This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

EchoTraffic: Enhancing Traffic Anomaly Understanding with Audio-Visual Insights

Zhenghao Xing[1]*, Hao Chen[1]*, Binzhu Xie[1], Jiaqi Xu[1], Ziyu Guo[1],
Xuemiao Xu[2], Jianye Hao[3], Chi-Wing Fu[1], Xiaowei Hu[4†], and Pheng-Ann Heng[1]
[1] The Chinese University of Hong Kong [2] South China University of Technology
[3] Tianjin University [4] Shanghai Artificial Intelligence Laboratory

Abstract

Traffic Anomaly Understanding (TAU) is essential for improving public safety and transportation efficiency by enabling timely detection and response to incidents. Beyond existing methods, which rely largely on visual data, we propose to consider audio cues, a valuable source that offers strong hints to anomaly scenarios such as crashes and honking. Our contributions are twofold. First, we compile AV-TAU, the first large-scale audio-visual dataset for TAU, providing 29,865 traffic anomaly videos and 149,325 Q&A pairs, while supporting five essential TAU tasks. Second, we develop EchoTraffic, a multimodal LLM that integrates audio and visual data for TAU, through our audio-insight frame selector and dynamic connector to effectively extract crucial audio cues for anomaly understanding with a two-phase training framework. Experimental results on AV-TAU manifest that EchoTraffic sets a new SOTA performance in TAU, outperforming the existing multimodal LLMs. Our contributions, including AV-TAU and EchoTraffic, pave a new direction for multimodal TAU.

1. Introduction

Traffic Anomaly Understanding (TAU) is receiving increasing attention, given the practical need of detecting, analyzing, and responding to traffic anomalies such as accidents. Powered by large language models, the goal of TAU is not only to recognize relevant entities but also to understand the underlying causes and produce timely responses. Access to these information is critical for alerting emergency services, lifesaving, reducing damage, enhancing public safety, and improving transportation efficiency [58, 68].

At present, current methods [10, 19, 44, 46, 47, 49, 54] focus largely on detecting anomalies in traffic scenarios, often neglecting the underlying causes. Leveraging large language models, recent methods [8, 50, 67] are able to extend beyond detection to contextual understanding of general anomaly events. Yet, the existing methods remain limited to visual inputs, overlooking audio, which is a readily accessible but crucial modality for TAU. In real-world situations, traffic anomalies usually involve audio cues such as

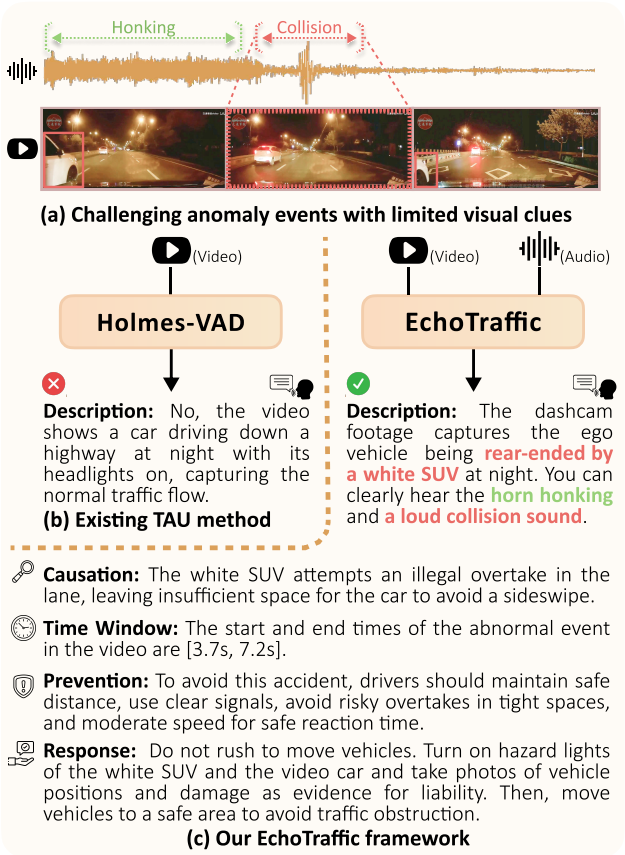

Figure 1. Comparing traffic anomaly understanding in a challenging scenario with limited visual cues. (b) Holmes-VAD [67], a very recent approach, still falls short of identifying the out-of-view anomaly. (c) Our EchoTraffic can provide a comprehensive understanding by simultaneously considering audio and visual signals.

*Joint first authors.
†Corresponding author (huxiaowei@pjlab.org.cn).

crashes, braking, or honking, which occur largely beyond the camera's field of view, as depicted in Fig. 1 (a). Without audio data, it is hard to fully capture and interpret the anomaly events, limiting our ability to recognize and analyze the anomaly effectively, as shown in Fig. 1 (b).

Also, detailed textual annotations are crucial for obtaining accurate reasoning and appropriate responses to traffic anomalies. Yet, as Table 1 shows, *existing datasets for TAU largely lack audio information and comprehensive textual annotations.* Importantly, *"how to effectively leverage and integrate audio and visual signals in a deep model for traffic anomaly understanding"* remains an unsolved challenge.

In this work, we first present **AV-TAU**, the first dataset compiled for Audio-Visual Traffic Anomaly Understanding. AV-TAU provides **29,865** real-world traffic anomaly videos with audio-visual insights and comprehensive annotations in the form of **149,325** Question & Answer (Q&A) pairs across **five** key TAU tasks, illustrated in Fig. 1 (c): (i) identifying and describing anomalies through multimodal data, (ii) explaining the underlying causes, (iii) pinpointing the time windows of anomalies, (iv) recommending effective prevention strategies, and (v) providing suitable response solutions. To our knowledge, AV-TAU is the largest TAU dataset and the first that incorporates audio data.

Second, we develop **EchoTraffic**, a multimodal large language model, specifically designed for understanding and interpreting traffic anomalies. EchoTraffic encompasses several key innovations: an *audio-insight frame selector* to identify critical frames based on audio cues, *specialized visual and audio encoders* to capture essential features, and an *audio-visual dynamic connector* to adaptively fuse these features for a balanced representation. The fused multimodal output is then processed by a large language model decoder, enabling EchoTraffic to generate natural language descriptions of the detected anomalies.

