Leveraging Video-LLMs for Crash Detection and Narrative Generation: Performance Analysis and Challenges

Ibne Farabi Shihab[a,*], Benjir Islam Alvee[2] and Anuj Sharma[3]

[a] Department of Computer Science, Iowa State University of Science and Technology, Ames, USA; ishihab@iastate.edu

[b] Department of Computer Science, Stony Brook University, Engineering Dr, Stony Brook, USA; benjirislam.alvee@stonybrook.edu

[c] Department of Civil Construction and Environmental Engineering, Iowa State University of Science and Technology, Ames, USA; anujs@iastate.edu,

* Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 29, 2024

Keywords: (Large Language Model, Video-LLM, Video-ChatGPT,Crash detection, Crash narrative)

1 INTRODUCTION

The surge in road fatalities in the United States in 2020, with 38,824 deaths reported by the National Highway Traffic Safety Administration (NHTSA) (Wells & Toffin, 2005), highlights the urgent need for effective accident investigation and analysis. Traffic cameras, crucial for monitoring and recording incidents, are hampered by storage limitations leading to the loss of vital footage (Schrank & Lomax, 2007). This underscores the importance of Automatic Incident Detection (AID) techniques in managing traffic incidents efficiently and reducing costs (Shi & Abdel-Aty, 2015).

Initially, AID algorithms utilized sensor data (Xu *et al.*, 2016) but have evolved to include CCTV cameras for real-time crash detection and offline data analysis to improve road safety (Oliaee *et al.*, 2023). An emerging field is crash narrative generation, which aids in rapid response and safety enhancements (Das *et al.*, 2021). AID techniques are divided into Explicit Event Recognition (supervised learning) and Anomaly Detection (unsupervised learning), with the latter gaining momentum through the development of large language models (LLMs) for robust natural language processing (NLP) (Thirunavukarasu *et al.*, 2023). This progress has paved the way for Image-LLMs (Alayrac *et al.*, 2022) and Video-LLMs (Makridakis *et al.*, 2023), indicating a potential leap towards artificial general intelligence (AGI) in video understanding. Video LLMs, integrating language models with video data interpretation, represent a major advancement in understanding multimedia content. These models employ various innovative techniques to overcome challenges in video analysis, suggesting their significant potential in enhancing traffic management systems through improved crash detection and narrative generation (Ling *et al.*, 2023).

In this study, a novel approach utilizing an off-the-shelf Video-LLM model to detect crash times and generate narratives from videos has been developed. The key contributions include:

1. We developed a crash time detection system for recorded videos using VideoChatGPT (Maaz *et al.*, 2023) with lengths of 2, 20, and 40 minutes, sub-sampled to 20-minute intervals using the video's crash time.

2. Using the VideoChatGPT, corresponding crash narratives have been generated, which, to our knowledge, have not been previously attempted, and we analyzed them based on eleven criteria.

3. These eleven criteria have been further divided into five types of context: Environmental Analysis, Collision-Specific Analysis, Vehicle Dynamics Analysis, Contextual Analysis, and Analytical Analysis, to provide insights into how the model performed for the crash narrative of 2-minute videos.

2 Methodology

2.1 Data Processing, Crash Time Definition, and Model Commands

The crash time is defined as the moment a vehicle is about to collide or lose control. A dataset of 500 crash videos from the Iowa Department of Transportation for 2021-2023 was analyzed, each lasting about an hour. Half of these videos showed actual crashes; a CSV file documented the crash times and details. Videos were trimmed to lengths of 2, 10, 20, and 40 minutes, ensuring each included footage from before and after the crash.

A Python script was created to process these videos, asking the VideoChatGPT model to pinpoint the crash time and describe the crash for up to 250 videos. The outputs were saved in CSV files for crash times and narratives linked to video names for easy reference. The remaining 250 videos were also processed using the same methodology, but no accidents were visible.

