IMPACT: Integrated Multimodal Pipeline for Rapid Accident Causality Tracking (Student Abstract)

Vashu Chauhan[1], **Avinash Anand**[1,2], **Manisha Luthra**[3], **Uelison Jean Lopes dos Santos**[3], **Carsten Binnig**[3], **Rajiv Ratn Shah**[1]

[1]IIIT Delhi, India
[2]Nvidia Joint AI Center, Singapore Institute of Technology, Singapore
[3]Technical University of Darmstadt, Germany
vashu22606@iiitd.ac.in, avinasha@iiitd.ac.in, manisha.luthra@dfki.de, uelison.santos@dfki.de, carsten.binnig@cs.tu-darmstadt.de, rajivratn@iiitd.ac.in

Abstract

Traffic accidents pose a significant societal challenge, with many fatalities being avoidable through timely emergency response. We introduce **IMPACT** (Integrated Multimodal Pipeline for Rapid Accident Causality Tracking), a scalable AI framework designed for autonomous, rapid traffic incident analysis using existing urban CCTV infrastructure. IMPACT combines a low-latency CPU-based vision module for real-time key-frame filtering (24 FPS) with the causal reasoning capabilities of MLLMs, reducing costly MLLM calls by over **92%** compared to naive sparse sampling. We further present **TRACE10K**, a dataset featuring three-tier textual annotations that describe accident dynamics at the frame-sequence level.

Code & Dataset: https://github.com/endeavorXx/STREET

Introduction and Related Work

Road traffic injuries cause over 1.19 million deaths annually, and delayed emergency medical services (EMS) are linked to a 46% mortality increase in motor vehicle crashes (WHO 2023). While cities are equipped with vast CCTV networks, current accident analysis systems are often computationally expensive, limited to simple detection, and lack the causal reasoning needed for actionable insights. Powerful Multimodal Large Language Models (MLLMs) (Liu et al. 2023) can provide this reasoning but are too costly for continuous, city-wide deployment.

Prior work has focused on forecasting collisions (Shah, Sinha, and Wang 2018). While many recent methods feed video segments or short clips to an MLLM for analysis (Zhang et al. 2025), this approach is resource-intensive and does not scale to practical, large-scale deployments. The uniform nature of common sparse-sampling methods (Lei et al. 2021) forces a high MLLM invocation rate to avoid missing critical events, resulting in significant yet avoidable computational overhead.

To address these gaps, we propose **IMPACT**, a hybrid framework that balances efficiency and intelligence by extracting the most salient frames through fast pre-processing.

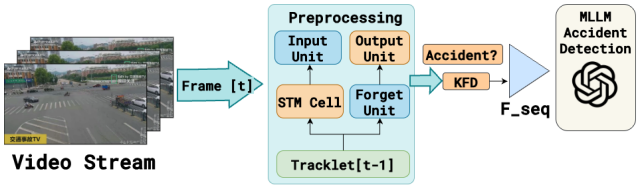

Figure 1: The IMPACT Framework Pipeline: (1) A lightweight KFD algorithm analyzes the video stream in real-time. (2) Upon detecting a potential collision, a critical frame is flagged. (3) The sequence is passed to an MLLM for detailed, tiered causal reasoning.

We also introduce **TRACE10K** dataset of **10,079** frame sequences with 3-tier free-form textual descriptions.

Method

The IMPACT framework operates in a three-stage pipeline, as illustrated in Figure 1 and demonstrated on a real-world example in Figure 2. The stages are: (1) real-time preprocessing and key-frame detection, (2) targeted frame sequence sampling, and (3) multimodal causal reasoning.

Preprocessing and Key-Frame Detection(KFD) The core of our efficiency lies in the preprocessing module, which runs on a standard Core-i3 11th Gen. CPU. It uses classical computer vision techniques, GFTT for corner detection and Lucas-Kanade optical flow for tracking to maintain a dynamic "tracklet" of keypoints on moving objects.

The KFD algorithm analyzes these tracklets to anticipate collisions. For each tracked keypoint, it calculates a **motion vector** based on its recent displacement. A **vicinity circle (vc-circle)** is then defined for each keypoint, using its motion vector as the diameter. A potential collision is flagged if the vc-circles of two keypoints on different motion paths intersect at a significant angle, indicating an imminent impact rather than parallel movement, as visualized in Figure 2. This geometry-based approach is extremely fast, processes video streams in near real-time (∼24 FPS).

