VAGU & GtS: LLM-Based Benchmark and Framework for Joint Video Anomaly Grounding and Understanding

Shibo Gao[1,2], Peipei Yang[2,3](✉)*, Yangyang Liu[2,3], Yi Chen[3], Han Zhu[2,3], Xu-Yao Zhang[2,3], Linlin Huang[1]

[1]Beijing Jiaotong University
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences

Abstract

For video anomaly detection, it's both important to detect when the event happens and what the event is. The tasks of temporal grounding and semantic understanding can benefit from joint learning, but no existing work support it. To address this problem, we introduce VAGU (Video Anomaly Grounding and Understanding), the first benchmark designed to jointly evaluate semantic understanding and precise temporal grounding of anomalies, with comprehensive annotations and objective multiple-choice Video QA. Besides, we propose Glance then Scrutinize (GtS), the first training-free framework that achieves the best balance performance in both accuracy and efficiency. GtS uniquely balances high temporal precision and semantic interpretability while meeting practical speed requirements, outperforming previous methods in real-world scenarios. Furthermore, we introduce the JeAUG metric for holistic evaluation of both speed and accuracy. Extensive experiments demonstrate the superior effectiveness and practicality of our benchmark, framework, and metric.

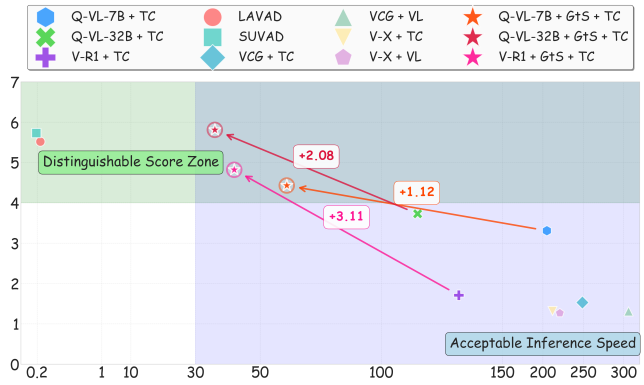

Figure 1: Comparison of our proposed framework with other LLM-based VAD methods in terms of inference speed and performance. It can be seen that our proposed framework achieves the best balance. The specific methods corresponding to the abbreviations can be found in Table. 2.

Introduction

Video Anomaly Detection (VAD) aims to understand and temporally ground anomalous events in video sequences. In recent years, driven by the growing demand for real-time monitoring in industrial automation, intelligent surveillance, and smart transportation systems, VAD has emerged as a critical research frontier in computer vision and multimedia analytics (Luo, Liu, and Gao 2017; Lu et al. 2020; Park et al. 2022; Xu et al. 2022; Liu et al. 2025a,b).

The current video anomaly detection domain exhibits significant methodological bifurcation: traditional DNN-based methods (semi-supervised (Wang et al. 2022b; Liu et al. 2021; Wang et al. 2022a), weakly-supervised (Zhou, Yu, and Yang 2023; Zaheer et al. 2022; Joo et al. 2023; Wu et al. 2023), open-set VAD (Ding, Pang, and Shen 2022; Zhu, Bao, and Yu 2022; Zhu et al. 2023) and else) and LLM-based methods respectively focus on temporal grounding and semantic understanding of anomalies, creating a "when-what" capability dissociation (Tang et al. 2024; Ye, Liu, and He 2024; Bharadwaj et al. 2024). On one hand, traditional DNN-based methods learn normal/anomalous patterns in video sequences through video-level classification labels to capture anomalous events, yet they merely output temporal grounding results ("when anomalies occur") while lacking semantic understanding. On the other hand, emerging LLM-based methods can generate natural language descriptions leveraging LLMs' open-domain knowledge to answer "what anomalies occur", but universally neglect precise temporal grounding of anomaly onset/offset boundaries.

Although some methods attempt joint grounding and understanding through vision-language models (VLMs) via frame/segment-wise video analysis (Gao, Yang, and Huang 2025; Zanella et al. 2024; Ahn et al. 2025), their prohibitively high computational overhead renders them unsuitable for meeting VAD's stringent real-time processing requirements. Notably, due to the lack of a consistent standard for defining anomalous events, existing VQA/VTG models exhibit limited applicability when directly transferred to the VAD domain, as they face challenges with the openness of the problem (Lv and Sun 2024).

In light of these research, we systematically investigate three pivotal questions:

*First author: shibo.gao@nlpr.ia.ac.cn, Corresponding author: ppyang@nlpr.ia.ac.cn