We train EchoTraffic in two phases: first to align the visual, auditory, and language modalities using the large-scale InternVid-10M dataset [56], then to instruction-tune the model on our AV-TAU dataset to enhance its adaptability and accuracy on anomaly understanding in diverse traffic scenarios. Compared to other baseline models, EchoTraffic achieves state-of-the-art (SOTA) performance consistently for both traditional and GPT [1] evaluated metrics, underscoring its exceptional capability in traffic anomaly understanding. Further, experimental results also confirm the effectiveness of incorporating audio-visual information, which significantly boosts the performance beyond visual modality alone. Below, we summarize our contributions:

- We formulate **AV-TAU**, the first large-scale audio-visual traffic anomaly dataset, comprising 29,865 videos and 149,325 Q&A pairs to support traffic anomaly understanding through both visual and auditory cues.
- We develop **EchoTraffic**, a multi-modal large language model, integrating audio-visual insights for advanced traffic anomaly understanding in various traffic scenarios.
- Experimental results show that EchoTraffic sets a new SOTA performance in traffic anomaly understanding tasks, paving a new direction on multimodal TAU. Our dataset and code are publicly available at https://github.com/HarryHsing/EchoTraffic.

2. Related Work

2.1. Traffic Anomaly Datasets

Traditional traffic anomaly datasets focus on detecting and locating accidents by finding relevant time frames. Dashcam datasets from vehicle-mounted cameras, like DAD [5], A3D [61], and CCD [3], capture accident-related frames, with CCD further categorizing accident causes. Extending A3D, DoTA [62] provides detailed category labels and supports unsupervised detection, while DADA2000 [9] assesses driver attention during incidents, and ROL [25] aids early risk detection. ADV [30] and TRA [37] offer video-level labels without specific incident details. Surveillance datasets, though less common, offer broader views from fixed cameras, as in CADP [48], TAD-2 [59], and TADS [4]. Hybrid datasets [41, 58, 64] combine dashcam and surveillance footages for anomaly analysis. Synthetic datasets [2, 27, 39, 54, 55] supplement real data, though bridging the gap to real scenes remains challenging [11, 12]. Overall, existing datasets lack detailed text annotations, which are crucial for causal reasoning in TAU.

The rise of Multi-modal Large Language Models has led to new general video anomaly datasets [8, 50, 67] enriched with text descriptions, though traffic-specific datasets remain limited. While some traffic anticipation datasets [22, 26, 29, 42] include text annotations, they focus mainly on near-miss incidents rather than actual accidents. SUTD-TrafficQA [58] stands out as the first large-scale dataset for TAU, providing accident reasoning and prevention insights through a Q&A format. MM-AU [12] further advances TAU research with text annotations.

Our dataset advances traffic anomaly research by integrating audio with visual data for richer insights. As the largest dataset to date, it includes the most video samples, Q&A pairs, and unique response annotations, bridging detection and practical response strategies. Table 1 provides a detailed comparison of traffic anomaly datasets.

2.2. Traffic Anomaly Understanding Methods

TAU seeks to detect unexpected events and analyze their causes in traffic scenarios using video data. However, prior methods [10, 19, 44, 46, 47, 49, 54] have focused mainly on detection, often neglecting the description and causal analysis of anomalies. Recently, researchers have developed MLLMs for enhanced anomaly understanding [8, 34,

Table 1. Comparing datasets for traffic anomaly detection and understanding. Note that existing datasets differ in the number of clips, frames, audio availability, temporal annotations, anomaly categories, Q&A pairs, and observation views. Observation views include Dashcam (car-mounted cameras), Surveillance (fixed, static cameras), and Camera (handheld devices). Our AV-TAU dataset provides the largest number of clips, frames, and Q&A pairs, along with annotated audio information across multiple observation views.

Datasets	Years	Clips	Frames	Audio	Temporal Annotation	Anomaly Categories	Q&A Pairs	Observation View	Real/Synthetic
DAD [5]	2016	1,750	175K	-	✓	✓	-	Dashcam	Real
CADP [48]	2018	1,416	518K	-	✓	✓	-	Surveillance	Real
A3D [61]	2019	3,757	208K	-	✓	✓	-	Dashcam	Real
GTACrash [27]	2019	7,720	154K	-	✓	-	-	Dashcam	Synthetic
VIENA^2 [2]	2019	15,000	2.25M	-	✓	-	-	Dashcam	Synthetic
DVAD [23]	2019	60	10K	-	-	-	-	Dashcam	Real
Drive-Anomaly106 [73]	2019	106	11K	-	✓	✓	-	Dashcam	Real
RetroTrucks [20]	2020	474	36K	-	-	-	-	Dashcam	Real
ADV [30]	2020	100	10K	-	-	✓	-	Dashcam	Real
CTA [64]	2020	1,935	-	-	✓	✓	-	Dashcam, Surveillance	Real
TaskFix [24]	2021	-	1K	-	-	-	-	Surveillance	Real
DADA-2000 [9]	2021	2,000	658K	-	✓	✓	-	Dashcam	Real
DADA-seg [69]	2021	313	12K	-	✓	-	-	Dashcam	Real
TAD-1 [41]	2021	500	540K	-	-	✓	-	Dashcam, Surveillance	Real
CCD [3]	2021	1,381	75K	-	✓	-	-	Dashcam	Real
SUTD-TrafficQA [58]	2021	10,080	1.90M	-	-	-	62,535	Dashcam, Surveillance	Real
YouTubeCrash [28]	2021	122	2K	-	✓	-	-	Dashcam	Real
USDC [63]	2022	180	43K	-	✓	-	-	Dashcam	Real
Traffics [15]	2022	29	5K	-	✓	✓	-	Surveillance	Real
MP-RAD [54]	2022	2,000	366K	-	✓	-	-	Surveillance	Synthetic
TRA [37]	2022	560	43K	-	✓	✓	-	Dashcam	Real
DoTA [62]	2022	5,586	732K	-	✓	✓	-	Dashcam	Real
TAD-2 [59]	2022	333	298K	-	✓	✓	-	Surveillance	Real
DeepAccident [55]	2023	-	57K	-	✓	✓	-	Dashcam	Synthetic
ROL [25]	2023	1,000	100K	-	✓	-	-	Dashcam	Real
CTAD [39]	2023	1,100	792K	-	✓	-	-	Surveillance	Synthetic
TADS [4]	2024	966	259K	-	✓	✓	-	Surveillance	Real
MM-AU [12]	2024	11,727	2.19M	-	✓	✓	58,650	Dashcam	Real
AV-TAU (Ours)	**2024**	**29,865**	**3.16M**	✓	✓	✓	**149,325**	**Dashcam, Surveillance, Camera**	**Real**

40, 50, 60, 65, 67, 72]. For example, Zanella et al. [65] present LAVAD, a language-based training-free approach using LLMs to interpret scene descriptions, while Zhang et al. [67] introduce Holmes-VAD by fine-tuning MLLMs with LoRA [21] for enhanced detection. Tang et al. [50] develop HAWK, which incorporates motion data to improve video anomaly understanding with MLLMs.

These existing methods, however, overlook audio, which is critical for understanding various anomalies detectable only through auditory cues [14]. To this end, we develop EchoTraffic to integrate visual and auditory data to enhance the understanding of traffic anomalies in videos.