3 Results

3.1 Time related work

VideoChatGPT was identified as the most effective for traffic incident detection, outperforming others due to its consistent performance across video lengths from concise 2-minute clips to extensive 40-minute sessions. Other models were too conservative or failed to generate accurate text outputs and crash time detection. VideoChatGPT's superior video analysis capabilities ensured its adaptability and precision in various contexts.

3.2 Video Analysis of Different Durations

Analysis of 2-minute videos has established a baseline for the performance of VideoChatGPT. The results indicate that most errors are within four seconds, demonstrating that the model effectively interprets dynamic events, as seen in Plot 1 of Figure 1. However, when the video duration is extended to 10 minutes, there is an increased variability in the model's predictions, and most errors occur within a 15-second range (Plot 2, Figure 1). This suggests that although the model can handle longer videos, there is a noticeable decline in precision, highlighting areas requiring sustained attention and accuracy improvement.

Further testing with 40-minute videos has revealed even more pronounced discrepancies between predicted and actual events, highlighting a notable decline in accuracy as the length of videos increases, as illustrated in Plot 3, Figure 1. A sub-sampling strategy was implemented to address this issue, wherein 40-minute videos were shortened to 20 minutes based on initial predictions. This process involved posing the question first and then using the preliminary prediction to trim the video accordingly before posing the question again. The improved accuracy observed in these trimmed videos, as depicted in Plot 4, Figure 1, indicates that concentrating on

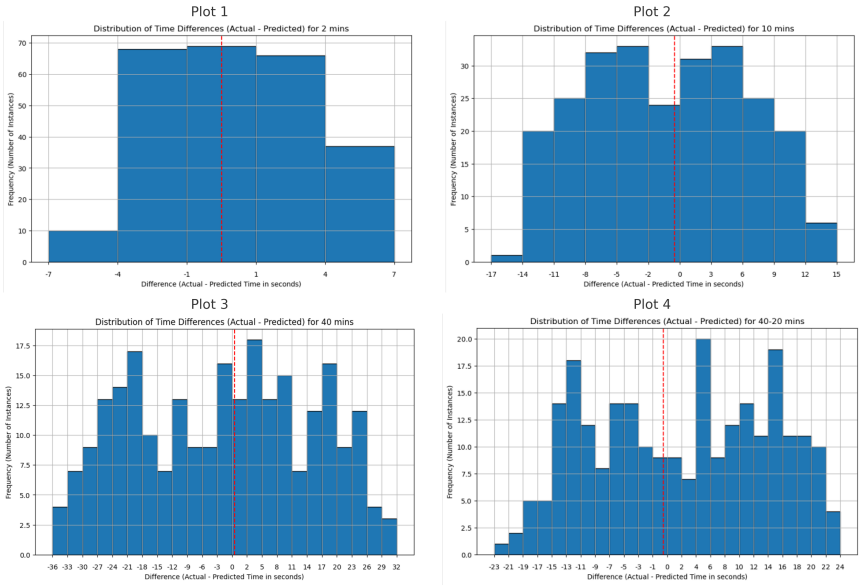

Figure 1 – *Distribution of time differences for 2-minute videos*

key segments can significantly enhance performance. Drawing inspiration from existing research, this approach underscores the model's capability to effectively analyze extensive video content, which is vital for applications such as real-time traffic monitoring.

3.3 Analysis of 2-Minute Generated Crash Report with Annotations

The analysis focused on 11 criteria of crash narratives generated by VideoChatGPT. These criteria covered various aspects, including weather conditions, collision types, driver profiles, road types, vehicle behavior, time of day, causal reasoning, and description accuracy. A manual review of videos was conducted to assess the model's performance for each criterion. The key findings revealed that the model performed well in identifying vehicle dynamics with a high accuracy rate of 73.53% in the "Car Control" criterion. However, it struggled with recognizing specific collision types, such as "Head-On Collision" and "Flip/Rollover," with correct identification rates of only 3.92% and 4.90%, respectively. The analysis also highlighted inconsistencies in "Causal Reasoning" and "Precision" indicating the need for improved accuracy in generating explanations and descriptions. The high percentage of unstated information in "Weather" and "Day Time" suggests potential areas for environmental and temporal factors enhancement. The evaluation categorized the criteria into five areas: environmental analysis, collision-specific analysis, vehicle dynamics analysis, contextual analysis, and analytical analysis, as shown in Table 1. While the model excelled in vehicle dynamics, it showed weaker performance in identifying specific collision types and moderate accuracy in context-related aspects. The findings highlight the model's variable performance across different analytical categories and suggest areas for further training and refinement to improve crash narrative generation capabilities.

