Remedying the Curse of Autonomous Driving: VLM Driven Training-Free Framework for Efficient Long-Tail Video Detection

Anonymous CVPR submission

Paper ID ***

Abstract

Autonomous driving has made remarkable progress, and we have finally witnessed its real-world deployment. These advancements have been driven by training models on an increasing scale of data. Nowadays, data collection is streamlined, with vast amounts—amounting to $\sim$ 100 years of driving data—collectible in a single day. However, most of this data is routine and does not help the model to generalize; on the contrary, it might bias models towards routine driving scenarios. This creates a critical bottleneck in model generalization, as systems remain vulnerable to long-tail scenarios that are rare but safety-critical. Hence, training models on such scenarios remains critical for safe and scalable deployment.

Despite extensive research on end-to-end driving models, a systematic method for efficiently detecting these long-tail events from large-scale dataset is missing. In this work, we propose a novel training-free two-stage framework based on vision language models. In the first stage, a small language model is used to process videos for scenario summarization. In the second stage, the scenario summarization from the first stage is processed by a large language model to rank the video's long-tail relevance. Our two-stage framework is designed to efficiently process industry-scale video datasets and accurately classify the relevance of video segments to long-tail events. Experiment results show our proposed method surpasses counterpart methods by 24% in AUC and provides a $\sim 14\times$ faster inference compared to counterpart multimodal large language models, enabling scalable and targeted data mining for autonomous systems, which is critical for generalizing driving models for deployment. We hope our study paves the way for further research in this critical field.

1. Introduction

In recent years, autonomous driving systems have evolved rapidly, reaching a stage where commercial deployment has finally become a reality after decades of research and development. Traditionally, these systems have followed a modular design paradigm—comprising separate perception, prediction, and planning components—each trained and optimized independently [2, 26]. However, this disjoint optimization often leads to compounding errors, motivating a shift toward end-to-end approaches that directly map raw sensory inputs to driving actions [10, 11, 13, 19, 22]. End-to-end systems themselves have undergone significant evolution, transitioning from cascaded, planning-oriented architectures to Vision-Language-Action (VLA) models—large transformer-based neural networks that integrate perception, reasoning, and control within a single framework [5, 12, 14, 21, 28]. VLAs have recently gained increasing attention and are emerging as a preferred architecture for real-world deployment [24]. With billions to trillions of parameters, these models demonstrate strong generalization capabilities by learning rich, transferable representations.

The success of such generalizable systems stems from large-scale training enabled by fleet-collected driving data. Modern pipelines can accumulate the equivalent of $\sim$100 years of driving data in a single day [24]. Within these datasets, long-tail scenarios—rare yet safety-critical events—are key to achieving robust and deployable autonomous driving. However, these cases constitute $<< 1\%$ of the data, making their identification and utilization extremely challenging [23]. Several prior works have explored building learnable systems to identify long-tail scenarios from large-scale driving data [3, 4]. These works leveraged various techniques such as reconstruction, prediction and generative models for identification of long-tail videos [3, 8, 15]. However, these traditional methods lack generalization, as they are tailored to specific types of long-tail cases, whereas real-world long-tail distributions are inherently open-ended and diverse. Moreover, these conventional vision-only systems lack reasoning capabilities, which further limits their applicability to generalized and complex driving scenarios. An alternative to traditional methods is to leverage powerful multimodal large language

models (MLLMs). However, these models incur substantial latency and memory overhead, resulting in high computational costs that hinder deployment at fleet scale.

To overcome these challenges, we propose long-tail video anomaly detection for autonomous driving (LTVAD-AD), a systematic training-free framework based on vision language models (VLMs) for efficiently detecting long-tail videos from massive real-world driving datasets in an open-ended manner, enabling scalable discovery beyond predefined anomaly types. In particular, our framework adopts a reasoning-driven two-stage system that integrates scalable multimodal understanding with high-level semantic reasoning. In the first stage, a small language model summarizes driving videos into concise textual descriptions that capture key events and context. In the second stage, a large language model reasons over these summaries to assess and rank their relevance to long-tail scenarios. This framework bridges low-level perception and high-level abstraction, enabling efficient large-scale analysis and open-ended discovery of rare, safety-critical driving events.

