Holistically Identifying Road Complexity and Relating it to Fatal Crashes

February 2025

Region 1:
New England University Transportation Center

Shannon C. Roberts
University of Massachusetts Amherst

In cooperation with U.S. Department of Transportation,
Office of the Assistant Secretary for Research and Technology (OST-R)

Grant #: 69A3552348301

TECHNICAL DOCUMENTATION

1. Project No. 161133	**2. Government Accession No.** 01905080	**3. Recipient's Catalog No.**
4. Title and Subtitle Holistically Identifying Road Complexity and Relating it to Fatal Crashes		**5. Report Date** February 2025 **6. Performing Organization Code** N/A
7. Author(s) Meng Wang ORCiD - 0000-0002-3304-0610 Shannon Roberts-ORCiD- 0000-0002-0052-7801		**8. Performing Organization Report No.** N/A
9. Performing Organization Name and Address New England University Transportation Center 181 Presidents Drive University of Massachusetts - Amherst Amherst, MA 01003		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** 69A3552348301
12. Sponsoring Agency Name and Address United States Department of Transportation Research and Innovative Technology Administration 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Research Report **14. Sponsoring Agency Code** USDOT OST-R

15. Supplementary Notes

Report uploaded and accessible at the NEUTC Website (www.umass.edu/neutc)

16. Abstract

Understanding the context of crash occurrence in complex driving environments is essential for improving traffic safety and advancing automated driving. Previous studies have used statistical models and deep learning to predict crashes based on semantic, contextual, or vehicle kinematic features, but none have examined the combined influence of these factors. In this study, we term the integration of these features ``roadway complexity". This paper introduces a two-stage framework that integrates roadway complexity features for crash prediction. In the first stage, an encoder extracts hidden contextual information from these features, generating latent complexity features. The second stage uses both original and latent complexity features to predict crash likelihood, achieving an accuracy of 87.98% with original features alone and 90.46% with the added latent complexity features. Ablation studies confirm that a combination of semantic, kinematic, and contextual features yields the best results, which emphasize their role in capturing roadway complexity. Additionally, complexity index annotations generated by the Large Language Model outperform those by Amazon Mechanical Turk, highlighting the potential of AI-based tools for accurate, scalable crash prediction systems.

17. Key Words Roadway complexity, scene perception evaluation, naturalistic driving, large language models (http://trt.trb.org)	**18. Distribution Statement** No restrictions.

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	14	

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized.**

About NEUTC

The New England Regional University Transportation Center (NEUTC) is a diverse, multidisciplinary consortium committed to addressing the pressing issue of traffic safety. Our objective, in line with the Infrastructure Investment and Jobs Act (IIJA), is to drive transformative research, education, and technology transfer to address critical traffic safety needs in a time when roadway fatalities are distressingly high.

Our research and educational activities at NEUTC are guided by four principal safety themes, each addressing a critical challenge in transportation safety. These themes capture the various integral components of the transportation system, focusing on technology, infrastructure, vehicles, and users with a commitment to safety and public engagement. Our overarching theme is promoting safety, with the common underlying science being the study of behavioral, systemic, environmental, and mobility-driven factors on safety.

Disclaimer

All matching funds were University salaries and graduate tuition waivers.

Motivation

Understanding factors contributing to crashes is essential for improving traffic safety and advancing automated vehicle design. A 2015 NHTSA report [1] found that 94% of 5,470 crashes from 2005 to 2007 were due to human errors like inattention, excessive speed, and misjudgment.

Research has sought to predict crash frequency and rates based on environmental factors and driver behavior. Statistical models have linked crash occurrences to variables such as speed, traffic volume, weather, and road design [2], [3], [4], [5], [6], while recent studies have used deep learning to analyze real-time images and sensor data from vehicles and infrastructure to detect hazards [7], [8]. Though promising, these models rely on either imagery or behavioral data, without integrating both to predict crash occurrences or density.

To standardize the imagery of roadway scenes, we define roadway scene complexity as a combination of semantic information (e.g., number of objects) and contextual variation (e.g., road curvature, roadway type, weather) in this paper. Scene complexity affects driver behavior, with factors like object density and road type impacting drivers' situational awareness [9], [10]. Visual and environmental complexity, such as high traffic volumes or urban versus rural settings, also influences cognitive demands on drivers [11], [12], [13], [14].