3. AV-TAU Dataset

We build AV-TAU, the first large-scale Audio-Visual dataset specifically designed to enhance Traffic Anomaly Understanding (TAU). Altogether, AV-TAU provides **29,865** anomalous traffic videos and **149,325** targeted Q&A pairs. To our best knowledge, AV-TAU stands as the largest dataset for TAU, uniquely integrating audio and visual data and setting a new benchmark for multimodal TAU.

3.1. Task Formulation

To comprehensively tackle the TAU task, we define five key subtasks to identify, analyze, and respond to traffic anomalies. Each subtask is detailed below.

(i) What - Multimodal Understanding. It assesses the model's ability to interpret both visual and audio information, aiming to accurately describe anomalies, identify involved objects, analyze the environment, and detect sounds such as collisions or honking. It evaluates the model's comprehensive, multimodal perception of traffic anomalies.

(ii) Why - Causal Inference. This subtask challenges the model to analyze and infer the causes of traffic anomalies, testing its ability to integrate audio-visual information for effective causal reasoning about anomaly origins.

(iii) When - Anomaly Timing. This subtask requests the model to identify the start and end points of the anomalies in traffic videos, calling for the model's abilities on fine-grained detection and temporal perception.

(iv) How - Prevention Strategy. This subtask prompts the model to propose preventive strategies for the detected anomalies, challenging it to perform reasoning and predictive analysis to suggest practical safety measures.

(v) How - Event Response. This subtask assesses the model's ability to formulate appropriate responses to the detected anomalies, requiring advanced reasoning to suggest suitable actions and interventions for specific traffic events.

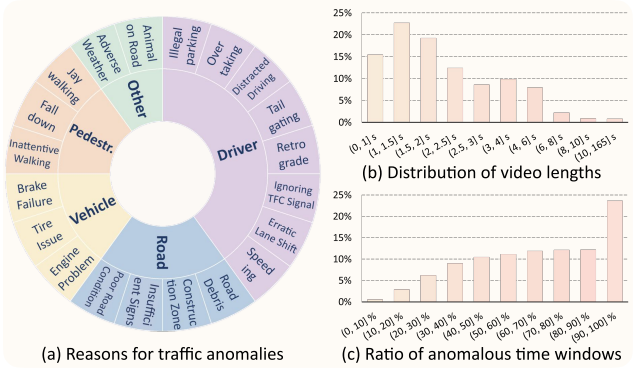

Figure 2. Statistics of the AV-TAU dataset. (a) shows various anomaly reasons, (b) shows video length distributions, and (c) shows the proportion of anomalous time windows.

3.2. Dataset Construction and Statistics

To create a high-quality resource for TAU research, we implement a rigorous, multi-stage dataset construction process, employing 65 trained annotators to ensure thorough coverage, accurate annotation, and validation at each stage.

Data Collection and Filtering. We sourced videos from various online platforms (YouTube, Bilibili, Twitter, Tencent, etc.) using targeted textual queries to obtain a broad spectrum of traffic anomalies (see keywords in Fig. 2 (a)). Annotators then manually filtered this initial collection, retaining only videos that clearly depict traffic anomalies with relevant audio, leading to a focused and informative dataset.

Annotation Quality Control. Each video was meticulously trimmed to capture the causes and consequences of anomalies, with annotators creating text-based answers for the five key tasks (see Fig. 1 (c)) to form Q&A pairs. Also, a cross-validation process is conducted for consistency: each set of labels was independently reviewed by three annotators, with any discrepancies resolved by the authors.

Dataset Statistics. AV-TAU provides 29,865 videos with 3.16 million frames, totaling 31.17 hours, with an average of 108.5 frames per video. The dataset spans five main anomaly categories, further divided into 20 subcategories (Fig. 2 (a)). It features diverse distributions in video lengths (Fig. 2 (b)) and anomaly time windows (Fig. 2 (c)), offering a rich foundation for the research and development of robust, multimodal traffic anomaly analysis.

4. Methodology

Existing video understanding models [35, 66] typically use uniform frame sampling, overlooking the fact that anamolous scenarios happen only in certain frames that are critical. Often, these key moments stand out not only visually but also audibly, with distinct audio signals. To advance traffic video anomaly understanding, we introduce

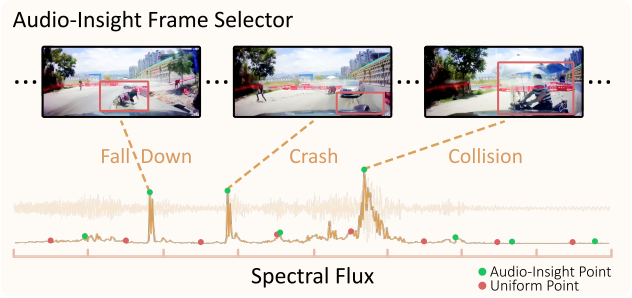

Figure 3. Our Audio-Insight Frame Selector (Green point) employs spectral flux to capture key traffic events, effectively highlighting critical frames missed by the uniform selector (Red point).

EchoTraffic, a novel approach that integrates audio cues to enhance anomaly comprehension.

4.1. Audio-Insight Frame Selector

We design the Audio-Insight Frame Selector to identify anomalous moments in traffic videos by adopting spectral flux [7, 53] for TAU. Spectral flux measures changes in an audio signal's power spectrum by comparing consecutive frames, revealing sudden variations. These audio shifts often align with key traffic events like braking, collisions, or alarms. By focusing on these audio-based cues, our method accurately selects relevant video frames, ensuring essential anomalies are captured. This alignment between audio and visual cues provides a more efficient and reliable approach to detecting traffic events, as shown in Fig. 3.

Spectral Flux Computation. Given a traffic anomaly video V, we extract its audio time series A at a sampling rate of α kHz, where α is empirically set as 44.1 (sampling rate of compact disc). We then compute the magnitude spectrogram of the audio signal using the Short-Time Fourier Transform (STFT) with a Hamming window $\omega(m)$. Denoting $\hat{A}(n,k)$ as the k-th frequency bin of the n-th point in the audio time series A, the STFT is computed as:

$$\hat{A}(n,k) = \sum_{m=-\frac{S}{2}}^{\frac{S}{2}-1} A(hn+m)\, w(m)\, e^{-\frac{2j\pi mk}{S}}, \quad (1)$$

where the window size S is 2048 and the hop size h is 512.

Spectral flux ($SF(n)$) measures the change in magnitude for each frequency bin over time. By focusing on the positive changes and summing up the changes across all frequency bins, the spectral flux is computed as:

$$SF(n) = \sum_{k=-\frac{S}{2}}^{\frac{S}{2}-1} H\left(|\hat{A}(n,k)| - |\hat{A}(n-1,k)|\right), \quad (2)$$

where $H(x) = \frac{x+|x|}{2}$ is the half-wave rectifier function, ensuring that only positive changes contribute to the sum.