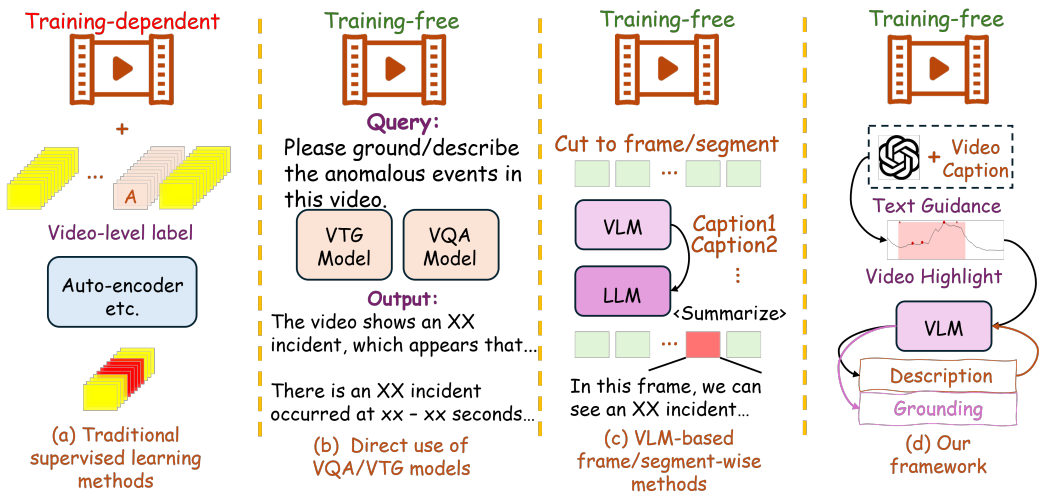

Figure 2: Comparison of our proposed framework with other VAD paradigms. Our proposed GtS framework can construct anomaly grounding and anomaly understanding capabilities based on existing multi-modal large language models under training-free conditions, while maintaining an acceptable inference speed.

Q1:Why both importance of grounding & understanding?

A1: VAD task necessitates simultaneous resolution of both temporal grounding ("when") and semantic comprehension ("what") queries (Gao, Yang, and Huang 2024; Zanella et al. 2024; Gao, Yang, and Huang 2025; Du et al. 2024). Additionally, these dual objectives exhibit intrinsic interdependence: temporal grounding provides critical contextual cues for semantic understanding, while semantic understanding feedback validates the plausibility of grounding results. Nevertheless, both DNN-based methods and LLM-based methods still fall short in establishing effective mechanisms for mutual assistance between these two dimensions. Unified annotation ensures alignment between semantics and temporal segments, avoids inconsistencies from merging datasets, and enables joint training and fair evaluation to improve overall model performance.

Q2: Why do VQA/VTG models perform poorly when directly applied to VAD tasks?

A2: The fundamental reason lies in the absence of a consistent standard for defining the scope of anomalies. In VAD tasks, there are typically no explicit labels indicating specific anomalies within each video, and temporal grounding prompts are inherently missing. This makes it difficult for VQA models, which rely on predefined questions, to effectively handle open-ended anomaly-related queries such as "What anomalies are present in the video?" Similarly, VTG models depend on textual queries to ground relevant video segments, but in VAD, such queries are usually unavailable (Lin et al. 2023; Shu et al. 2024; Maaz et al. 2023; Huang et al. 2024; Ye et al. 2023).

Q3:Why training-free framework?

A3: Firstly, VAD tasks face data scarcity challenges: real-world normal/anomalous data exhibits extremely imbalanced distribution, with anomalous data collection difficulties and prohibitively high annotation costs. Furthermore, the collected datasets often fail to encompass all anomalous categories, which constrains the scalability of supervised learning methods(Gao, Yang, and Huang 2024; Kim, Angelova, and Kuo 2023). Secondly, real-world VAD applications demand exceptional generalizability, whereas supervised learning methods frequently exhibit degraded robustness in cross-scenario deployments. Notably, although frame/segment-wise VLM-based training-free methods can concurrently output grounding and comprehension, its computational overhead remains prohibitively high — processing a video of several tens of seconds may require tens of hours — thus fundamentally conflicting with VAD's real-time processing requirements.

Building upon these critical questions, we rethink the VAD task and propose the VAGU benchmark (Video Anomaly Grounding and Understanding) — the first benchmark that integrates both anomaly grounding and anomaly understanding. VAGU comprises over 7,567 real-world videos spanning 21 major anomaly categories (covering human criminal activities, natural disasters, animal-related injuries, traffic accidents, etc.), with an average duration exceeding 2,700 frames. In addition, we provide over 20,000 anomaly-related QA pairs to facilitate comprehensive anomaly understanding.

Furthermore, we innovatively propose the GtS (Glance then Scrutinize) framework, which achieves "coarse-grained temporal grounding $\rightarrow$ fine-grained anomaly comprehension $\rightarrow$ fine-grained anomaly grounding" through dynamic and static textual guidance under training-free conditions. This framework constructs anomaly grounding and anomaly understanding capabilities by leveraging existing multi-modal large language models (MLLMs).

Fig. 1 illustrates performance and inference speed comparisons between our framework and existing methods, while Fig, 2 delineates distinctions between our framework and other VAD paradigms. The proposed framework

achieves an optimal balance between model performance and computational efficiency.

Additionally, we introduce the JeAUG metric, which jointly quantifies semantic accuracy and grounding precision while incorporating video duration as a weighting factor. This enables more equitable evaluation of VAD capabilities across diverse data scenarios compared to conventional evaluation systems(AUC, AP and others). Extensive experiments on the proposed benchmark demonstrate that VAGU effectively supports complex VAD task assessments. The GtS framework exhibits significant performance advantages on VAGU, and JeAUG provides more comprehensive and fairer evaluation than traditional metrics.