Driving behavior, such as speed adjustments in response to poor visibility or narrow lanes, is also influenced by scene complexity. Speed and acceleration patterns adjust based on obstacles and conditions [15]. Integrating behavior data with scene information deepens our understanding of driver interactions with their environment, improving crash risk modeling. Thus, we define roadway complexity as the combination of scene complexity and driving behavior, where driving behavior is represented by vehicle kinematic features.

Recent advancements in transformer-based models have further improved performance on scene understanding tasks. Among these, OneFormer [16] has achieved state-of-the-art panoptic quality scores [17] while maintaining computational efficiency, making it suitable for real-time scene understanding in automated driving.

LLMs have been applied to extract contextual information from driving environments by mapping raw sensory data (e.g., images, LiDAR) to higher-level contextual descriptions [18]. The application of LLMs indicates a leap forward as they can go beyond object detection to understand the context of the driving scenarios.

To address the challenges of (1) studying roadway complexity more holistically by incorporating both imagery and behavioral data, and (2) investigating the direct and indirect relationships between roadway complexity and crash density and rates, we propose a two-stage fusion prediction model framework. The framework learns hidden features from the fusion of the semantic representation of the driving scene, vehicle kinematic features, and contextual features. Our model, which is trained on a naturalistic dataset and historical crash data, consists of an encoder that captures the hidden context of the roadway complexity and a prediction module that uses these features to investigate the relationship between the crash density.

Executive Summary

Methods

This study utilizes a subset dataset derived from the MIT-AVT naturalistic driving studies [19]. The dataset includes a variety of multimodal data sources, such as 20-second raw video clips from the forward-facing camera, 30 Hz CAN bus data, 30 Hz GPS data, and contextual metadata features. The contextual features are sourced from a subsequent study [20].

For the purposes of this research, 500 video clips were selected from the dataset developed from Ding et al. [20], categorized as follows: 100 highway scenarios, 100 rural scenarios, 100 urban scenarios, 75 bridge scenarios, 75 overpass scenarios, and 75 crash hotspot scenarios. From each selected video clip, frames were extracted based on a fixed-distance sampling method, with one frame captured for every 20 meters traveled. The extraction began with the first frame of each clip and continued until the distance to the subsequent frame was less than 20 meters. Overall, there were 10,407 frames extracted.

1) Semantic Features: The semantic features were generated using the OneFormer algorithm [16]. The model outputs pixel-level semantic classifications for various objects such as cars, pedestrians, bicycles, roads, traffic signs, sidewalks, buildings, vegetation, and sky. To better understand the effect of complexity on driver behavior, a lead-car region was defined [21].

For each image, the percentage of pixels corresponding to each class relative to the total frame size was calculated. Similarly, the percentage of pixels for each class within the lead-car region was computed. Additionally, the number of cars, pedestrians, buses, bicycles, and motorcycles was counted in both the full frame and the lead-car regions. This process resulted in 50 initial semantic features. After closer examination, features with minimal variability (more than 90% being 0) were removed, reducing the final set to 17 semantic features. The final set of features includes 10 full-frame features and 7 lead car region features: car count, road, vegetation, sky, terrain, car, sidewalk, building, traffic light, person, and lead car-specific features: traffic sign, road, vegetation, sky, car, fence, and car count.

2) Driving Features: The vehicle kinematic features were extracted from the CAN bus data. For each 20-meter segment, the following 9 features were computed: current speed, mean speed, standard deviation of speed, mean longitudinal acceleration, standard deviation of longitudinal acceleration, minimum longitudinal acceleration, maximum longitudinal acceleration, raw deviation from the speed limit, and normalized deviation from the speed limit.

3) Contextual Features: The contextual features include road characteristics, which were generated using the Multimodal LLM (MLLM) GPT model [22]. The prompt used in this step is illustrated in Fig. 1. In the prompt, several questions were asked to gather contextual road characteristics, including information on weather conditions, road conditions, traffic conditions, visibility levels, time of day, road layout, road type, and lane width. Initially, each question was presented in an open-ended format to generate a pool of possible answers. The prompt was then refined based on these responses, providing predefined options to ensure consistency across the answers. During the prompting process, the GPT model was asked to generate a single output option for each question in the prompt. The prompt was run three times on each image to ensure