4 Discussion

This study on VideoChatGPT Video-LLM has shown that it is highly effective in analyzing short traffic videos for crash detection and narrative generation. However, it struggles when dealing with longer clips and certain types of collisions. The study found that implementing a sub-sampling strategy can improve its performance. Despite these improvements, challenges still exist in causal reasoning and recognizing contextual factors like weather, indicating a need for more comprehensive training and data integration. The research underscores the importance

Table 1 – *Reanalyzed Performance of the Model in Different Categories*

Category	**Criteria**	**Correct Occurrences**	**Incorrect Occurrences**	**Correct Percentage**	**Incorrect Percentage**
Environmental Analysis	Weather, Day Time	70	6	92.11%	7.89%
Collision Specific Analysis	Guardrail Hit, Head-On Collision, Flip/Rollover	43	64	40.19%	59.81%
Vehicle Dynamics Analysis	Car Control, Spinning	88	29	75.21%	24.79%
Contextual Analysis	Driver Description, Road Type	80	39	67.23%	32.77%
Analytical Analysis	Causal Reasoning, Precision	99	84	54.10%	45.90%

of fine-tuning or building from scratch the video-LLMs to continue improving traffic incident analysis with video-LLMs.

References

Alayrac, Jean-Baptiste, Donahue, Jeff, Luc, Pauline, Miech, Antoine, Barr, Iain, Hasson, Yana, Lenc, Millican, *et al.* 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, **35**, 23716–23736.

Das, Subasish, Dutta, Anandi, & Tsapakis, Ioannis. 2021. Topic models from crash narrative reports of motorcycle crash causation study. *Transportation research record*, **2675**(9), 449–462.

Ling, Chen, Zhao, Xujiang, Lu, Jiaying, Deng, Chengyuan, Zheng, Can, Wang, Junxiang, Chowdhury, Tanmoy, Li, Yun, Cui, *et al.* 2023. Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models. *arXiv preprint arXiv:2305.18703*.

Maaz, Muhammad, Rasheed, Hanoona, Khan, Salman, & Khan, Fahad Shahbaz. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*.

Makridakis, Spyros, Petropoulos, Fotios, & Kang, Yanfei. 2023. Large language models: Their success and impact. *Forecasting*, **5**(3), 536–549.

Oliaee, Amir Hossein, Das, Subasish, Liu, Jinli, & Rahman, M Ashifur. 2023. Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types. *Natural Language Processing Journal*, **3**, 100007.

Schrank, D. L., & Lomax, T. J. 2007. *The 2007 urban mobility report.* Tech. rept. Texas Transportation Institute, The Texas AM University System.

Shi, Q., & Abdel-Aty, M. 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, **58**, 380–394.

Thirunavukarasu, Arun James, Ting, Darren Shu Jeng, Elangovan, Kabilan, Gutierrez, Laura, Tan, Ting Fang, & Ting, Daniel Shu Wei. 2023. Large language models in medicine. *Nature medicine*, **29**(8), 1930–1940.

Wells, T., & Toffin, E. 2005. Video-based automatic incident detection on San-Mateo bridge in the San Francisco bay area. *In: 12th World Congress on Intelligent Transportation Systems*. Citeseer.

Xu, C., Liu, P., Yang, B., & Wang, W. 2016. Real-time estimation of secondary crash likelihood on freeways using high-resolution loop detector data. *Transportation Research Part C: Emerging Technologies*, **71**, 406–418.