest

Contains events showing police arresting individuals. These are legitimate arrests, in contrast with police brutality, considered in *Abuse*.

Road Accident

Contains events showing traffic accidents involving vehicles, pedestrians or cyclists.

Vandalism

Contains events involving deliberate destruction of or damage to public or private property, including property damage, such as graffiti and defacement directed towards any property without permission of the owner.

Normal

Contains events where no crime occurred. These videos include both indoor (such as a shopping mall) and outdoor scenes as well as day and night-time scenes.

FineVAU-Judge

You are an impartial AI evaluator for a video anomaly understanding task. Your role is to compare a model's textual description of a video ('MODEL_ANSWER') with a structured JSON representation of the ground truth ('GROUND_TRUTH_JSON').

Your task is to meticulously evaluate the 'MODEL_ANSWER' against the 'GROUND_TRUTH_JSON' and produce a single JSON object as your output. Do not include any explanatory text before or after the JSON object.

Follow these evaluation criteria precisely:

- Location Evaluation:** For each key in the 'Location' object of the ground truth (e.g., "environment", "time_of_day"), determine if the 'MODEL_ANSWER' correctly identifies its value.
 - Score '1' if the information is present and correct.
 - Score '0' if the information is missing or incorrect.
- Entity Evaluation:** For each entity listed in the 'Entity' array of the ground truth:
 - 'presence_score': Score '1' if the entity (identified by its "id") is mentioned in the 'MODEL_ANSWER'. Score '0' if it is not mentioned.
 - 'attributes': For each attribute of the entity (e.g., "color", "clothing"), score '1' if the 'MODEL_ANSWER' correctly states the attribute. Score '0' if it's missing or incorrect.
- Event Evaluation:** For each event in the 'Event' array of the ground truth:
 - 'score': Compare the ground truth 'event_description' with the 'MODEL_ANSWER'.
 - Score '2' (Perfect Match): The 'MODEL_ANSWER' accurately and completely describes the event.
 - Score '1' (Partial Match): The 'MODEL_ANSWER' describes some aspects of the event correctly but misses key details or contains minor inaccuracies.
 - Score '0' (Mismatch/Miss): The 'MODEL_ANSWER' fails to describe the event or provides a completely incorrect description.

Your output **MUST** be a single, valid JSON object structured exactly as follows:

```
{
  "location": {
    "environment": <0 or 1>,
    "time_of_day": <0 or 1>,
    "lighting_conditions": <0 or 1>,
    "crowd_density": <0 or 1>,
    "salient_feature": <0 or 1>
  },
  "entity": [
    {
      "id": "<ground_truth_entity_id>",
      "presence_score": <0 or 1>,
      "attributes": {
        "<attribute_name>": <0 or 1>,
        ...
      }
    },
    ...
  ],
  "event": [
    {
      "event_description": "<ground_truth_event_description>",
      "score": <0, 1, or 2>
    },
    ...
  ]
}
```

<GROUND_TRUTH_JSON>

<MODEL_ANSWER>

Figure 7: **FineVAU-Judge prompt.** Prompt used to elicit a structured output with the membership scores for each ground truth element from FineVAU-Judge. We use Gemini-2.5-flash (Team 2025) as the judge.

Event Decomposition and Entity Linking

Role: Senior Crime Scene Analyst

Objective: Your task is to re-create a log of events for a video, based on visual evidence. You will be given a raw, often incomplete, log as a starting point. Your goal is to produce a new, high-quality log that is detailed, accurate, and complete.

Input: Video file and a JSON object with a raw log of `timestamps` and `sentences`.

Output: A single JSON object. Your entire response must be ONLY the JSON object.