data quality. For each question, the final answer was determined by selecting the response that was agreed upon by the majority of the three outputs.

```
"""This is the front-camera view image that you, as a driver,
can see. I'm interested in the complexity and demanding level
of the roadway scene for drivers, so you are asked to answer
the following questions. Please give your answers in JSON
format, including the following fields:

From this image, can you tell me how complex and demanding this
environment is for you to navigate and drive on? (1 - 10),

What is the weather like in the image? Please choose one of the
following: clear, cloudy, rainy, snowy, foggy, night.
What is the road condition like in the image? Please choose one
of the following: dry, wet, icy.
What is the traffic condition like in the image? Please choose
one of the following: light, moderate, heavy.
What is the visibility like in the image? Please choose one of
the following: clear, low visibility.
What is the time of day in the image? Please choose one of the
following: day, night, dusk/dawn.
What is the road layout like in the image? Please choose one of
the following: straight, curved, slight curve.
What is the road type like in the image? Please choose one of
the following: highway, city street, rural road, residential
area.
What is the road width like in the image? Please choose one of
the following: narrow, medium, wide.
"""
```

Fig. 1. The prompt used in collecting contextual features with GPT-4o model.

4) Ground Truth: There were two ground truths in this study. As illustrated in Figure 2, for the encoder, the complexity index was collected to quantify the overall complexity level in a given roadway scene. For the prediction, the crash density value was computed and serves as the ground truth for estimating crash rates.

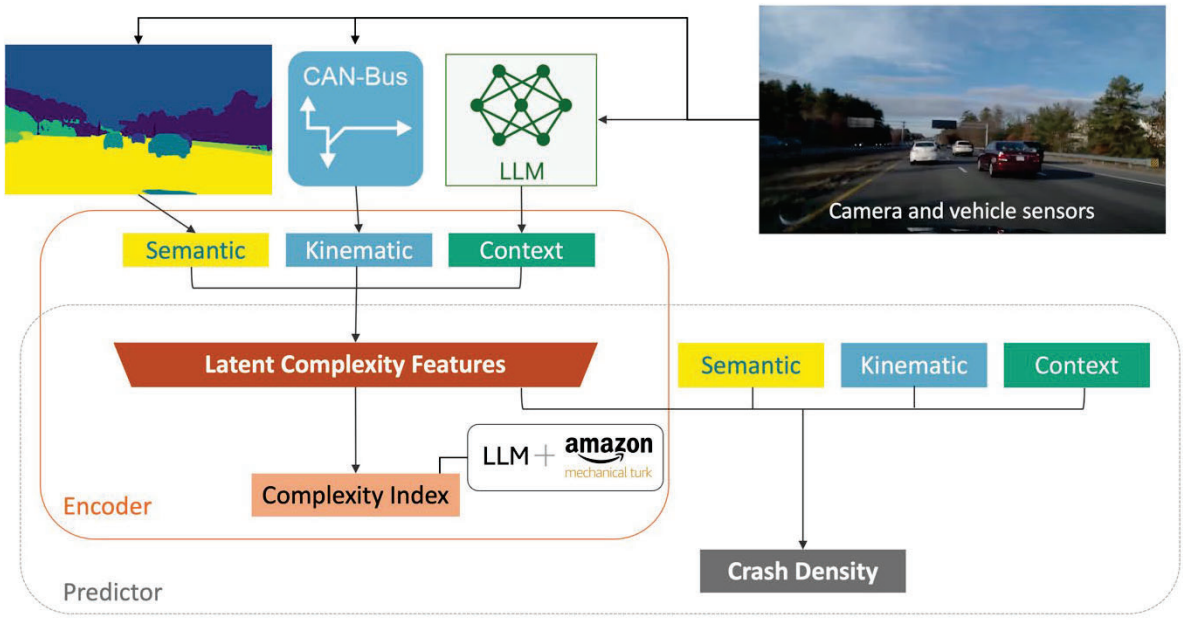

Fig. 2. The model structure. The model takes raw images and CAN-Bus signals as inputs to generate semantic, contextual, and kinematic features, which are then used to investigate their relationship with crash density estimates, serving as the model's output. It consists of an encoder that learns hidden features from the semantic, kinematic, and contextual data, which are infused with the complexity index. The prediction model then utilizes all the available features, including the latent complexity features, to predict the crash density and rates. Example data is shown above each feature source.

1) Complexity Index: The complexity index was generated from two sources: AI and humans. For AI, the GPT model was used along with the contextual feature generation process, as shown in Fig. 1. In this approach, the model generated a complexity score on a scale from 0 to 10 to describe the complexity and demand level of the roadway scenes.