To evaluate the effectiveness of our approach, we conduct experiments on a proprietary dataset comprising both long-tail and routine driving scenarios, following established practices in the broader video anomaly detection literature. The experimental results demonstrate that our proposed method significantly outperforms counterpart methods by 24% in AUC, and provides a $\sim 14\times$ faster inference compared to larger video-language models ($\sim 10\times$ more parameters) that are impractical for large scale deployment due to their latency and memory requirements. Our main contributions are summarized as follows:

- We propose a systematic training-free framework based on vision language models for efficiently detecting long-tail driving scenarios from large-scale real-world datasets in an open-ended manner.
- We design a reasoning-driven two-stage architecture that integrates a small language model for scenario understanding&summarization, and a large language model for semantic reasoning and long-tail relevance ranking.
- We provide detailed guidelines for constructing proprietary video anomaly detection datasets tailored to autonomous driving applications and future research in this critical domain.
- Extensive experimental results show that our approach significantly outperforms comparable baselines in terms of both long-tail video detection and inference latency, making it ideal for large scale applications.

2. Related Work

2.1. End-to-End Autonomous Driving

End-to-end autonomous driving methods, which map raw sensor inputs directly to control outputs, have received remarkable attention in recent years and have rapidly evolved, paving the way for scalable deployment of autonomous driving systems. UniAD [11] is one of the seminal works that inspired this new wave of research in end-to-end autonomous driving. Unlike traditional stand-alone modular systems, UniAD leverages rasterized representations and integrates perception, prediction, and planning within a unified framework, enabling joint optimization across modules. VAD [13] which uses vectorized scene representation is another important line of work for end-to-end autonomous driving. Following these two works, a plethora of end-to-end works have been proposed. Some notable works include PARA-Drive [22], SparseDrive [19], and Diffusion-Drive [16]. Despite the notable progress , these vision-only end-to-end methods face generalization issues as it lacks the multi-modality and reasoning.

To address the challenges in vision-only E2E systems, multimodal large language models have recently been investigated. Initial work in the field distilled features from large vision language models into vision-only end-to-end systems to enhance the generalization performance [9, 20]. More recent works have focused on vision language action models (VLAs) which takes input as videos, language (as well as other information such as audio or vehicle kinematics) and directly outputs actions [5, 12, 24, 28]. These models have achieved state-of-the-art performance and strong generalization as these large transformer neural networks have $\sim$ trillion number of parameters and trained on vast scale of diverse fleet data.

2.2. Video Anomaly Detection

Long-tail scenario detection in autonomous driving is a video anomaly detection task where any non-routine interesting driving scenario can be considered as anomaly. Early works in the field leveraged classical CNN or GAN based methods for this task. Among these works, [8] employs an autoencoder and use reconstruction error as an indicator for anomaly score as reconstruction error for non-routine driving scenarios is anticipated to be high. In another work, [3] performs future frame prediction, and assign a high anomaly score if there is a non-predictable relevant object in a relevant location. However, such works lack reasoning and limit the long-tail scenarios to a subset, thus are not able to generalize to open-ended nature of long-tail scenarios.

Given the limitations of conventional methods, more recent works [23] leverages the multimodal LLMs for extracting long-tail scenarios, but does not provide any methodological aspects on how MLLMs are leveraged. In this work, we provide a systematic and detailed approach for efficiently mining long-tail videos using vision language models. Moreover, proposed approach is training free, further enabling generalization to open-ended nature of long-tail scenarios.

3. Methodology

3.1. Problem Set-up

Video anomaly detection (VAD) systems traditionally aim to *learn* a classifier model by training on a ground truth dataset labeled video-level as normal or anomaly. However, challenge in curating long-tail videos and open-ended nature of long-tail videos renders such methods impractical as they bias towards training scenarios and fail to well-generalize to unseen long-tail scenarios [4].