Overall, our contributions are summarized as follows:

- We have developed VAGU, a novel VAD benchmark that jointly addresses anomaly grounding and anomaly understanding. To the best of our knowledge, VAGU is the first large-scale benchmark that combines anomaly grounding and anomaly understanding, and also the first to provide an objective multiple-choice benchmark related to anomalies. Compared to existing datasets, our dataset is more comprehensive, challenging, and features higher annotation quality.
- We propose GtS, a training-free VAD framework that leverages text guidance to build capabilities for anomaly grounding and anomaly understanding on existing multimodal large language models.
- Based on VAGU, we introduce an evaluation metric that jointly quantifies semantic accuracy and grounding precision.
- Extensive experiments have been conducted on the proposed VAGU, demonstrating the superiority of our benchmark, framework, and evaluation metric.

Related Work

Traditional DNN-based VAD Paradigms

Semi-supervised and weakly-supervised methods remain dominant in video anomaly detection (VAD). Semi-supervised approaches use self-supervised tasks to learn normal patterns (Lu et al. 2020; Dong, Zhang, and Nie 2020; Zhao et al. 2017; Astrid, Zaheer, and Lee 2021; Wu, Moore, and Shah 2010), such as auto-encoder reconstruction error and temporal prediction models (Wang and Cherian 2019; Wu, Liu, and Shen 2019). Some recent works further improve robustness via multi-task learning (Wang et al. 2022a; Gao, Yang, and Huang 2024), but still struggle with adapting to new scenarios, as minor viewpoint changes can reduce performance.

Weakly-supervised methods use video-level annotations, often with multiple instance learning or semantic priors for anomaly inference (Zhou, Yu, and Yang 2023; Zaheer et al. 2022; Joo et al. 2023; Wu et al. 2023). While these improve detection accuracy, their reliance on manual annotation limits its scalability in real-world deployments.

Unsupervised and open-set VAD methods (Acsintoae et al. 2022; Ding, Pang, and Shen 2022; Zhu, Bao, and Yu 2022; Zhu et al. 2023) seek to reduce annotation needs, but often perform poorly across varying scenarios due to architectural limitations. Across all supervised paradigms, a key challenge remains: limited anomaly understanding, which restricts deployment in complex real-world settings.

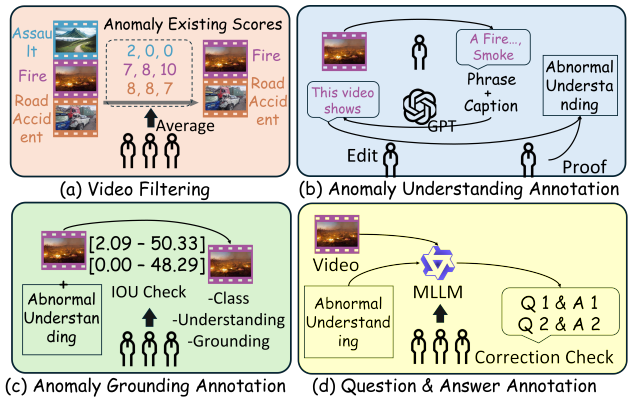

Figure 3: The annotation workflow of the VAGU dataset. Each sample is processed by at least three annotators collaboratively at every stage.

LLM-Based VAD Paradigms

Recently, LLM-based VAD methods have advanced significantly (Feng et al. 2025a,b). Vision-language models (VLMs) combine LLM reasoning with visual feature extraction, showing strong potential for VAD. Current approaches fall into two main categories:

The first uses external LLMs with frozen VLMs. Videos are segmented, described by VLMs, and processed by LLMs for semantic integration and detection (Du et al. 2024; Tang et al. 2024; Ye, Liu, and He 2024; Lv and Sun 2024; Bharadwaj et al. 2024; Zanella et al. 2024; Gao, Yang, and Huang 2025; Ahn et al. 2025; Yang et al. 2024). LAVAD (Zanella et al. 2024) provides a training-free pipeline, while SU-VAD (Gao, Yang, and Huang 2025) and FtR (Yang et al. 2024) introduce rule-mining and hallucination mitigation. Although these methods offer precise grounding and anomaly understanding, their frame/segment-wise processing incurs high computational costs, limiting real-time applications.

The second direction enhances VLMs via instruction tuning for interpretable predictions. Holmes-VAD (Zhang et al. 2024) uses QA datasets and temporal sampling. VAD-LLaMA (Lv and Sun 2024) improves contextual modeling with LTC modules and a three-stage training strategy. CUVA (Du et al. 2024) adds a MIST selector for feature extraction, and VERA (Ye, Liu, and He 2024) generates instructional questions under weak supervision. While these methods boost detection performance, they require large domain-specific datasets for fine-tuning, leading to high computational cost and limited generalization.

Proposed VAGU Benchmark

In this section, we will first introduce the anomaly grounding and anomaly understanding tasks required by the VAGU

Table 1: A detailed comparison between the proposed VAGU dataset and other datasets. A.U. and A.G. mean Anomaly Understanding and Anomaly Grounding. The A.G. annotations in the CUVA and HIVAU-70k dataset are accomplished using VLM, which results in significant errors in application.