The human-generated complexity indices relied on Amazon Mechanical Turk (MTurk) for annotations. The task was designed to assess the complexity level of roadway scenes. Workers were shown image frames and asked to rate the complexity of each scene on a scale from 1 to 10. Only workers with a high approval rating, at least 500 completed tasks, and residing in the US were selected. Each scene for the study was annotated by three workers, and the final complexity score was determined by averaging their responses.

2) Crash Density: To generate the crash density, GPS data from crashes over a 5-year period (2018 to 2022) was obtained and aggregated from the Massachusetts Department of Transportation's IMPACT app. A Kernel Density Estimation (KDE) method was then applied to this GPS data to create a continuous scale representing crash density. The heatmap of the crash hotspots is shown in Fig. 3 (a).

To keep the scale consistent with the Complexity Index, the crash density was normalized and represented on a scale from 0 to 10. The 10,407 frames were categorized into three levels based on the calculated crash density value. As shown in Fig. 3 (b), the crash density distribution was

skewed and unbalanced. There was a noticeable trend where crash density ranges between 0 and 0.5 are highly frequent, decrease somewhat in the range of 0.5 to 2, and become increasingly rare beyond 2. Therefore, density values between 0 and 0.5 were categorized as Low, those between 0.5 and 2 as Medium, and those between 2 and 10 as High.

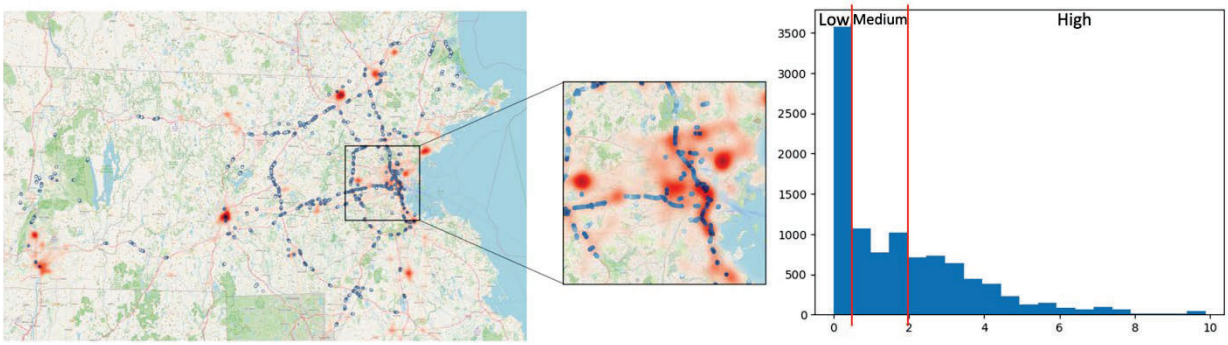

Fig. 3. (a) Crash density heatmap (2018-2022) in Massachusetts, displayed in red, where darker colors indicate a higher crash density. Five hundred video clips from the MIT-AVT dataset are marked in blue. (b) The distribution of crash density value.

1) Complexity Encoder: The complexity encoder used a fully connected neural network structure with either 32 hidden neurons. The input to the network was threefold: (1) the 17 semantic features, (2) the combination of the 17 semantic features and 9 kinematic features, or (3) the combination of all features—17 semantic features, 9 kinematic features, and 19 contextual features. The input variables were normalized to a 0-1 range to ensure consistency across features and improve the stability of the model during training. The output of the network was the complexity index, which was treated as either a continuous or categorical variable for data obtained from the MLLM, and as a continuous variable for data obtained from MTurk. The Root Mean Square Error (RMSE) was used as the evaluation metric for the complexity index when treated as a continuous variable, while accuracy was used as the metric when the complexity index was treated as a categorical variable.

2) Crash Density Prediction Model: After generating the latent complexity features from the encoder, they were used to predict the level of crash density value in combination with the corresponding input feature sets. For example, if the latent complexity features were trained on only semantic features, the input for the crash prediction model would consist of both the latent complexity and semantic features. Similarly, if the latent complexity features were trained on all available features, the input to the crash prediction model would include the latent complexity, semantic, kinematic, and contextual features.