Hence, in this study, we present a *training-free* framework leveraging pretrained MLLMs that is not bounded by training distribution, and can leverage advanced reasoning capabilities to generalize to any long-tail scenario. In particular, given a driving scenario video, $\mathbf{V}$, our goal is to design a training-free framework f to directly classify a given video as normal or anomaly.

3.2. Vanilla LTVAD-AD

In a vanilla setup, multimodal large language models (MLLMs) capable of processing videos [1] can be directly employed for long-tail video classification through carefully designed prompts. Given a video $\mathbf{V}$ and a scoring prompt P_{score}, the MLLM can be instructed to produce an anomaly score s, reflecting the likelihood that the video depicts a rare or interesting event (Fig. 1 baseline). To achieve this, we craft the P_{score} (template in Fig. 1, full prompt is provided in Appendix) to enforce the MLLM to output a score $s \in 0.0, 0.1, 0.2, \ldots, 1.0$, representing 11 uniformly sampled values in the interval $[0, 1]$, where 0 denotes normal and 1 denotes anomaly, formulated as:

$$s = f_{\text{MLLM}}(\mathbf{V}, P_{\text{score}}), \tag{1}$$

where $f_{\text{MLLM}}(\cdot)$ denotes the multimodal LLM inference function. The resulting score s is then thresholded to classify the video as either normal or anomaly:

$$\hat{y} = \begin{cases} 1, & \text{if } s \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where τ is a predefined decision threshold.

While this setup (with a well-crafted P_{score}) can enable large MLLMs to achieve good performance, processing videos through such models is computationally expensive and impractical for large-scale deployment due to their high inference cost and memory demands. Conversely, small language models (SLMs) are computationally efficient and scalable, but their performance remains limited even with carefully designed prompts, owing to weaker reasoning and representation capabilities. This trade-off between capability and scalability motivates the need for a more efficient yet reasoning-aware framework for long-tail video understanding.

3.3. LTVAD-AD

To address the trade-offs described in 3.2, we propose a systematic training-free two-stage framework based on multimodal language models for efficiently detecting long-tail videos in an open-ended manner. As illustrated in Fig. 1c, the core idea is to decouple video understanding and reasoning across two complementary stages, thereby achieving both scalability and interpretability.

Stage 1: SLMs as Scenario Descriptor. In the first stage, SLMs summarizes raw driving videos into concise textual descriptions that capture key events and context. Long-tail scenarios—such as emergency vehicles, wildlife crossings, or other unexpected events—naturally emerge as non-routine information in these summaries. Given the limited capability and context window of SLMs, we split each long video into shorter, non-overlapping segments before processing to maintain temporal coherence and scalability.

Given a video segment $\mathbf{V}_i$ and a description prompt P_{desc}, the SLM generates a textual scenario description

$$T_i = f_{\text{SLM}}(\mathbf{V}_i, P_{\text{desc}}), \tag{3}$$

where $f_{\text{SLM}}(\cdot)$ denotes the SLM model inference function and P_{desc} is simply "Please describe the given video.".

This stage enables efficient large-scale video summarization by offloading heavy visual processing to a lightweight model. Our segmenting strategy further enhances summarization quality by improving temporal focus and reducing information dilution across long sequences. Although these SLMs lack advanced reasoning capabilities, the resulting descriptions capture salient scene elements and temporal context, providing a compact yet informative representation for the second-stage reasoning model.

Stage 2: LLM-based Reasoning and Scoring. In the second stage, a large language model (LLM) performs high-level reasoning over the textual scenario descriptions generated by the SLM. Each description T_i is paired with a reasoning scoring prompt $\hat{P}_{\text{score}}$ (almost identical to P_{score}, see appendix for details), instructing the model to evaluate whether the segment depicts a rare or anomalous driving event. The LLM produces a relevance score s_i representing the likelihood of the segment belonging to the long-tail distribution:

$$s_i = f_{\text{LLM}}(T_i, \hat{P}_{\text{score}}), \tag{4}$$

where $f_{\text{LLM}}(\cdot)$ denotes the LLM. For each video, we aggregate segment-level scores by taking the maximum score across all segments, i.e. $s = \max_i s_i$. The videos are then scored using the same thresholding strategy as in the vanilla setup, where videos with $s \geq \tau$ are identified as anomaly.