Instructions & Mandates:

- 1. Re-evaluate Every Event:** For each `timestamp` and `sentence` provided, you must re-watch the corresponding video segment and write a **new, more detailed** `event\_description`. Enhance it with specific details about actions and entities. **It is extremely important that you do not remove any details about events or entities present in the log that are anomalous or dangerous.** Assume that existing log is true, but may be incomplete.
 - * **Poor:** "A person is on the ground."
 - * **Good:** "A person in a black hoodie lies motionless on the pavement."
- 2. Add Missing Critical Events:** Scrutinize the entire video, especially the gaps between timestamps. If you see a critical event that was completely missed (e.g., a weapon being drawn, a vehicle arriving, an explosion), you must add it as a new event object with its correct timestamp.
- 3. Identify All Involved Entities:** For every event, new or re-written, you must populate the `involved\_entities` list.
 - * An entity is any person, vehicle, animal, or key object involved in the event.
 - * **Describe entities with detail:** Instead of "a man," use "a tall man in a green jacket." Instead of "a car," use "a blue sedan with a broken window." This detail is mandatory.
- 4. Define Event Type:** For every event, classify its `event\_type` as one of: `Abnormal`, `Normal`.
- 5. Do not repeat events or entities.** Sometimes videos are replay loops. If so, do not repeat the same event several times.

Output Schema:

```
{
  "refined_events": [
    {
      "timestamp": [10.2, 15.5],
      "event_type": "Interaction",
      "event_description": "A man wearing a red t-shirt and a black baseball cap violently shoves another man against a brick wall.",
      "involved_entities": ["man in red t-shirt and black cap", "man shoved against wall", "brick wall"]
    },
    {
      "timestamp": [18.1, 22.5],
      "event_type": "Action",
      "event_description": "The man in the red t-shirt runs away down the street.",
      "involved_entities": ["man in red t-shirt and black cap"]
    }
  ]
}
```

<VIDEO>

<GROUND_TRUTH_JSON>

Figure 8: **Stage 1 Prompt for Dataset Annotation.** Prompt used to augment and refine existing event annotations. We use Gemini-2.5-Flash (Team 2025) as the LVLM annotator.

Entity Grounding and Scene Description

Role: Expert Visual Analyst

Objective: Analyze a video, an entity list, and event descriptions to create a detailed JSON catalogue. Your goal is to refine the provided entity data by enriching and specifying details based on visual evidence.

Output Format: Your entire response must be a single, valid JSON object and nothing else.

Instructions:

1. **Refine `entity\_description`:**

- * **Simplify:** Shorten long descriptions to their core identity (e.g., "woman with white top" becomes "a woman"). Move attributes to their dedicated fields.

- * **Specify and Identify:** Replace vague descriptions with precise ones (e.g., "an object on the table" becomes "a red book"). For any entity previously described as "unidentified" or "something" or similar, visually identify the object and update its id to reflect its concrete category.

- * **New ID:** Put the description (including the ones you changed) into the id.

2. **Enrich Attributes:** Use the video and event context to meticulously fill in all physical attribute fields (`clothing`, `color`, etc.) with specific details. Increment on existing knowledge whenever possible.

3. **Use `salient\_feature` Correctly:**

- * **For people:** Note a single, unique, distinguishing feature (e.g., "has a prominent tattoo").

- * **For the location:** Note the single most notable element of the overall environment.

- * **For all other entities:** Do not include.

4. **Mitigate Hallucinations:** Use "N/A" or `[]` for any attribute that you cannot confirm visually.

5. **Respect the schema:** Do not add additional entries to the JSON. If the field is a list, output a list. If it is a string, output a string.

Reference Output Schema:

```
{
  "location": {
    "environment": "outdoor parking lot",
    "time_of_day": "Daytime",
    "lighting_conditions": "Bright/Well-lit",
    "crowd_density": "Few people",
    "salient_feature": "Overturned trash can"
  },
  "entities": [
    {
      "id": "a man",
      "category": "person",
      "clothing": ["red shirt", "blue jeans"],
      "age_group": "adult",
      "gender": "male",
      "salient_feature": "has a beard"
    },
    {
      "id": "a watch",
      "category": "object",
      "color": "silver"
    }
  ]
}
```

Figure 9: **Stage 2 Prompt for Dataset Annotation.** Prompt used to augment entity and location attributes. We use Gemini-2.5-Flash (Team 2025) as the LVLM annotator.



What?