Key-Frame Sampling If KFD algorithm flags a critical frame, a short, temporally focused sequence of frames ($\mathcal{F}_{seq}$) is extracted. This sequence includes a few frames before the critical event (to provide context), the critical frame

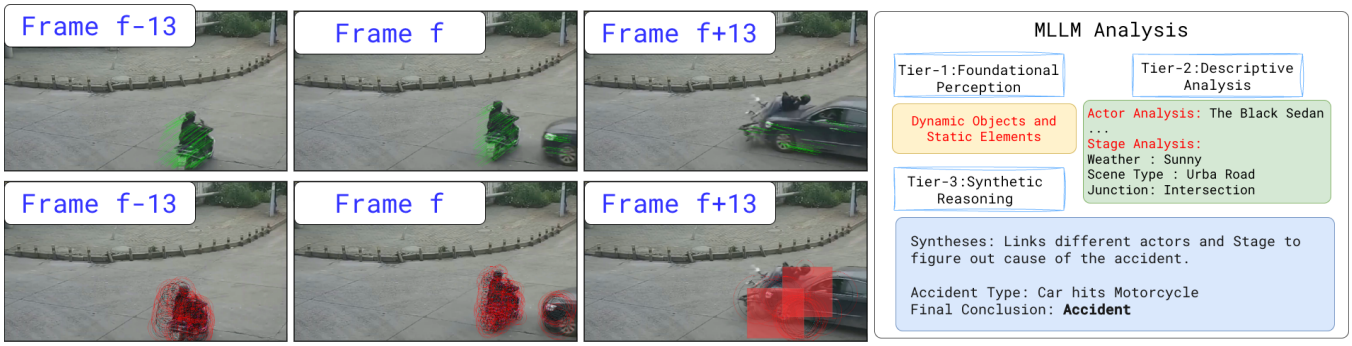

Figure 2: An example of the IMPACT framework processing a traffic incident. The pre-processing module uses the KFD algorithm to track keypoint trajectories (green lines) on the moving sedan and motorbike. As the vehicles converge, their intersecting vicinity circles (vc-circles, shown in red circles) and the resulting highlighted impact area (red box) trigger the detection of a critical event. The flagged $\mathcal{F}_{seq}$ is then passed to the MLLM for hierarchical reasoning. The MLLMs output on the right-hand side is a structured analysis comprising Tier-1 (Perception), Tier-2 (Description), and Tier-3 (Causal Reasoning), providing a comprehensive understanding of the accident.

itself, and a few frames after (to show the immediate outcome). This targeted sampling ensures the MLLM receives a concise, information-dense input optimal for reasoning.

Model	Precision	Recall	F1
Gemini-Flash$^{\dagger}_{Impact}$	**0.823**	0.810	**0.817**
GPT-4o-mini$_{\text{Impact}}$	0.655	0.719	0.678
LLaMA Mav$_{\text{Impact}}$	0.694	0.811	0.730
Phi-4-vision$_{\text{Impact}}$	0.724	0.525	0.523
Video-only Baselines			
LSTM+Attn (Chan et al. 2016)	0.788	0.868	0.818
V-JEPA2 (Assran et al. 2025)	**0.993**	0.964	**0.978**
ViViT (Arnab et al. 2021)	0.917	0.874	0.893
Gemini$_{\text{Video}}$	0.757	**0.886**	0.799

Table 1: Accident recognition performance across IMPACT models using $\mathcal{F}_{seq}$ of size 3, compared with video-only baselines finetuned on 60% and tested on 30% data-splits.

Multimodal Causal Reasoning In the final stage, IMPACT employs multimodal large language models (MLLMs) to generate structured causal explanations. The sampled frame sequence $\mathcal{F}_{seq}$ is provided to the MLLM with a structured prompt that elicits a three-stage reasoning process: (i) **Perception**, where dynamic and static agents are identified; (ii) **Event Description**, where the model produces a temporally grounded account of the scene; and (iii) **Causal Inference**, where the model synthesizes visual and temporal cues to explain how and why the incident unfolded. We hypothesize that a causal chain can be constructed from free-form natural language descriptions (Xiong et al. 2022; Sun, Chao, and Li 2024). This design ensures interpretability and preserves the causal chain needed for downstream emergency decision-making.