Dataset	Domain	Dataset Statistical Information		Dataset Annotation			
		Video Samples	Anomaly Categories	Audio	A.U.	A.G.	QA
UCF-Crimes	Crime	1900	13			Frame/Human	
XD-Violence	Violence	800	6	✓		Frame/Human	
ShanghaiTech	Streetscape	437	13			BBbx/Human	
UCSD Ped1	Streetscape	70	5			BBbx/Human	
UCSD Ped2	Streetscape	28	5			BBbx/Human	
CUHK Avenue	Streetscape	37	5			BBbx/Human	
Street Scene	Traffic	81	17			BBbx/Human	
CUVA	Multiple	1000	11	✓	Caption/Human	Period/VLM	
VANE-Bench	Multiple	325	19				✓
HIVAU-70k	Multiple	5443	15		Caption/Human	Period/VLM	
VAGU(Ours)	Multiple	7567	21	✓	Caption/Human	Period/Human	✓

dataset. Subsequently, we will describe the data collection process of VAGU and its semi-automatic annotation workflow with human verification. Finally, we will present detailed statistical information of the dataset along with comparative analyses against existing datasets.

Task Definition

Video Anomaly Understanding. This task comprises two objectives: anomaly classification and anomaly understanding. In the anomaly classification subtask, the model is expected to output the category of the anomalous event present in the video, selected from a predefined anomaly category database. In the anomaly understanding subtask, the model must analyze the given video content to comprehensively describe the subject, process, causes, and consequences potentially involved in the anomalous event.

Video Anomaly Grounding. This task requires the model to detect precise temporal intervals of anomalous events based solely on video data without external semantic information. This task's core constraints include: (1) *Prompt Ambiguity:* The use of any video-level labels or fine-grained semantic labels provided in the dataset, except for the predefined anomaly type list, is strictly prohibited. (2) *Information Interactivity:* The model permits the incorporation of anomaly describe derived from the upper-level video anomaly understanding task as grounding evidence.

Data Collection

Compared to normal videos, those containing anomalous events are exceptionally scarce. We integrated existing datasets (CUVA (Du et al. 2024), UCF-Crime(Sultani, Chen, and Shah 2018), and XD-Violence (Wu et al. 2020)) and collected over 12,000 videos potentially containing anomalies from major platforms such as YouTube, BiliBili, and TikTok. Through rigorous analysis and filtering, we curated 7,567 high-quality videos spanning 21 distinct anomaly categories across domains including human criminal activities, natural disasters, traffic accidents, and animal-inflicted injuries.

Manual Annotation

The construction of our VAGU dataset consists of three main steps: video filtering, semi-automatic anomaly annotation, and anomaly grounding. The process took about 650 hours and involved 20 annotators, with at least three annotators working on each video per stage. We collected videos from public datasets and online platforms, removed inappropriate content, and used a three-level quality review. For annotation, we adopted a human-AI collaboration: annotators labeled key phrases, generated descriptions with vision-language models, and expanded them with ChatGPT. All annotations underwent multi-person cross-checks for accuracy and consistency. For anomaly grounding, at least three annotators independently marked the time intervals of anomalies, achieving consensus through IoU-based aggregation; ambiguous cases were iteratively re-annotated. Finally, we used advanced multimodal models to create multiple-choice QA tasks for each video, with annotators verifying correctness. Fig. 3 illustrates this process.

Dataset Detailed Information

The VAGU dataset comprises 7,567 high-quality anomaly or normal videos (3520 anomalies and 4047 normal), each annotated with triple labels for anomaly classification, understanding, and grounding. This dataset covers four major domains—human criminal activities, natural disasters, traffic accidents, and animal-related injuries—encompassing 21 fine-grained categories. Table. 1 provides a horizontal comparison with existing datasets through multi-dimensional metrics. Notably, certain categories (e.g., "Fire," "Arson," "Burning") exhibit semantic similarities; detailed definition criteria for differentiation can be found in the appendix. It is worth noting that for videos classified as "Normal," we treat them as video understanding tasks by providing semantic understanding annotations and QA annotations, while the grounding annotation is set as default (absent).

Additionally, due to the limitations of existing VLMs, we encourage the use of multiple VLMs for ensemble-based completion of VAD tasks on the VAGU dataset.

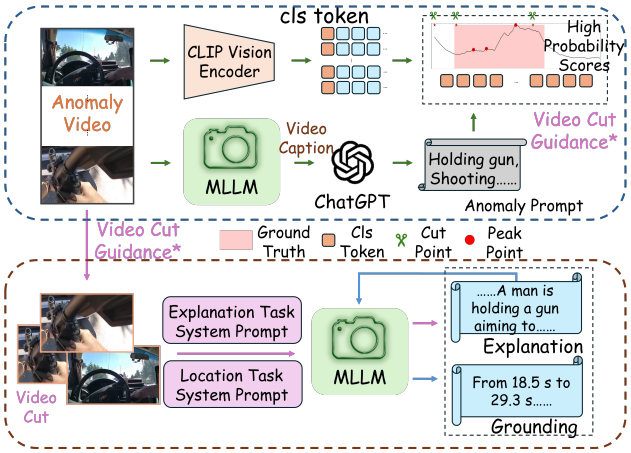

Figure 4: The flowchart of the GtS framework. In the first stage, GtS helps the model ground the time when the main event of the video occurs through dynamic and static text guidance. Then, in the second stage, GtS conducts fine-grained anomaly grounding and understanding based on the video segmentation in the previous stage.