By reasoning directly over textual scenario descriptions that capture key events and contextual details, this stage enables efficient large-scale inference without processing raw videos.

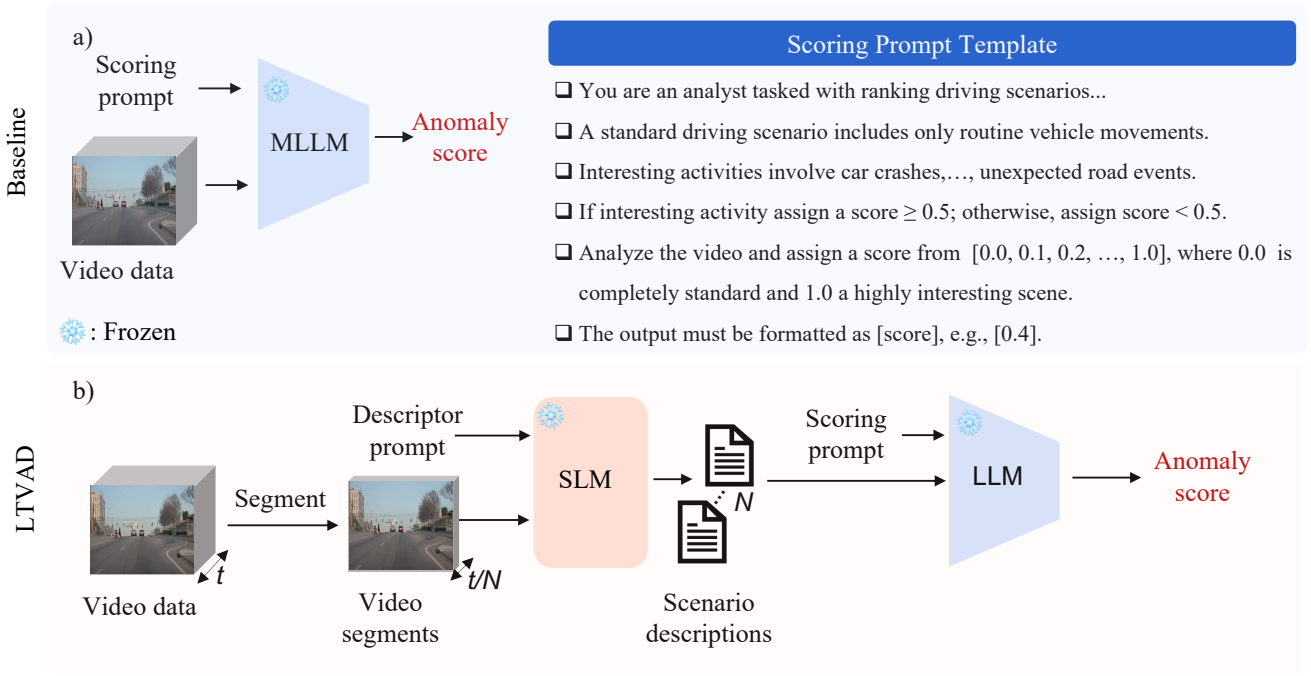

Figure 1. **Overview of proposed framework.** a) Baseline set-up for long-tail video anomaly detection. b) Proposed two stage framework, where first stage temporally segments the video into smaller lengths and perform scenario description using SLMs; second stage employs an LLM to perform anomaly scoring, followed by max pooling over segments for the final anomaly score.

4. Experiments and Results

4.1. Dataset

Given the lack of publicly available datasets for video anomaly detection in autonomous driving, we develop a detailed labeling framework that combines model-based pseudo-labeling with subsequent manual review.

Data Sources and Preprocessing. Our dataset, illustrated in Fig. 2, is collected from real-world driving under diverse conditions—ranging from day to night and across varying weather—using front-view camera recordings. Raw videos are 30 second each, recorded at a frame rate of 30 Hz and a resolution of 1220 × 1936 pixels.