Dataset We evaluate IMPACT on the **TRACE10K** benchmark, which extends the TADS dataset (Chai et al. 2024) with **10,079** human-annotated frame-sequence descriptions collected from **966** accident videos captured under diverse environmental conditions. TRACE10K includes three-tier hierarchical causal reasoning labels (Tier-1, Tier-2, Tier-3), frame-level accident boundaries [F_{start}, F_{end}], and metadata covering multiple weather types, junction structures, scene categories, and accident categories (see Supplementary). To ensure balanced evaluation, we additionally include **110** non-accident videos from the IEEE Video Dataset (Adewopo et al. 2023). This diversity enables the study of both accident recognition and structured causal reasoning.

Model	MLLM Calls(%)	F1 Score
LLaMA Mav.$_{\text{Impact}}$	**0.317**	0.73
LLaMA Mav$_{\text{Uniform}}$	4.1	**0.765**

Table 2: F1 score and efficiency performance on LLaMA MAVERICK under IMPACT vs Uniform sampling

Discussion Although several video-based baselines (Table 1) achieve strong performance—particularly in recall—our analysis shows that these models exhibit **poor generalizability** when evaluated across diverse environmental conditions on **OOD Datasets** (see Supplementary). These models rely heavily on spatiotemporal correlations rather than causal cues, leading to overfitting on dataset-specific motion patterns. In contrast, IMPACT's MLLM-driven tiered reasoning provides robustness across conditions while using only **few frames**, enabling structured causal explanations beyond mere classification. This hybrid design strikes a balance between recognition performance and computational efficiency, motivating its use in real-world traffic monitoring systems. This selective filtering further reduces MLLM usage from **4.1%** to just **0.317%** (a 93% reduction) relative to an optimized uniform sampling baseline, while the F1 score drops by only 3 points (Table 2), demonstrating the scalability of IMPACT.

Conclusion and Future Work

IMPACT offers a scalable and economically efficient framework for real-time traffic accident understanding and causal analysis. Future work will explore fine-tuning compact domain-adapted MLLMs and integrating broader multi-sensor inputs. **TRACE10K** provides causal effect annotations at the frame-sequence level for **966 accident videos.**

References

Adewopo, V.; Elsayed, N.; ElSayed, Z.; Ozer, M.; Zekios, C.; Abdelgawad, A.; and Bayoumi, M. 2023. Traffic Accident Detection Video Dataset for AI-Driven Computer Vision Systems in Smart City Transportation.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Assran, M.; Bardes, A.; Fan, D.; Garrido, Q.; Howes, R.; Muckley, M.; Rizvi, A.; Roberts, C.; Sinha, K.; Zholus, A.; et al. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*.

Chai, Y.; Fang, J.; Liang, H.; and Silamu, W. 2024. TADS: a novel dataset for road traffic accident detection from a surveillance perspective. *The Journal of Supercomputing*, 80(18): 26226–26249.

Chan, F.-H.; Chen, Y.-T.; Xiang, Y.; and Sun, M. 2016. Anticipating accidents in dashcam videos. In *Asian conference on computer vision*, 136–153. Springer.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7331–7341.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Shah, S.; Sinha, A.; and Wang, L. 2018. CADP: A Novel Dataset for CCTV Traffic Accident Analysis. *arXiv preprint arXiv:1807.01992*.

Sun, Y.; Chao, Q.; and Li, B. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3493–3511.

WHO. 2023. Road Traffic Injuries.

Xiong, K.; Ding, X.; Li, Z.; Du, L.; Liu, T.; Qin, B.; Zheng, Y.; and Huai, B. 2022. Reco: Reliable causal chain reasoning via structural causal recurrent neural networks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6426–6438.

Zhang, R.; Wang, B.; Zhang, J.; Bian, Z.; Feng, C.; and Ozbay, K. 2025. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. *Accident Analysis & Prevention*, 219: 108077.