Proposed Framework: Glance then Scrutinize

In this section, we will introduce Glance then Scrutinize (GtS), an novel training-free VAD framework designed to address the three core challenges outlined previously. Built upon existing VLMs, our framework employs a dual textual guidance mechanism encompassing both dynamic and static contexts to ground potential anomalous video segments, thereby establishing comprehensive video anomaly grounding and understanding capabilities. Fig. 4 illustrates the workflow of the proposed framework.

Glance: Static & Dynamic Text Guidance and Video Highlight Cut

Our proposed framework accomplishes video anomaly grounding and understanding through a dual-phase pipeline. During the "Glance" phase, we initially leverage existing VLMs to generate video captions Cap_V for input videos V_{input}. While current VLMs constrained by computational resources often misinterpret genuine anomalies as normal behaviors due to insufficient anomaly-specific token extraction, our experimental findings reveal that the generated descriptions maintain high accuracy in subject recognition. Building upon this insight, we construct multi-source textual inputs incorporating video captions, dataset-provided anomaly lists $\mathcal{A}$, and LLM-pregenerated contextual phrase banks $\mathcal{B}_p$ to drive GPT in parsing latent dynamic information (actions/events) and static information (subjects/scenes) from videos, thereby generating prompt lists for potential anomaly detection:

$$\mathcal{PL}_s, \mathcal{PL}_d = \text{LLM}(\, Cap_V \,,\, \mathcal{A} \,,\, \mathcal{B}_p \,). \tag{1}$$

Crucially, it is noteworthy that the prompt lists returned by GPT may not inherently contain anomalous content, as this phase fundamentally aims to identify principal video subjects while filtering irrelevant background segments through semantic guidance.

We subsequently employ dual encoders (CLIP Φ_{image} and Video-CLIP Φ_{video}) to perform feature encoding on video frames/segments, generating a temporal anomaly probability curve through cross-modal similarity computation:

$$S_{s/d}(t) = \frac{\exp\left(\frac{1}{N}\sum_{n=1}^{N}\langle \Phi_x(pl_x^n),\, \Phi_x(V_t)\rangle\right)}{\sum_{t'}\exp\left(\frac{1}{N}\sum_{n=1}^{N}\langle \Phi_x(pl_x^n),\, \Phi_x(V_{t'})\rangle\right)} \tag{2}$$

where $x \in image, video$, and N denotes the number of text descriptions in the corresponding branch; pl_x^n represents the static or dynamic text descriptions; V_t refers to the input frame or segment; Φ_x is the corresponding encoder; and $\langle\cdot,\cdot\rangle$ indicates the cosine similarity. N is the number of static or dynamic text descriptions

Considering that events are continuous, we derive the overall similarity curve by combining the two similarity curves and apply the Savitzky-Golay filter to smooth the scores:

$$S(t) = \frac{1}{Q}\sum_{p=-h}^{+h}(\alpha\cdot S_s(t) + (1-\alpha)\cdot S_d(t))\cdot q_p, \tag{3}$$

where q_p/Q is the smoothing coefficient, determined through polynomial fitting using the least squares method, α is constant term.

Based on this curve, we implement a three-stage segmentation strategy: firstly detecting local extrema points within the similarity curve, secondly screening top-K candidate peaks according to inter-peak distances and magnitude thresholds:

$$\mathcal{P} = \{t \in [0,T) \;\mid\; S(t) = argmaxima(S)\}, \tag{4}$$

$$\mathcal{P}^* = \text{TopK}\Big(\{p \in \mathcal{P} \;\mid\; |p_i - p_j| > \theta, S(p_i) >= \tau\}\Big), \tag{5}$$

where θ and τ are thresholds.

Finally, we perform dynamic window partitioning around the selected peaks while considering the total video duration, thereby segmenting the original video into high/low anomaly probability segments:

$$\mathcal{H} = \bigcup_{p\in\mathcal{P}^*}[\max(0, p-\beta T), \min(T, p+\beta T)], \tag{6}$$

where $\beta \in (0,1)$ controlling the window size proportion relative to T.

Scrutinize: Fine-grained Anomaly Grounding and Understanding

In the "scrutinize" phase, we establish a closed-loop integration of anomaly understanding and grounding through existing VQA and VTG models. For high-probability anomalous segments, we deploy the VQA model to detect and describe anomalous events based on predefined anomaly catalogs from dataset.

To more precisely capture anomalous cues, we perform non-uniform sampling on each segmented video clip based on the similarity curve. Specifically, we select frames according to the cumulative distribution of similarity scores, ensuring that sampling density is proportional to local similarity values. By partitioning the cumulative similarity scores into N equal intervals, we determine the corresponding timestamps as sampling points.

$$P_i = \min \left\{ m \in \{a, ..., b\} \;\middle|\; \sum_{t=a}^{m} S(t) \geq \frac{i}{N} \sum_{t=a}^{b} S(t) \right\} \tag{7}$$

where $S(t)$ is the similarity score, $\{a, ..., b\}$ denotes the segment interval, P_i are the timestamps of the i-th sampled frame, and N is the total number of sampled frames.