A man in a black suit and a woman in a long patterned dress enter the waiting area from the lower right. **The man in the black suit immediately walks directly towards the man in the white short-sleeved shirt and the woman in the black dress, appearing to confront them.**

The man in the black suit initiates a physical altercation with the man in the white short-sleeved shirt, pushing him. The woman in the black dress becomes involved in the struggle. The woman in the patterned dress attempts to pull the man in the black suit away

The man in the black suit, the man in the white short-sleeved shirt, and the woman in the black dress to fall to the ground.

A man wearing a brown cowboy hat rushes towards the struggling individuals on the ground, attempting to separate them.

Two security guards rush into the scene and actively intervene to separate the fighting individuals. The man in the black suit is pulled away, and the woman in the black dress is helped to her feet. The man in the white short-sleeved shirt is also helped up from the floor.

Who?



Woman in patterned dress

category : person,
clothing : [long black and white patterned dress, black high heels]
age_group : adult
gender : female
salient_feature : Long dark hair



Man in white short-sleeved shirt

category : person,
clothing : [white short-sleeved shirt, dark pants]
age_group : adult
gender : male
salient_feature : N/A



Woman in black dress

category : person,
clothing : [black short dress, black high heels]
age_group : adult
gender : female
salient_feature : N/A



Two security guards

category : person,
clothing : [white uniform shirt, dark pants]
age_group : adult
gender : male
salient_feature : N/A



Man in brown cowboy hat

category : person,
clothing : [brown cowboy hat, white t-shirt, light-colored shorts]
age_group : adult
gender : male
salient_feature : N/A



Man in black suit

category : person,
clothing : [black suit jacket, white dress shirt, black pants]
age_group : adult
gender : male
salient_feature : N/A

Where?



environment : Indoor waiting area

time_of_day : Daytime

lighting_conditions : Well-lit

crowd_density : Moderate

salient_feature : Long rows of waiting area chairs

Figure 10: **Annotation example from FineW³**. Our dataset contains highly detailed and fine-grained information regarding events (*What*), their involved entities and their physical attributes (*Who*), and the location where events happen, as well as physical attributes of this location (*Where*). To mitigate hallucinations, common in synthetic VAU data, we ground events in preexisting data from an high quality human labeled source (UCA (Yuan et al. 2024)), and augment event, entity and location data by instructing the LVLM annotator to only annotate information that is clearly visible in the video. For information that may not clearly be visible, we instruct the LVLM to annotate "N/A" for "Not Applicable". In the example illustrated in the image, the model can't observe distinguishing features for some of the entities annotated. This may indicate: 1) the distinguishing features are already present elsewhere in the data (e.g., in "man in brown cowboy hat", the brown cowboy hat is already in the clothing) or they are simply not clearly visible (e.g., "woman in black dress").

LVLM

Please provide a comprehensive description of the video.

In your description, detail:

- the setting, including the location, the time_of_day, the lighting conditions, the crowd density and the main salient feature of the location.
- any objects (e.g. people, vehicles, animals, or other objects in general) and physical attributes:
 - for people: their clothing, approximate age, gender and any prominent salient features
 - for other objects: category, colors
- the sequence of events in the video, with a particular emphasis on abnormal, criminal, dangerous or illegal events.

Ensure your description is as comprehensive and precise as possible.

Figure 11: **Prompt for LVLM evaluation in FineVAU.** We use a unified prompt across all five open-source LVLMs evaluated. This prompt is designed to extract maximum performance in our benchmark, deliberately encouraging the models to provide a report that addresses all aspects of FineVAU evaluation.



Model Report (VideoLLaMA3)

**Static Coarse
Grained
Understanding**

**Fine-Grained
Spatial and
Temporal
Understanding**

Normalcy Bias

The video shows a crowded waiting room with many people sitting on chairs. The room is well-lit, and there are large windows in the background. A man in a black shirt is pushing a woman in a wheelchair towards the center of the room. As they move, a group of people, including a man in a white shirt and a woman in a black dress, walk past them. The man in the black shirt continues to push the woman in the wheelchair towards the center of the room, where a group of people is gathered. The people in the group are standing and talking, and some of them are looking at the man and the woman in the wheelchair. The man in the black shirt stops pushing the woman in the wheelchair and stands next to her. The group of people in the center of the room continues to talk and look at the man and the woman in the wheelchair.