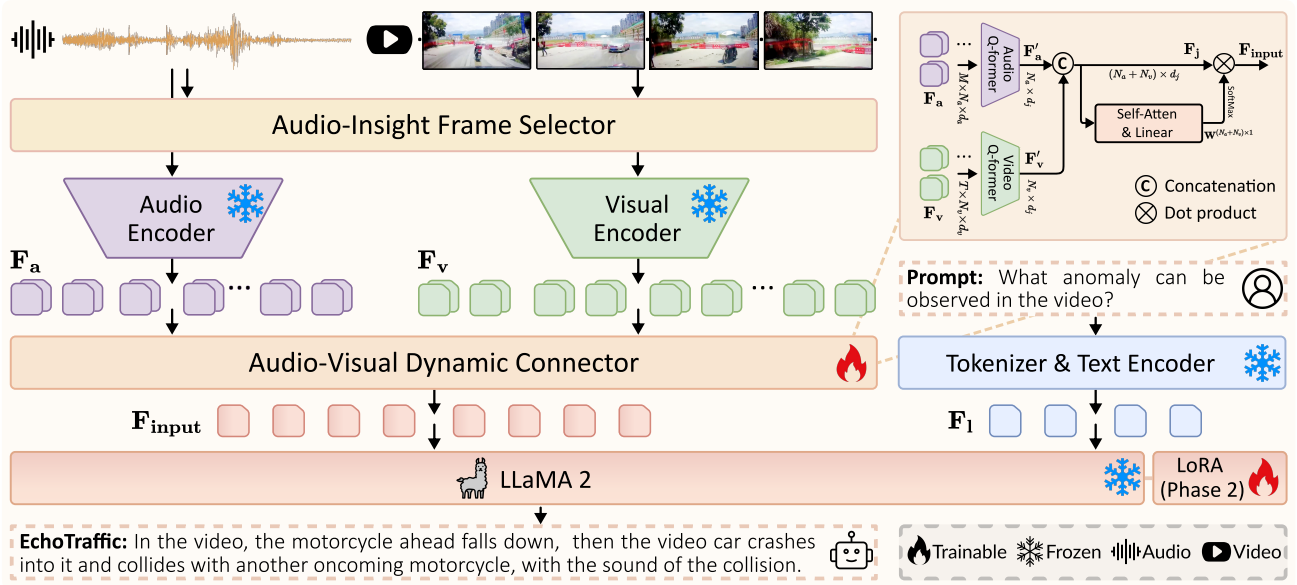

Figure 4. Overview of EchoTraffic. EchoTraffic processes video and audio inputs, using the Audio-Insight Frame Selector to identify key frames from audio cues. These frames and audio segments are encoded by the Visual and Audio Encoders, prioritized by the Audio-Visual Dynamic Connector, and aligned in the large language model embedding space to enhance traffic anomaly understanding.

Frame Selection. To capture key events throughout the audio duration, we divide the Spectral Flux sequence into T equal-length segments. In each segment, we locate the point with the highest Spectral Flux value, marking the most significant spectral change, *i.e.*,

$$n_i^* = \underset{n \in [n_i^{\text{start}}, n_i^{\text{end}}]}{\arg\max} \; SF(n)\,, \tag{3}$$

where $i = 1, 2, \ldots, T$ and n_i^* is a significant spectral change within segment i. These point n_i^* represent moments of potential importance in the traffic scenario, as they infer critical audio anomalies. For each n_i^*, we retrieve the corresponding video frames, identifying key frames in the original traffic anomaly video. The final set of selected frames is represented as $V = \{V^t\}_{t=1}^{T} \in \mathbb{R}^{T \times H \times W \times 3}$, where H and W denote the height and width of each frame.

Note that our frame selector does not solely focus on frames with significant audio signals. It uniformly divides the video into segments and selects frames most likely to exhibit anomalies based on audio cues. This ensures that even segments with weak or no audio signals contribute relevant information to the frame selection process.

4.2. Model Architecture

This section introduces the EchoTraffic architecture, detailing how video and audio data are fused for understanding traffic anomalies. As illustrated in Fig. 4, the visual and audio encoders first process data from each modality separately. The audio-visual dynamic connector then merges these two data streams and aligns them with the LLM's embedding space. Finally, the LLM decoder generates a textual description of the identified traffic anomalies.

Visual Encoder. We employ the ViT-g/14 model from EVA-CLIP [13] as the visual encoder, widely adopted in video understanding models [31, 45, 66]. The selected key frames are fed into the visual encoder and the latent features are processed through a pre-trained image-level Q-former [31], yielding N_v tokens per frame. This visual information is represented as $\mathbf{F_v} = \{\mathbf{F_v^t}\}_{t=1}^{T} \in \mathbb{R}^{T \times N_v \times d_v}$, where d_v denotes the visual feature dimension. Temporal structure is further enhanced by adding learnable position embeddings to the T dimension of $\mathbf{F_v}$.

Audio Encoder. For the audio time series A extracted from the video, we segment it into M parts, each spanning two seconds. Inspired by [17], we convert these segments into spectrograms with 128 mel-spectrogram bins. Treating spectrograms as 2D signals akin to images, we encode them using a ViT, following [16, 18]. After encoding, the audio is represented as $\mathbf{F_a} = \{\mathbf{F_a^m}\}_{m=1}^{M} \in \mathbb{R}^{M \times N_a \times d_a}$, where N_a denotes the number of tokens per audio segment, and d_a is the auditory feature dimension. We then add learnable position embeddings to the M dimensions of $\mathbf{F_a}$ to capture temporal information.

Audio-Visual Dynamic Connector. This module has three main subcomponents: a video Q-former, an audio Q-former, and a self-attention mechanism. The video and audio Q-formers process visual tokens $\mathbf{F_v}$ and audio tokens $\mathbf{F_a}$ to produce fixed-length token sets, *i.e.*, N_v for video and N_a for audio. These tokens are projected into the LLM's embedding space. The video and audio tokens

Table 2. Comparing different approaches on the task of traffic anomaly understanding over the four key subtasks in terms of four evaluation metrics. **Bold** values indicate the best performance, while underlined values indicate the second-best performance. The scores of all metrics range from 0 to 1. Methods incorporating audio signals are marked with $*$. EchoTraffic obtains the best metric scores for most cases.