Furthermore, when performing anomaly detection on each non-initial video, we provide the model with the understanding results from the previous segment. This helps to maintain subject consistency in the subsequent integration process.

Compared to holistic video analysis, segmented processing significantly enhances VLMs' ability to capture anomaly-relevant features, effectively identifying subtle anomalies overlooked in full-length videos. For low-probability segments, we use VQA models to generate captions while extracting anomaly-associated clues.

We then employ LLMs to integrate these captions, removing redundant and irrelevant descriptions while establishing semantic connections between segments. This enables detection of causally dependent anomalous behaviors (e.g., theft requiring concealment-escape sequences, arson involving material placement-ignition procedures) through multi-segment evidence fusion.

After anomaly characterization, the GtS framework employs VTG models for temporal grounding. By using fine-grained semantic understanding as contextual prompts, our framework achieves superior grounding precision. This pipeline establishes synergy between anomaly understanding and grounding tasks, fundamentally enhancing overall performance through cognitive-visual alignment.

Experiment

The Proposed JeAUG Metric

The existing VAD evaluation metrics generally have the limitation of single-dimensional assessment. Some studies regard VAD as a VQA task and adopt text similarity measures such as ROUGE (Chin-Yew 2004), BLEU (Papineni et al. 2002), and METEOR(Banerjee and Lavie 2005), or generation quality evaluation metrics based on GPT series models, focusing on the assessment of anomaly semantic understanding. Another category of methods focuses on anomaly spatiotemporal grounding and mainly relies on traditional metrics in the field of computer vision such as AUC and AP. We propose an evaluation metric that jointly assesses anomaly understanding and detection (Joint Evaluation of Anomaly Understanding and Detection, JeAUG).

This metric includes a dual-module evaluation framework: in the dimension of anomaly understanding, it guides external large language models (LLMs) to conduct multi-dimensional scoring on the semantic integrity and logical consistency of anomaly descriptions by constructing structured natural language prompt templates. Specifically, we set up prompts to ask the LLM to score the results from four aspects: subject, scene, course of events, and impact. A score of 1/10 represents the lowest score, indicating that the response is almost entirely unrelated to the ground truth, while a score of 10/10 represents the highest score, indicating that the response is very appropriate in every aspect compared to the ground truth. For the problem of subjective video boundary annotations in the dimension of anomaly grounding, we design an evaluation function that integrates video length weights based on human cognitive levels of abnormal events:

$$F(IoU) = \frac{0.63}{ln10} ln(0.7 \cdot \min(\lfloor 10 \cdot IoU \rfloor, 7) + 1) + 0.5. \tag{8}$$

Finally, we obtained the overall calculation equation for JeAUG:

$$JeAUG = \min(\gamma \cdot \mathrm{F}(IoU), 1) \cdot Score_{A.U.}, \tag{9}$$

where γ is the video length factor. ***For further rational discussion on JeAUG, please refer to the appendix.***

Implementation Details

Our GtS framework employs different VLMs as the anomaly understanding model and anomaly grounding model. Meanwhile, we use CLIP-L/14 to encode video frames. All experiments were conducted using 14 A6000 GPUs, which took approximately 210 hours in total. For the LLM responsible for integration, we use Llama-3.1-8B. In Equation 4, $\alpha = 0.4$, and q_p/Q is obtained through least squares polynomial fitting, in Equation 6, γ is the mean of all peak points, and θ=total frame num / 12, in Equation 7, β=total frame num / 20. The above hyperparameters were derived by randomly sampling 100 videos from the anomalous videos that we discarded (not included in the evaluation videos) and calculating the proportion of anomalous timestamps. For the number of sampled frames N, we uniformly set it to 32. We run all experiments three times and report the median result.

Quantitative Evaluation of GtS

We conducted a systematic comparative experiment on the proposed GtS framework on the VAGU dataset, and the results are shown in the Table. 2. The experimental results indicate that GtS achieves a good balance between computational efficiency and detection accuracy. We note that due to the current limitations of VLMs, directly using VTG/VQA models to handle VAD tasks performs poorly.

To further verify the effectiveness of the framework, we conducted a fine-grained category analysis on the anomaly understanding sub-task. As shown in Fig. 5, compared with the baseline, GtS shows a significant improvement in anomaly categories that require more anomaly clues (such as Arrest, Arson, Riots, Shoplifting, etc.), which fully demonstrates the superiority of GtS.