Figure 12: **Qualitative example of a weak report.** In this example, VideoLLaMA3 (Zhang et al. 2025a) provides an answer that exhibits the patterns identified in the results discussion. While the model can understand the general context of the video (i.e., a "crowded waiting room with many people sitting on chairs"), it struggles to understand specific details, conflating the man in the wheelchair, with one of the women, and hallucinating normal actions such as "talking" and "pushing the woman in the wheelchair". Critically, it completely fails to report any fight or physical altercation between the man in suit and the man in white shirt, as well as the involvement of security officers and the man in a brown cowboy hat.



Model Report (Qwen2.5-VL)

Static Coarse Grained Understanding

The video begins with a wide-angle view of a road that appears to be in a semi-urban area. The setting suggests a daytime scene under clear skies, with ample natural light illuminating the environment. The road stretches into the distance, flanked by utility poles and some greenery on either side. In the background, there are structures that resemble small buildings or houses, indicating a residential or mixed-use area.

The road itself has a few vehicles moving along it, including cars and what seems to be a truck. The traffic is light, suggesting a calm day without heavy congestion. The overall atmosphere is peaceful, with no immediate signs of distress or unusual activity.

Events with strong visual cues

As the video progresses, there is a sudden and dramatic shift. A massive explosion occurs, sending a large plume of dark smoke and debris into the sky. This event is abrupt and unexpected, marking a significant change from the initial calmness. The smoke billows outwards, obscuring much of the view, and the sound of the explosion likely accompanies this visual change, though it is not explicitly described here.

Following the explosion, the video shows a rapid transition to a completely different scene, which appears to be a glitch or a technical error, resulting in a distorted and fragmented image. This suggests a disruption in the video feed, possibly due to an issue with the camera or transmission system.

Figure 13: **Qualitative example of a performant report.** Despite the low accuracy of Qwen2.5-VL (Bai et al. 2025), the model can successfully identify the overall scene of the video and accurately describe the explosion, the dark smoke, the debris and even the camera glitch. This example further validates our hypothesis that LVLMs can accurately report more static and coarse information, or events with strong visual cues.



Model Report (Intern-VL3)

Static Coarse Grained Understanding

The video takes place in a small, well-lit electronics shop with a narrow aisle. The shop is filled with various electronic items and boxes stacked on shelves and tables. The lighting is bright, and the floor is tiled. A black office chair is positioned in the center of the aisle.

A woman wearing a pink blouse and a dark saree with floral patterns is present throughout the video. She has long hair and is carrying a brown shoulder bag. Initially, she stands in the aisle, looking around. She then starts to interact with the items on the shelves, picking up and examining them. She moves back and forth along the aisle, occasionally bending down to look at items on lower shelves.

Fine-Grained Spatial and Temporal Understanding

As the video progresses, the woman continues to handle the items, **occasionally placing them back on the shelves**. She appears to be searching for something specific. **Eventually, she picks up a large, dark-colored bag from the floor and starts to pack items into it**. She continues to move items from the shelves into the bag, **methodically organizing them**.

Normalcy Bias

The video concludes with the woman walking out of the frame, leaving the shop empty except for the black office chair and the items on the shelves. There are no other people visible in the video, and no abnormal, criminal, dangerous, or illegal activities are observed.

Figure 14: **Qualitative example of a mixed report.** Despite being the strongest baseline, InternVL3 still struggles with very sudden and discrete events. In this case, while the model can accurately report the context of the scene, where a woman is looking at items in an electronics shop; and the physical attributes of the woman, such as the long hair. However, it conflates the skirt of the woman with a dark-colored bag and the act of hiding the items in her purse with methodically organizing items into the bag, revealing difficulty in understanding details and a strong hallucination bias towards normal situations.