Task	Metric	Video-LLaVA	LLaMA-ViD	TimeChat	VideoChat2	InternVideo2	VideoLLaMA2*	A-Guardian	Holmes-VAD	EchoTraffic*
Description	BLUE	0.0522	0.0525	0.0167	0.0198	0.0358	0.1066	0.0752	0.0930	**0.2541**
	Rouge	0.3297	0.3518	0.3210	0.4369	0.4012	0.3257	0.2813	0.2436	**0.4380**
	MoverScore	0.5616	0.5637	0.5293	0.5562	0.5616	0.5799	0.5620	0.5614	**0.6205**
	BERTScore	0.8792	0.8818	0.8637	0.8838	0.8843	0.8822	0.8745	0.8696	**0.9102**
	GPT-Eval	0.4109	0.4697	0.2371	0.5418	0.4954	0.6241	0.5464	0.5959	**0.6843**
Causation	BLUE	0.0873	0.0878	0.0421	0.0266	0.0326	0.1037	0.1022	0.0928	**0.2464**
	Rouge	0.2334	0.2402	0.2723	**0.3517**	0.2986	0.2095	0.1349	0.1742	0.3509
	MoverScore	0.5410	0.5421	0.5214	0.5280	0.5230	0.5432	0.5420	0.5381	**0.6077**
	BERTScore	0.8629	0.8641	0.8536	0.8668	0.8589	0.8589	0.8493	0.8548	**0.9017**
	GPT-Eval	0.3500	0.3799	0.1656	0.3243	0.2182	0.6066	0.7389	0.6551	**0.8301**
Prevention	BLUE	0.1009	0.1043	0.0832	0.0700	0.0956	0.1012	0.0855	0.0904	**0.1902**
	Rouge	0.1554	0.1635	0.1020	0.2133	0.1584	0.1343	0.0715	0.1921	**0.4095**
	MoverScore	0.5502	0.5541	0.5268	0.5477	0.5464	0.5457	0.5151	0.5527	**0.6120**
	BERTScore	0.8638	0.8665	0.8425	0.8701	0.8639	0.8585	0.8272	0.8655	**0.8974**
	GPT-Eval	0.7324	0.7125	0.4437	0.6045	0.6683	0.6803	0.2672	0.6053	**0.8462**
Response	BLUE	0.0712	0.0810	0.0431	0.0124	0.0430	0.0870	0.0994	0.0832	**0.3171**
	Rouge	0.2627	0.2410	0.2752	0.3318	0.2864	0.1992	0.1238	0.1934	**0.3994**
	MoverScore	0.5509	0.5511	0.5304	0.5234	0.5340	0.5391	0.5417	0.5354	**0.6310**
	BERTScore	0.8594	0.8579	0.8510	0.8527	0.8532	0.8482	0.8368	0.8463	**0.9071**
	GPT-Eval	0.5233	0.5497	0.3715	0.3982	0.4194	0.5847	0.6746	0.6942	**0.8678**
Overall	BLUE	0.0779	0.0814	0.0463	0.0322	0.0517	0.0996	0.0906	0.0899	**0.2520**
	Rouge	0.2453	0.2491	0.2426	0.3335	0.2861	0.2172	0.1529	0.2008	**0.3994**
	MoverScore	0.5509	0.5528	0.5270	0.5388	0.5412	0.5520	0.5402	0.5469	**0.6178**
	BERTScore	0.8663	0.8676	0.8527	0.8684	0.8651	0.8619	0.8469	0.8591	**0.9041**
	GPT-Eval	0.5091	0.5317	0.3082	0.4672	0.4523	0.6243	0.5540	0.6384	**0.8103**

are then concatenated to form $\mathbf{F_j} \in \mathbb{R}^{(N_v+N_a)\times d_l}$, where d_l is the LLM's hidden dimension. Recognizing that the significance of different video/audio tokens may vary, a self-attention module is designed to assign importance weights for each token. It comprises multiple self-attention layers followed by a linear layer, which takes $\mathbf{F_j}$ as input and generates weights $\mathbf{W} \in \mathbb{R}^{(N_v+N_a)\times 1}$. The final input to the LLM is computed as $\mathbf{F_{input}} = \mathbf{F_j} \times \mathbf{W}$, which effectively prioritizes tokens by the learned importance.

LLM Decoder. We take [52] as the backbone of our LLM decoder, given its strong performance across domains. When language instructions are provided, LLaMA2's tokenizer converts them into language features, represented as $\mathbf{F_l} \in \mathbb{R}^{N_l \times d_l}$, where N_l is the number of tokens in the language instructions. These language features are then concatenated with the audio-visual input $\mathbf{F_{input}}$ to form the full input to the LLM for generating the final textual response.

4.3. Training Strategy

Phase 1: Multi-Modal Alignment with InternVid-10M. We start to train our model on the large-scale InternVid-10M [56] dataset, which contains ten million video and audio samples for producing video descriptions. This substantial dataset size plays a crucial role in effectively aligning the visual and auditory modalities with the language modality, laying a strong foundation for deeper video reasoning tasks. To focus on alignment, we freeze the parameters of the visual encoder, audio encoder, and LLM decoder and train only the audio-visual dynamic connector. Additionally, we initialize the audio encoder with pre-trained weights from ImageBind [16], boosting its capacity to fuse audio and visual inputs. We trained the model for one epoch on the full InternVid-10M dataset with a batch size of 128, using the AdamW optimizer. The learning rate follows a cosine decay schedule with an initial warm-up period, starting at 1e-4. Training in this phase was conducted on eight NVIDIA GTX A6000 GPUs, taking around 240 hours.

Phase 2: Instruction-Tuning for Traffic Anomaly Understanding with AV-TAU. The second phase targets traffic anomaly detection using our AV-TAU dataset. The inputs include video, audio, and a related question, whereas the output is an answer aimed at understanding and responding to traffic anomaly. We train both the audio-visual dynamic connector and the LoRA [21]-based LLM, keeping the visual and audio encoders frozen. This instruction-tuning phase improves the model's accuracy in identifying and interpreting anomalies. In this phase, we trained for one epoch on our own dataset with a batch size of 16, using the AdamW optimizer with a cosine learning rate decay and a warm-up period, starting at a learning rate of 2e-5. The LoRA parameters were set to $r = 64$ and $\alpha = 128$. Training in this phase was conducted on eight NVIDIA GTX A6000 GPUs, taking approximately three hours.

Table 3. Evaluation on the EchoTraffic framework design.

Model	BLEU	Rouge	Mover.	BERTScore	GPT-Eval
w/o Audio Information	0.2243	0.3777	0.5791	0.8594	0.7190
w/o Audio-Insight Frame Selector	0.2395	0.3870	0.5862	0.8730	0.7713
w/o Audio-Visual Dynamic Connector	0.2469	0.3976	0.6092	0.8876	0.7946
EchoTraffic	**0.2520**	**0.3994**	**0.6178**	**0.9041**	**0.8103**

5. Experimental Results

We conduct experiments on the AV-TAU dataset with a 90%-10% train-test split. The test set includes 3,000 videos and 15,000 Q&A pairs, providing a robust basis for evaluating EchoTraffic's performance in the traffic anomaly understanding (TAU) task against comparable models.