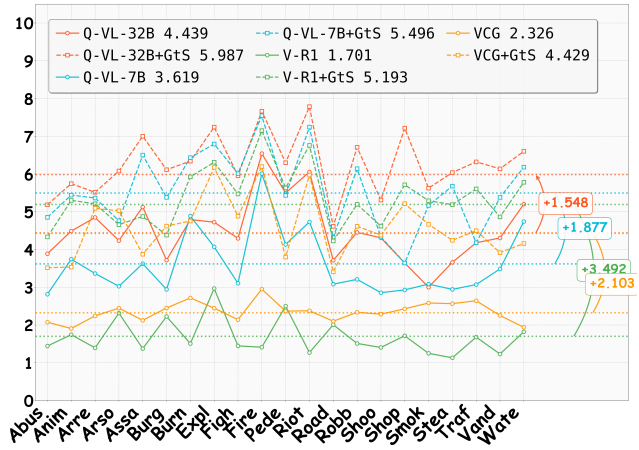

Figure 5: Fine-grained category comparison experiments on the video anomaly understanding sub-task.

Ablation & Case Study

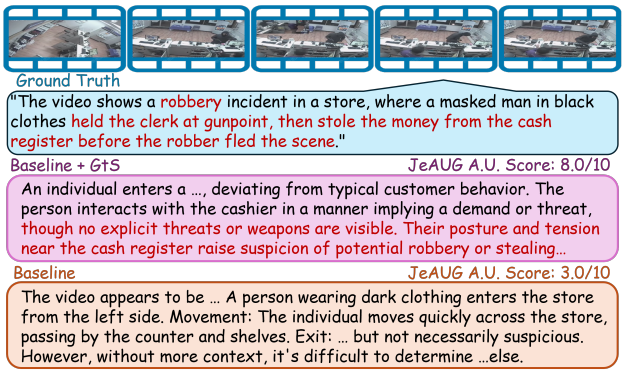

Figure 6: The case study of the GtS on the VAGU.

In Fig. 6, we present a case study of the GtS framework on the VAGU dataset. It can be clearly observed that GtS accurately identified the primary anomaly event and provided a detailed analysis of the entire process. Although the output still shows some hallucination effects and contains redundant descriptions, the training-free VAD framework of GtS undoubtedly exhibits tremendous potential.

In addition, we also investigated the impact of modules such as dynamic texts and integral non-uniform sampling on the anomaly detection performance of the GtS framework, and the results are shown in Table 3. It can be seen that each module enhances the anomaly detection capability of the GtS framework.

Conclusion

In this paper, we introduce VAGU, the first benchmark in the VAD field that simultaneously considers video anomaly understanding and grounding. Compared with existing datasets, VAGU is more comprehensive, more challenging, and has higher annotation quality, with all annotations undergoing multiple rounds of manual checks. Additionally, we propose GtS, which uses dynamic and static text guidance. GtS can follow the general setting of VAD, first grounding the approximate time when the subject event occurs in the video using only the abnormal list provided by the dataset, and then conducting fine-grained anomaly understanding and grounding, achieving a great balance between computational cost and detection performance. We also propose a joint metric, JeAUG, for evaluating the performance of anomaly understanding and grounding, which can more comprehensively and fairly assess model performance. Experimental results show that the introduction of VAGU brings new research directions to VAD.

Table 2: The comparative experiments on the VAGU dataset. The GtS framework we proposed achieved the best balance between inference speed and detection accuracy. In the table, TC stands for TimeChat, and VT stands for VTimeLLM.

Methods	A.U	JeAUG	QA	FPS
Frame/Segment-wise				
LAVAD	5.52	4.47	/	**0.24**
SUVAD	**5.73**	**4.58**	/	0.19
Direct use of VQA/VTG models				
VideoChatGPT + TC	2.32	1.47	60.3%	229
VideoChatGPT + VT	2.10	1.34	same↑	**286**
Video-XL + TC	2.31	1.55	58.6%	192
Video-XL + VT	2.13	1.38	same↑	201
Qwen2.5-VL-7B + TC	3.61	2.28	68.0%	185
Qwen2.5-VL-32B + TC	**4.43**	**2.78**	71.9%	95
Video-R1 + TC	1.70	1.08	**80.0%**	112
Ours GtS				
Qwen2.5-VL-7B + TC*	5.50	4.04	73.5%	61
Qwen2.5-VL-32B + TC*	**5.99**	**4.30**	76.8%	36
Video-R1 + TC*	5.19	3.69	**88.9%**	42
VideoChatGPT + TC*	4.42	3.26	65.1%	**71**

Table 3: The experiment on the impact of dynamic and static text.

Model	JeAUG A.U. Scores
Qwen2.5-VL-7B + GtS	**5.50**
- dynamic text guidance	5.27
- static text guidance	5.30
- integral non-uniform sampling	5.38
- using contextual understanding	5.41

Acknowledgments

This work has been supported by "Scientific and Technological Innovation 2030" Program of China Ministry of Science and Technology (2021ZD0113803) and the National Natural Science Foundation of China under Grants 62222609.

References

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

Ahn, S.; Jo, Y.; Lee, K.; Kwon, S.; Hong, I.; and Park, S. 2025. AnyAnomaly: Zero-Shot Customizable Video Anomaly Detection with LVLM. *arXiv preprint arXiv:2503.04504*.

Astrid, M.; Zaheer, M. Z.; and Lee, S.-I. 2021. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 207–214.

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*, 65–72.

Bharadwaj, R.; Gani, H.; Naseer, M.; Khan, F. S.; and Khan, S. 2024. VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs. *arXiv preprint arXiv:2406.10326*.