Model-Based Labeling. To determine whether a clip contains an interesting (abnormal) scenario, we adopt a model-based labeling approach. Each 30-second MP4 clip is analyzed using Qwen2.5-VL-72B-Instruct, guided by a structured prompt defining a taxonomy of ten high-level categories of rare or safety-critical driving situations. The model is instructed to:

1. Output `Scenario Detected: [Category] – [Subtype]` only when highly confident, or
2. Return `"no"` if no scenario is detected or confidence is insufficient.

This conservative instruction ensures that only high-confidence detections are retained, effectively reducing label noise from uncertain cases.

Category Taxonomy. The prompt defines the following categories and example subtypes:

1. Unpredictable Pedestrian Behavior: jaywalking, pedestrians revealed from occlusion, children running, impaired pedestrians.
2. Unusual Vehicle Actions: abrupt stops or lane changes, wrong-way driving, emergency vehicles with active sirens.
3. Complex Traffic Interactions: complex roundabouts, non-standard intersections.
4. Road Infrastructure Anomalies: active construction zones, unlit or malfunctioning signals, detours, degraded lane markings.
5. Obstacles and Debris: fallen cargo, road debris, cones, barriers, or temporary objects.
6. Unusual Road Users: tricyclists, e-scooters, agricultural vehicles, non-standard vehicles.
7. Animals on the Road: wildlife crossings, domestic pets, livestock on rural roads.
8. Adverse Weather Conditions: heavy rain, fog, flooding, snow, or intense sun glare.
9. Pickup/Drop-Off Challenges:congested curb zones, long passenger wait times.
10. Emergency Situations and Environmental Extremes: active accident scenes, authority-imposed road closures, blocked traffic.

Each 30-second clip is processed independently. All clips labeled as “interesting” are then manually reviewed

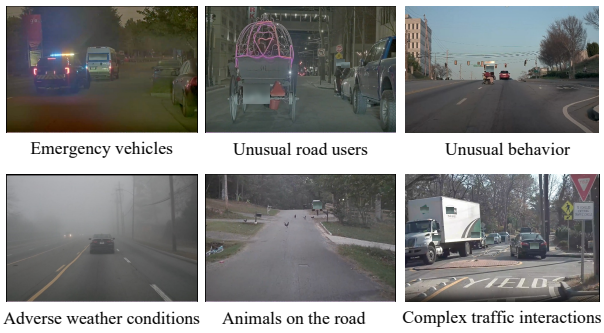

Figure 2. Examples from the category taxonomy.

to confirm correctness and remove false positives. A subset of clips labeled as "no" is also audited to estimate false negatives and refine the model threshold if necessary.

Dataset Composition. Following established practices in video anomaly detection [18, 25], we construct a validation dataset of 300 video clips from a large corpus of aforementioned preprocessed driving videos, tailored for anomaly detection in autonomous driving. We note that training based models requires a much larger annotated dataset, but such large training datasets are not needed for training-free methods as proposed LTVAD [25]. Consistent with [18], our proprietary dataset is balanced, containing 150 normal clips (label = 0) and 150 long-tail edge-case clips (label = 1). The edge-case subset in validation set covers all variety of rare or critical driving scenarios listed in the taxonomy, selected to achieve balanced coverage across categories. This ensures that the dataset reflects a diverse set of long-tail conditions relevant to autonomous driving safety.

4.2. Experimental Set-up and Metrics

Set-up. In this study, we leverage open-source pretrained MLLMs for our experiments. In particular, for stage 1 of our framework, we use Qwen2.5-VL-7B as small MLLM [1] , and we employ Qwen2.5-72B-Instruct for stage 2. For baseline comparisons described in Sec. 3.2, we include Qwen2.5-VL-7B and InternVL3-8B as SLM baselines; Qwen2.5-VL-72B, and InternVL3-78B as large MLLM baselines. Given the training-free aspect of our methodology, we did not change any parameter such as fps (default 2) during inference. For scoring threshold, we use $\tau = 0.5$ to strike a balance between recall and precision. All experiments were conducted on four NVIDIA H100 GPUs (80 GB each) using the HuggingFace Transformers library with FlashAttention and BF16 precision.