5.1. Experimental Setup

Evaluation Metrics. To evaluate the models' performance in understanding traffic anomalies, we adopt five text-level metrics: BLEU [43], ROUGE [36], MoverScore [71], BERTScore [70], and GPT-Eval. The first three metrics are traditional text-based methods that rely on N-gram [43] matching. For GPT-Eval, similar to methods in [38, 50], We prompt GPT-4o [1] with the question and reference answer, allowing it to evaluate the model's output. The GPT-Eval evaluation focuses on three main aspects, *i.e.*, *reasonableness*, *detail*, and *consistency*.

Baselines. As no existing methods are specifically designed for the TAU task, we compare our approach with state-of-the-art models from two related domains: (i) General Video Understanding, including Video-LLaVA [35], LLaMA-VID [33], TimeChat [45], VideoChat2 [32], InternVideo2 [57], and VideoLLaMA2 [6], and (ii) General Video Anomaly Understanding, represented by A-Guardian [8] and Holmes-VAD [67]. This comparison evaluates the ability of these baseline models to effectively interpret and understand video anomalies in traffic scenarios.

5.2. Quantitative Comparison

Table 2 highlights EchoTraffic's superior performance across all evaluation metrics, underscoring its advanced capabilities in traffic anomaly understanding. Key findings include: (i) EchoTraffic consistently achieves the highest scores in text-level BLEU, ROUGE, MoverScore, BERTScore, and GPT-Eval metrics across all categories, establishing its dominance in the Description, Causation, Prevention, and Response subtasks; (ii) VideoLLaMA2 and Holmes-VAD show competitive performance, with VideoLLaMA2 benefiting from integrated audio signals and Holmes-VAD leveraging extensive training on anomaly videos. *Additionally, we fine-tune Video-LLaMA2 under EchoTraffic's settings and evaluate its performance. The results, along with additional comparisons to Gemini 2.0 [51], are available in the supplementary material.*

Table 4. Qualitative evaluation of the EchoTraffic framework. Red highlights semantic inconsistencies, Green shows strong ground truth alignment, and [YELLOW] marks textual issues.

Model	Description
w/o Audio Information	The video shows a person riding a tricycle ahead of the car, which unexpectedly comes to a complete stop on the road, creating a sudden obstruction. [FAILED TO IDENTIFY ANOMALY]
w/o Audio-Insight Frame Selector	This video captures a moment when the car unexpectedly hits a tricycle parked on the road [WRONG]. The sound of the crash is clear, followed by the shocked voice of a man.
w/o Audio-Visual Dynamic Connector	In this footage, the car collides with a tricycle turning around. The impact sound is clear, followed by the noise of items shifting inside the car from the collision.
EchoTraffic	In this video, the car collides with a tricycle approaching from the right while in motion. You can hear the impact of the collision, along with the passengers' startled screams.
Ground Truth	The video shows a car driving straight when it suddenly collides with a three-wheeled vehicle ahead that is preparing to turn left. You can clearly hear the sound of the impact during the collision, along with the startled voices of the passengers and the jarring noise of the car shaking as it brakes.

5.3. Qualitative Comparison

To provide qualitative insights into EchoTraffic's capabilities, we compare it with VideoLLaMA2 and Holmes-VAD. Fig. 5 illustrates a challenging scenario where limited visual cues make detecting a collision difficult, as it occurs in a blind spot without causing camera shake. Nonetheless, EchoTraffic excels at accurately detecting, interpreting, and focusing on traffic anomalies. For instance, in Fig. 5 (a), VideoLLaMA2 and Holmes-VAD fail to identify anomalies from visual cues, resulting in hallucinations, while EchoTraffic accurately infers a collision between the ego vehicle and a white car ahead using audio cues. In Fig. 5 (b) and (d), VideoLLaMA2 and Holmes-VAD do not provide the correct response recommendation or an accurate causal interpretation after the incident. In Fig. 5 (c), Holmes-VAD partially suggests a prevention strategy, while VideoLLaMA2 focuses on irrelevant information. These observations underscore EchoTraffic's effectiveness and capability in leveraging audio-visual insights to accurately perceive and analyze complex traffic anomalies.

5.4. Ablation Study

To assess the contribution of each component in EchoTraffic and the significance of audio information, we conducted an ablation study by systematically removing key elements one by one, as shown in Table 3. When audio information is removed, the model's ability to describe audio-related anomalies is impaired, resulting in incomplete descriptions, which negatively impacts overall performance; see the example in Table 4 "w/o Audio Information." The absence of the Audio-Insight Frame Selector reduces the model's accuracy in describing visual information, as the example shown in Table 4 "w/o Audio-Insight Frame Selector", which emphasizes the role of keyframe selection. Without the Audio-

(a) Description: What anomaly is presented in the video?

VideoLLaMA2: The video shows a car driving down the street in the rain. The camera is inside the car, providing a view of the rainy street and the cars passing by. The unusual incident in the video is that the car hits a pedestrian on the sidewalk. [HALLUCINATION]

Holmes-VAD: Yes, the video shows a car driving down a wet road, passing by other vehicles, and the driver appears to be in control, but suddenly, the car veers off the road and crashes into a tree [HALLUCINATION], which is an abnormal event.

EchoTraffic: In this video, the car in the video collides with a white car that is turning left. You can clearly hear the sound of the video car colliding with the white car, as well as the sound of the woman's screaming.

(b) Causation: Why did the unusual event occur in the video?

VideoLLaMA2: The video shows a car driving down the street in the rain. The sound of the rain hitting the car's roof and windshield is the main source of sound in the video. It appears that the anomaly in the video was caused by a mechanical failure in the car's engine and steering system. [WRONG REASON]

Holmes-VAD: The video shows a car driving down a wet road, passing by other vehicles, and the driver's attention is briefly drawn to the side of the road, where a pedestrian is walking, but the driver continues on their way without stopping. [NOT FOCUS ON CAUSATION]

EchoTraffic: The cause of the accident is that the driver of the video car fails to maintain a safe distance from the white car ahead, and the white car makes a sudden lane changes to the left, which leads to a collision.

(c) Prevention: How could the anomaly have been prevented in the video?

VideoLLaMA2: The video shows a car driving down the street in the rain. The risk of the anomaly can be minimized by ensuring that the car's windshield wipers are turned on and functioning properly. [UNRELATED INFORMATION] This will help to keep the driver's vision clear and reduce the chances of any accidents or collisions.

Holmes-VAD: To minimize the risk of this anomaly, the driver should slow down and maintain a safe distance from the car in front, as the car in front is driving slowly and cautiously, indicating a potential hazard or road condition that requires extra caution.