Chin-Yew, L. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

Ding, C.; Pang, G.; and Shen, C. 2022. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7388–7398.

Dong, F.; Zhang, Y.; and Nie, X. 2020. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8: 88170–88176.

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

Feng, T.; Wang, X.; Jiang, Y.-G.; and Zhu, W. 2025a. Embodied AI: From LLMs to World Models. In *IEEE Circuits and Systems Magazine*.

Feng, T.; Wang, X.; Zhou, Z.; Wang, R.; Zhan, Y.; Li, G.; Li, Q.; and Zhu, W. 2025b. EvoAgent: Agent Autonomous Evolution with Continual World Model for Long-Horizon Tasks. *arXiv preprint arXiv:2502.05907*.

Gao, S.; Yang, P.; and Huang, L. 2024. Scene-Adaptive SVAD Based On Multi-modal Action-Based Feature Extraction. In *Proceedings of the Asian Conference on Computer Vision*, 2471–2488.

Gao, S.; Yang, P.; and Huang, L. 2025. SUVAD: Semantic Understanding Based Video Anomaly Detection Using MLLM. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14271–14280.

Joo, H. K.; Vo, K.; Yamazaki, K.; and Le, N. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, 3230–3234.

Kim, D.; Angelova, A.; and Kuo, W. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11144–11154.

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

Liu, Y.; Liu, J.; Li, C.; Xi, R.; Li, W.; Cao, L.; Wang, J.; Yang, L. T.; Yuan, J.; and Zhou, W. 2025a. Anomaly Detection and Generation with Diffusion Models: A Survey. *arXiv preprint arXiv:2506.09368*.

Liu, Y.; Liu, S.; Zhu, X.; Li, J.; Yang, H.; Teng, L.; Guo, J.; Wang, Y.; Yang, D.; and Liu, J. 2025b. Privacy-preserving video anomaly detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–22.

Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; and Li, G. 2021. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13588–13597.

Lu, Y.; Yu, F.; Reddy, M. K. K.; and Wang, Y. 2020. Few-shot scene-adaptive anomaly detection. In *European Conference on Computer Vision*, 125–141.

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

Lv, H.; and Sun, Q. 2024. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*.

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

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*, 311–318.

Park, C.; Cho, M.; Lee, M.; and Lee, S. 2022. FastAno: Fast anomaly detection via spatio-temporal patch transformation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2249–2259.

Shu, Y.; Liu, Z.; Zhang, P.; Qin, M.; Zhou, J.; Liang, Z.; Huang, T.; and Zhao, B. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*.

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

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

Wang, G.; Wang, Y.; Qin, J.; Zhang, D.; Bao, X.; and Huang, D. 2022a. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision*, 494–511.

Wang, J.; and Cherian, A. 2019. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8201–8211.

Wang, Y.; Qin, C.; Bai, Y.; Xu, Y.; Ma, X.; and Fu, Y. 2022b. Making Reconstruction-based Method Great Again for Video Anomaly Detection. In *2022 IEEE International Conference on Data Mining (ICDM)*, 1215–1220.

Wu, P.; Liu, J.; and Shen, F. 2019. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems*, 31(7): 2609–2622.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 322–339. Springer.

Wu, P.; Zhou, X.; Pang, G.; Zhou, L.; Yan, Q.; Wang, P.; and Zhang, Y. 2023. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *arXiv preprint arXiv:2308.11681*.

Wu, S.; Moore, B. E.; and Shah, M. 2010. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *IEEE computer society conference on computer vision and pattern recognition*, 2054–2060. IEEE.

Xu, Z.; Zeng, X.; Ji, G.; and Sheng, B. 2022. Improved anomaly detection in surveillance videos with multiple probabilistic models inference. *Intelligent Automation & Soft Computing*, 31: 1703–1717.

Yang, Y.; Lee, K.; Dariush, B.; Cao, Y.; and Lo, S.-Y. 2024. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, 304–322. Springer.

Ye, M.; Liu, W.; and He, P. 2024. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. *arXiv preprint arXiv:2412.01095*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zaheer, M. Z.; Mahmood, A.; Khan, M. H.; Segu, M.; Yu, F.; and Lee, S.-I. 2022. Generative cooperative learning for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14744–14754.

Zanella, L.; Menapace, W.; Mancini, M.; Wang, Y.; and Ricci, E. 2024. Harnessing Large Language Models for Training-free Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18527–18536.

Zhang, H.; Xu, X.; Wang, X.; Zuo, J.; Han, C.; Huang, X.; Gao, C.; Wang, Y.; and Sang, N. 2024. Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm. *arXiv preprint arXiv:2406.12235*.

Zhao, Y.; Deng, B.; Shen, C.; Liu, Y.; Lu, H.; and Hua, X.-S. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1933–1941.

Zhou, H.; Yu, J.; and Yang, W. 2023. Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection. *arXiv preprint arXiv:2302.05160*.

Zhu, J.; Ding, C.; Tian, Y.; and Pang, G. 2023. Anomaly Heterogeneity Learning for Open-set Supervised Anomaly Detection. *arXiv preprint arXiv:2310.12790*.

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