Metrics. Following established practices in video anomaly detection, we report the Area Under the ROC (receiver operating characteristic) Curve (AUC) as our primary evaluation metric, which plots the true positive rate (TPR) against false positive rate (FPR) [4, 25]. *We note that AUC is agnostic to threshold value* τ. In addition, we report accuracy, precision, recall, and F1-score for completeness. We further highlight recall (i.e., TPR) and F1 as auxiliary metrics alongside AUC, as they are particularly important for evaluating performance in long-tail anomaly detection scenarios for autonomous driving, where missing rare events can have severe consequences.

4.3. Experimental Results

Main results. Tab. 1 shows the quantitative comparisons between proposed LTVAD-AD with counterpart SLMs and large MLLM models. The experimental results show that proposed framework significantly ourperforms counterpart SLM models (Qwen2.5-VL-7B) with a 24% improvement in AUC and further improvements in recall and F-1 score. Furthermore, proposed method able to achieve similar performance (slightly better) in terms of both AUC and Recall compared to Qwen2.5-VL-72B variant without any bells and whistles. More importantly, in terms of computational efficiency, our method achieves a substantial improvement over large MLLMs. In particular, the average processing time for a single 30-sec video clip using a large MLLM (Qwen2.5-VL-72B) is $\sim$22 s, whereas our proposed two-stage framework processes the same input in $\sim$1.5 s, resulting in a $\sim 14\times$ gain. This efficiency gain arises as the first stage offloads heavy visual processing to a lightweight SLM for video understanding, while the second stage performs text summarization in large batches without incurring significant memory overhead.

Method	Accuracy	Precision	Recall	F1	AUC
InternVL3-8B	0.69	0.83	0.48	0.61	0.78
InternVL3-78B	0.9	0.89	0.91	**0.9**	0.92
Qwen2.5-VL-7B	0.74	0.89	0.54	0.67	0.75
Qwen2.5-VL-72B	0.88	0.84	0.95	0.89	0.92
LTVAD-AD	0.85	0.79	**0.96**	0.87	**0.93**

Table 1. Quantitative comparison of the proposed LTVAD-AD model with counterpart SLM and MLLM methods.

Ablation study. In this section, we investigate the influence key components in our proposed LTVAD-AD framework, namely temporally segmenting videos, including ranking guidance and in-context examples in the scoring prompt.

Tab. 2 shows ablation for various segmenting lengths. In this study, raw videos (30 sec) are temporally segmented into smaller non-overlapping fixed size segments. Temporally segmenting video size by 2 provides improvement over no segmenting. Further segmenting decreases the AUC and F-1 scores.

Tab. 3 shows the comparison between scoring prompts with and without enforced guidance for ranking videos as normal or anomaly. In particular, we ablate the use of "If an interesting activity is present, make sure to assign a score $\geq$ 0.5;otherwise assign a score $<$0.5" in the prompt. Without

Method	Accuracy	Precision	Recall	F1	AUC
no segmenting	0.85	0.84	0.88	0.86	0.90
segmenting/2	0.85	0.79	**0.96**	**0.87**	**0.93**
segmenting/3	0.76	0.69	0.94	0.79	0.89

Table 2. Ablation study on the impact of video length on long-tail detection for the proposed framework.

using this explicit phrase in the prompt, the LLMs tend to be conservative in ranking videos as anomalies, resulting in a huge performance drop in recall and F-1 score. Thus, LLMs needs explicit guidance in correctly ranking videos.

Method	Accuracy	Precision	Recall	F1	AUC
w/o ranking guidance	0.64	0.96	0.29	0.45	0.88
w ranking guidance	0.85	0.84	**0.88**	**0.86**	**0.90**

Table 3. Ablation study on the impact of explicit ranking prompt on long-tail video anomaly detection.