EchoTraffic: To avoid this accident, the driver of the video car should maintain a safe distance from surrounding vehicles, slow down and ensure sufficient reaction time. The white car should use clear signals, avoid risky overtakes in tight spaces.

(d) Response: How should the event be managed after the anomaly?

VideoLLaMA2: The video does not show any other cars or people on the street, creating a sense of isolation. The car's headlights are on, illuminating the wet road ahead. The video ends with the car driving out of frame, leaving the viewer with a sense of anticipation for what might happen next. [NOT FOCUS ON RESPONSE]

Holmes-VAD: The abnormality in this video is the white van's excessive following distance and the driver's waving at the camera, which may indicate reckless or aggressive driving behavior, and it is important to address this anomaly to ensure the safety of the driver, passengers, and other road users. [NOT FOCUS ON RESPONSE]

EchoTraffic: After the accident, the driver should immediately turn on hazard lights to warn others, check for injuries, call emergency services if needed, and provide first aid. If possible, move the vehicle to a safe area to prevent further traffic obstruction.

Figure 5. Qualitative comparison on the TAU task. Red highlights key semantic inconsistencies, Green indicates strong alignment with the ground truth, and [YELLOW] marks textual issues. Additional examples are provided in the supplementary material.

Visual Dynamic Connector, the model struggles to learn the relative importance of visual and auditory information, potentially missing certain anomalies in the video; see Table 4 "w/o Audio-Visual Dynamic Connector."

6. Conclusion

We develop EchoTraffic, an audio-visual model for Traffic Anomaly Understanding (TAU) that enhances event detection and analysis by incorporating essential audio cues often missed by visual-only methods. EchoTraffic is instruction-tuned on our AV-TAU, the first large-scale dataset designed specifically for audio-visual TAU, using the audio-insight frame selector and the audio-visual dynamic connector to enable detailed traffic anomaly analysis. Evaluation shows its effectiveness in understanding complex incidents and enhancing urban transport safety. We further highlight long video understanding as a key challenge in TAU research.

Acknowledgment

This work was supported by the National Key R&D Program of China (Grant No. 2023YFE0202700), the Hong Kong Innovation and Technology Fund (Grant No. MHP/092/22), the Research Grants Council of Hong Kong (Grant No. 14200824), and Shanghai Artificial Intelligence Laboratory. We thank Zhenyu Ye for valuable discussions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 7

[2] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. VIENA^2: A driving anticipation dataset. In *Asian Conf. Comput. Vis.*, pages 449–466, 2018. 2, 3

[3] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM Int. Conf. Multimedia*, pages 2682–2690, 2020. 2, 3

[4] Yachuang Chai, Jianwu Fang, Haoquan Liang, and Wushouer Silamu. TADS: a novel dataset for road traffic accident detection from a surveillance perspective. *J. Supercomput.*, pages 1–24, 2024. 2, 3

[5] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Asian Conf. Comput. Vis.*, pages 136–153, 2017. 2, 3

[6] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. 7

[7] Simon Dixon. Onset detection revisited. In *Int. Conf. Digit. Audio Effects*, pages 133–137, 2006. 4

[8] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 18793–18803, 2024. 1, 2, 7

[9] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. DADA: Driver attention prediction in driving accident scenarios. *IEEE Trans. Intell. Transp. Syst.*, 23(6): 4959–4971, 2021. 2, 3

[10] Jianwu Fang, Jiahuan Qiao, Jie Bai, Hongkai Yu, and Jianru Xue. Traffic accident detection via self-supervised consistency learning in driving scenarios. *IEEE Trans. Intell. Transp. Syst.*, 23(7):9601–9614, 2022. 1, 2

[11] Jianwu Fang, Jiahuan Qiao, Jianru Xue, and Zhengguo Li. Vision-based traffic accident detection and anticipation: A survey. *IEEE Trans. Circuits Syst. Video Technol.*, 34(4): 1983–1999, 2023. 2

[12] Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 22030–22040, 2024. 2, 3

[13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 19358–19369, 2023. 5

[14] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3

[15] Hadi Ghahremannezhad, Hang Shi, and Chengjun Liu. Real-time accident detection in traffic surveillance using deep learning. In *IEEE Int. Conf. Imaging Syst. Tech.*, pages 1–6, 2022. 3

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 15180–15190, 2023. 5, 6

[17] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. OmniMAE: Single model masked pretraining on images and videos. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10406–10417, 2023. 5

[18] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. *Proc. Interspeech*, 2021. 5

[19] Feten Hajri and Hajer Fradi. Vision transformers for road accident detection from dashboard cameras. In *IEEE Int. Conf. Adv. Video Signal Based Surveill.*, pages 1–8, 2022. 1, 2

[20] Sanjay Haresh, Sateesh Kumar, M Zeeshan Zia, and Quoc-Huy Tran. Towards anomaly detection in dashcam videos. In *IEEE Intell. Veh. Symp.*, pages 1407–1414, 2020. 3

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *Int. Conf. Learn. Represent.*, 2022. 3, 6

[22] Xiaohui Huang, Pan He, Anand Rangarajan, and Sanjay Ranka. Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video. *ACM Trans. Spatial Algorithms Syst.*, 6(2):1–28, 2020. 2

[23] Earnest Paul Ijjina and Sanjay Kumar Sharma. Accident detection from dashboard camera video. In *Int. Conf. Comput. Commun. Netw. Technol.*, pages 1–4, 2019. 3

[24] Caitlienne Diane C. Juan, Jaira Rose A. Bat-og, Kimberly K. Wan, and Macario O. Cordel II. Investigating visual attention-based traffic accident detection model. *Philipp. J. Sci.*, 150(2):515–525, 2021. 3

[25] Muhammad Monjurul Karim, Zhaozheng Yin, and Ruwen Qin. An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos. *IEEE Trans. Intell. Veh.*, 9(1):1792–1803, 2023. 2, 3

[26] Hirokatsu Kataoka, Teppei Suzuki, Shoko Oikawa, Yasuhiro Matsui, and Yutaka Satoh. Drive video analysis for the detection of traffic near-miss incidents. In *Int. Conf. Robot. Autom.*, pages 3421–3428, 2018. 2

[27] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. Crash to not crash: Learn to identify dangerous vehicles using a simulator. In *AAAI Conf. Artif. Intell.*, pages 978–985, 2019. 2, 3

[28] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. Predicting vehicle collisions using data collected from video games. *Mach. Vis. Appl.*, 32(4):93, 2021. 3

[29] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, David C. Anastasiu, Yoichi Sato, and Norimasa Kobori. WTS: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In *Eur. Conf. Comput. Vis.*, pages 1–18, 2024. 2

[30] Trung-Nghia Le, Shintaro Ono, Akihiro Sugimoto, and Hiroshi Kawasaki. Attention R-CNN for accident detection. In *IEEE Intell. Veh. Symp.*, pages 313–320, 2020. 2, 3

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Int. Conf. Mach. Learn.*, pages 19730–19742, 2023. 5

[32] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 22195–22206, 2024. 7

[33] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *Eur. Conf. Comput. Vis.*, 2024. 7

[34] Haicheng Liao, Yongkang Li, Chengyue Wang, Yanchen Guan, KaHou Tam, Chunlin Tian, Li Li, Chengzhong Xu, and Zhenning Li. When, where, and what? a novel benchmark for accident anticipation and localization with large language models. *arXiv preprint arXiv:2407.16277*, 2024. 2

[35] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *Annu. Meet. Assoc. Comput. Linguist.*, 2024. 4, 7

[36] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annu. Meet. Assoc. Comput. Linguist.*, pages 74–81, 2004. 7

[37] Chunsheng Liu, Zijian Li, Faliang Chang, Shuang Li, and Jincan Xie. Temporal shift and spatial attention-based two-stream network for traffic risk assessment. *IEEE Trans. Intell. Transp. Syst.*, 23(8):12518–12530, 2021. 2, 3

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Adv. Neural Inf. Process. Syst.*, pages 34892–34916, 2023. 7

[39] Haohan Luo and Feng Wang. A simulation-based framework for urban traffic accident detection. In *IEEE Int. Conf. Acoust. Speech Signal Process.*, pages 1–5, 2023. 2, 3

[40] Hui Lv and Qianru Sun. Video anomaly detection and explanation via large language models. *arXiv preprint arXiv:2401.05702*, 2024. 3

[41] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE Trans. Image Process.*, 30:4505–4515, 2021. 2, 3

[42] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. DRAMA: Joint risk localization and captioning in driving. In *Winter Conf. Appl. Comput. Vis.*, pages 1043–1052, 2023. 2

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annu. Meet. Assoc. Comput. Linguist.*, pages 311–318, 2002. 7

[44] Karishma Pawar and Vahida Attar. Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, 8(3):379–387, 2022. 1, 2

[45] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 14313–14323, 2024. 5, 7

[46] Debaditya Roy, Tetsuhiro Ishizaka, C. Krishna Mohan, and Atsushi Fukuda. Detection of collision-prone vehicle behavior at intersections using siamese interaction LSTM. *IEEE Trans. Intell. Transp. Syst.*, 23(4):3137–3147, 2020. 1, 2

[47] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, Partha Pratim Roy, and Adway Mitra. Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid cnn-vae architecture. *IEEE Trans. Intell. Transp. Syst.*, 23(8):11891–11902, 2021. 1, 2

[48] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. CADP: A novel dataset for CCTV traffic camera based accident analysis. In *IEEE Int. Conf. Adv. Video Signal Based Surveill.*, pages 1–9, 2018. 2, 3

[49] Leonardo Taccari, Francesco Sambo, Luca Bravi, Samuele Salti, Leonardo Sarti, Matteo Simoncini, and Alessandro Lori. Classification of crash and near-crash events from dashcam videos and telematics. In *Int. Conf. Intell. Transp. Syst.*, pages 2460–2465, 2018. 1, 2

[50] Jiaqi Tang, Hao Lu, Ruizheng Wu, Xiaogang Xu, Ke Ma, Cheng Fang, Bin Guo, Jiangbo Lu, Qifeng Chen, and Ying-Cong Chen. Hawk: Learning to understand open-world video anomalies. *Adv. Neural Inf. Process. Syst.*, 2024. 1, 2, 3, 7

[51] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[53] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process*, 10(5):293–302, 2002. 4

[54] Thakare Kamalakar Vijay, Debi Prosad Dogra, Heeseung Choi, Gipyo Nam, and Ig-Jae Kim. Detection of road accidents using synthetically generated multi-perspective accident videos. *IEEE Trans. Intell. Transp. Syst.*, 24(2):1926–1935, 2022. 1, 2, 3

[55] Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. DeepAccident: A motion and accident prediction benchmark for v2x autonomous driving. In *AAAI Conf. Artif. Intell.*, pages 5599–5606, 2024. 2, 3

[56] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. InternVid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 2, 6

[57] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. InternVideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 7

[58] Li Xu, He Huang, and Jun Liu. SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 9878–9888, 2021. 1, 2, 3

[59] Yajun Xu, Chuwen Huang, Yibing Nan, and Shiguo Lian. TAD: A large-scale benchmark for traffic accidents detection from video surveillance. *arXiv preprint arXiv:2209.12386*, 2022. 2, 3

[60] Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: Reasoning for video anomaly detection with large language models. *Eur. Conf. Comput. Vis.*, 2024. 3

[61] Yu Yao, Mingze Xu, Yuchen Wang, David J. Crandall, and Ella M. Atkins. Unsupervised traffic accident detection in first-person videos. In *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 273–280, 2019. 2, 3

[62] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella M. Atkins, and David J. Crandall. DoTA: Unsupervised detection of traffic anomaly in driving videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):444–459, 2022. 2, 3

[63] Helton Agbewonou Yawovi, Masato Kikuchi, and Tadachika Ozono. Who was wrong? an object detection based responsibility assessment system for crossroad vehicle collisions. *MDPI*, 3(4):844–862, 2022. 3

[64] Tackgeun You and Bohyung Han. Traffic accident benchmark for causality recognition. In *Eur. Conf. Comput. Vis.*, pages 540–556, 2020. 2, 3

[65] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 18527–18536, 2024. 3

[66] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *Annu. Meet. Assoc. Comput. Linguist.*, 2023. 4, 5

[67] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Chuchu Han, Xiaonan Huang, Changxin Gao, Yuehuan Wang, and Nong Sang. Holmes-VAD: Towards unbiased and explainable video anomaly detection via multi-modal LLM. *arXiv preprint arXiv:2406.12235*, 2024. 1, 2, 3, 7

[68] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen. Data-Driven intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.*, 12(4):1624–1639, 2011. 1

[69] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. Exploring event-driven dynamic context for accident scene segmentation. *IEEE Trans. Intell. Transp. Syst.*, 23(3):2606–2622, 2021. 3

[70] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 7

[71] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. *Assoc. Comput. Linguist.*, 2019. 7

[72] Jiaqi Zhu, Shaofeng Cai, Fang Deng, and Junran Wu. Do LLMs understand visual anomalies? uncovering LLM capabilities in zero-shot anomaly detection. *arXiv preprint arXiv:2404.09654*, 2024. 3

[73] Rixing Zhu, Jianwu Fang, Hongke Xu, and Jianru Xue. Progressive temporal-spatial-semantic analysis of driving anomaly detection and recounting. *Sensors*, 19(23):1–16, 2019. 3