Tab. 4 shows the ablation study on the inclusion & exclusion of in-context examples for interesting activities (e.g., car crashes). Exclusion of in-context examples does not lead to a major performance difference highlighting the LLMs reasoning capability under an explicit scoring guidance.

Method	Accuracy	Precision	Recall	F1	AUC
w/o in-context examples	0.74	0.68	**0.91**	0.78	0.88
w in-context examples	0.85	0.84	0.88	**0.86**	**0.90**

Table 4. Ablation study on the impact of in-context examples in the prompt on long-tail video detection.

Distillation. SLMs are not specifically trained on autonomous driving data, which can lead to hallucinated scenario descriptions and suboptimal performance in downstream anomaly detection. To address this limitation, we perform supervised fine-tuning of the SLMs (Qwen2.5-VL-7B) on an autonomous driving dataset to improve domain alignment and description quality. For this study, we adopted a knowledge distillation approach [27] (illustrated in Fig. 3), where we first employ a stronger multimodal teacher model (Qwen2.5-VL-72B) to generate scenario descriptions for each training video. Subsequently, these generated descriptions are treated as pseudo-labels and used to fine-tune the SLMs. To preserve the generalization ability of the base model, we adopt parameter-efficient fine-tuning with LoRA (rank=64, α=64), optimizing a standard autoregressive language modeling loss [17]. The model is fine-tuned for 2 epochs on 5K video–description pairs. To illustrate the efficiency of the fine-tuning, we rerun the segmenting/3 experiment in Tab. 2 with a fine-tuned SLM backbone. Experimental results on Tab. 5 show that, fine-tuning SLM backbone enhance both accuracy and precision, while also showing improvements for F1 and AUC scores.

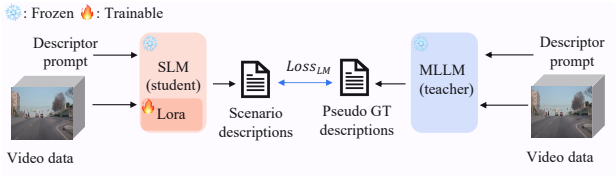

Figure 3. Distillation framework for scenario description generation. Pseudo ground truth (GT) descriptions from a teacher model is used for supervised fine-tuning of student SLM model, with objective of minimizing language modeling loss.

Method	Accuracy	Precision	Recall	F1	AUC
No Fine-tuning	0.76	0.69	**0.94**	0.79	0.89
Fine-tuning	**0.81**	**0.75**	0.93	**0.83**	**0.91**

Table 5. Quantiative comparison of SLM fine tuning vs without fine tuning for long-tail video anomaly detection.

5. Limitations

The presented framework currently only focuses on front-view cameras as front-view contain the most vital information for driving [7]. Given fleet-data generally collect multi-view data, additional views might positively help for better mining long-tail scenarios. However, a concern here is to find a balance as non-front-view cameras might also introduce large amount of false positives as they do not impact driving as much of the front-views. Another potential pitfall is the lack of explicit motion information extraction and incorporation into scenario descriptions that can be used for long-tail video anomaly detection [6]. This can be especially useful for understanding nuanced scenarios such as aggressive cut-ins. Lastly, in this study, we adopted a uniform sampling for frame selection. However, a lightweight method for selecting key frames that capture critical moments can further enhance downstream long-tail detection performance. We aim to explore these directions as part of our future work.

6. Conclusion

In this study, we presented LTVAD-AD, a novel long-tail video detection framework for autonomous driving. LTVAD-AD consists of two stages, where first stage leverages small language models for scenenario summarization and second stage employs an LLM to perform anomaly detection on scenario summarization from the first stage. Extensive experiments on curated driving dataset shows the efficiency and accuracy of proposed method in detecting long-tail driving videos over counterpart SLM & MLLM baselines. The proposed approach is versatile as information from first stage is generic and can be used various other tasks such as embedding and retrieval tasks.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 5

[2] Sagar Behere and Martin Törngren. A functional architecture for autonomous driving. In *Proceedings of the first international workshop on automotive software architecture*, pages 3–10, 2015. 1

[3] Jan-Aike Bolte, Andreas Bar, Daniel Lipinski, and Tim Fingscheidt. Towards corner case detection for autonomous driving. In *2019 IEEE Intelligent vehicles symposium (IV)*, pages 438–445. IEEE, 2019. 1, 2

[4] Jasmin Breitenstein, Jan-Aike Termöhlen, Daniel Lipinski, and Tim Fingscheidt. Corner cases for visual perception in automated driving: Some guidance on detection approaches. *arXiv preprint arXiv:2102.05897*, 2021. 1, 3, 5

[5] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 1, 2

[6] Yulu Gan, Ligeng Zhu, Dandan Shan, Baifeng Shi, Hongxu Yin, Boris Ivanovic, Song Han, Trevor Darrell, Jitendra Malik, Marco Pavone, et al. Foundationmotion: Auto-labeling and reasoning about spatial movement in videos. *arXiv preprint arXiv:2512.10927*, 2025. 6

[7] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems*, 2024. 6

[8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 1, 2

[9] Deepti Hegde, Rajeev Yasarla, Hong Cai, Shizhong Han, Apratim Bhattacharyya, Shweta Mahajan, Litian Liu, Risheek Garrepalli, Vishal M Patel, and Fatih Porikli. Distilling multi-modal large language models for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27575–27585, 2025. 2

[10] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 1

[11] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1, 2

[12] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 1, 2

[13] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 1, 2

[14] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. A survey on vision-language-action models for autonomous driving. *arXiv preprint arXiv:2506.24044*, 2025. 1

[15] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 1

[16] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12037–12047, 2025. 2

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 6

[18] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 5

[19] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8795–8801. IEEE, 2025. 1, 2

[20] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conference on Robot Learning*, pages 4698–4726. PMLR, 2025. 2

[21] Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025. 1

[22] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15449–15458, 2024. 1, 2

[23] Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, et al. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025. 1, 2

[24] Burhan Yaman, Yunsheng Ma, and Xin Ye. A peek into tesla's autonomous future: Core tech revealed by vp ashok

elluswamy at iccv 2025 wdfm-ad. *Medium*, 2025. Accessed: 2025-11-06. 1, 2

[25] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024. 5

[26] Jingyuan Zhao, Yuyan Wu, Rui Deng, Susu Xu, Jinpeng Gao, and Andrew Burke. A survey of autonomous driving from a deep learning perspective. *ACM Computing Surveys*, 57(10):1–60, 2025. 1

[27] Yingxue Zhou, Tan Zhu, Tao Zeng, Zigeng Wang, and Wei Shen. Learning from llm agents: In-context generative models for text casing in e-commerce ads. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1122–1133, 2025. 6

[28] Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025. 1, 2

A. Appendix

Prompts for LTVAD-AD. The complete scoring prompt (P_{score}) used for baselines in our study:

```
 You are an analyst tasked with
ranking driving scenarios based
on the presence of interesting
activities.
A standard driving scenario
includes only routine vehicle
movements.
Interesting activities
involve car crashes, emergency
vehicles, vulnerable road users,
unusual behavior of vehicles,
adverse weather conditions,
temporary road hazards, active
construction signs, or any other
unexpected road events.
Analyze the given video and
assign a score from the
following fixed list: [0.0,
0.1, 0.2, 0.3, 0.4, 0.5, 0.6,
0.7, 0.8, 0.9, 1.0], where 0.0
means the scene is completely
standard and 1.0 indicates a
highly interesting scene.
If any of the interesting
activity is present, make sure
to assign a high score (≥ 0.5);
otherwise, assign a low score
(< 0.5).
The output must be only one
number from the fixed list and
formatted as [score], e.g.,
[0.4].
```

In our proposed framework, the first stage description prompt (P_{desc}) was the following:

```
Please describe the given video.
```

The second-stage reasoning scoring prompt ($\hat{P}_{score}$) was identical to P_{score}, with the following modification to incorporate the scenario summarization from the first stage:

```
 ...
The video has been summarized as
follows: first-stage scenario
summarization.
Analyze the video summary,
and assign a score from the
following fixed list..
...
```