To train the crash prediction model, various algorithms were tested, including Random Forest (RF), Gradient-Boosted Decision Trees (GBDT), K-Nearest Neighbors (KNN), and fully connected neural networks (NN). The NN model used in this step is a fully connected neural network consisting of seven linear layers. Similar to the encoder, the input variables were normalized to a 0-1 range. The dataset was split into 70% for training and 30% for testing. During training, 5-fold cross-validation was used to determine the optimal parameters. The model performance was evaluated using accuracy as the primary metric. To ensure consistency,

the dataset was split in the same way as it was for the encoder. For best-performing tree-based models, the SHAP values [23] were used to analyze the impact of the most influential features associated with crash density.

Experiments

The model performance of the encoder is presented in Table I, showing the model performance on the complexity index obtained from the MLLM. As shown in the table, the three model settings exhibited similar performance across different input feature sets. However, several insights can be drawn from the results: first, the model with 32 hidden neurons and a continuous output produced the lowest RMSE overall. Second, models with more comprehensive input feature sets, which incorporated semantic, kinematic, and contextual features, consistently outperformed those with fewer feature sets. This pattern was observed across all three models.

Table I. Performance comparison of the complexity encoder models with different input feature sets and hidden neurons. The results are reported in terms of RMSE for continuous outputs and cross-entropy for categorical outputs. Downward arrows next to the values indicate that lower values are preferable.

Model Settings	**Input Features**	**RMSE/CE on Train ↓**	**RMSE/CE on Test ↓**
32 neurons, 1-d cont. output	**Semantic**	1.10	1.10
	Semantic + kinematic	1.05	1.06
	Sem. + Kin. + Cont.	0.84	0.86
16 neurons, 1-d cont. output	Semantic	1.11	1.12
	Semantic + kinematic	1.08	1.08
	Sem. + Kin. + Cont.	0.84	0.85
32 neurons, 10-d cat. output	Semantic	1.33	1.35
	Semantic + kinematic	1.31	1.34
	Sem. + Kin. + Cont.	1.15	1.18

Since the encoder with 32 hidden neurons and a continuous output variable yielded the best results, the 32 hidden features from this model were used as the latent complexity features in subsequent contents.

The performance of the crash prediction model is shown in Table II. The baseline model was trained using only the original feature sets, and the latent complexity features were subsequently added to investigate any improvement in model performance. Additionally, for comparison, the performance of the model trained using only the latent complexity features was evaluated, as well as the effect of adding the complexity index to the baseline model.

Table II. Performance comparison of different combinations of feature sets for crash rate prediction. The results are reported in terms of accuracy (%) on the test set. Upward arrows next to the values indicate that higher accuracy is preferable.

Input Features	Model	Baseline ↑	+ Latent Comp. ↑	Difference	Latent Comp. alone ↑	+ Comp. Index ↑
Semantic	RF	**73.23**	**78.35**	5.12	65.31	74.07
	GBDT	68.42	72.49	4.07	60.97	68.30
	KNN	66.09	70.86	4.77	63.85	67.49
	NN	67.81	73.11	5.30	61.07	68.83
Semantic + kinematic	RF	**83.56**	**86.32**	2.76	67.85	84.07
	GBDT	74.62	77.41	2.79	62.24	74.88
	KNN	74.15	75.98	1.83	67.42	73.53
	NN	75.60	77.95	2.35	66.08	72.45
Semantic + kinematic + Contextual	RF	**87.98**	**90.46**	2.17	78.49	87.35
	GBDT	80.13	82.24	2.11	68.03	80.04
	KNN	81.41	82.48	1.07	76.74	81.25
	NN	80.08	88.78	8.70	73.01	84.53

The results indicated that the Random Forest model consistently achieved the best performance across all combinations of input feature sets. Additionally, there was a clear trend of improved model performance as the number of input features increased, with the highest accuracy reaching 87.98%. Adding the latent complexity features led to further improvements in prediction performance, with the highest accuracy being 90.46%, when using all the three feature sets. This was followed by the neural network model, which achieved an accuracy of 88.78%. To assess the significance of these improvements, McNemar's test [24] was conducted, and the results indicated that the improvements were statistically significant. The results suggested that incorporating latent complexity features, particularly in models using a combination of semantic, kinematic, and contextual features, significantly enhances the accuracy of crash rate predictions.

For the best-performing model, which was trained on all available features and latent complexity features, the SHAP values of the 20 most influential features for each class are shown in Fig. 4. The SHAP values highlight the features that most strongly differentiate between each pair of classes. For example, when differentiating the low-density class from the medium-density class, the vegetation area in both the full frame and the lead-car region emerged as the most significant features. Higher values of vegetation area were associated with low crash density areas.

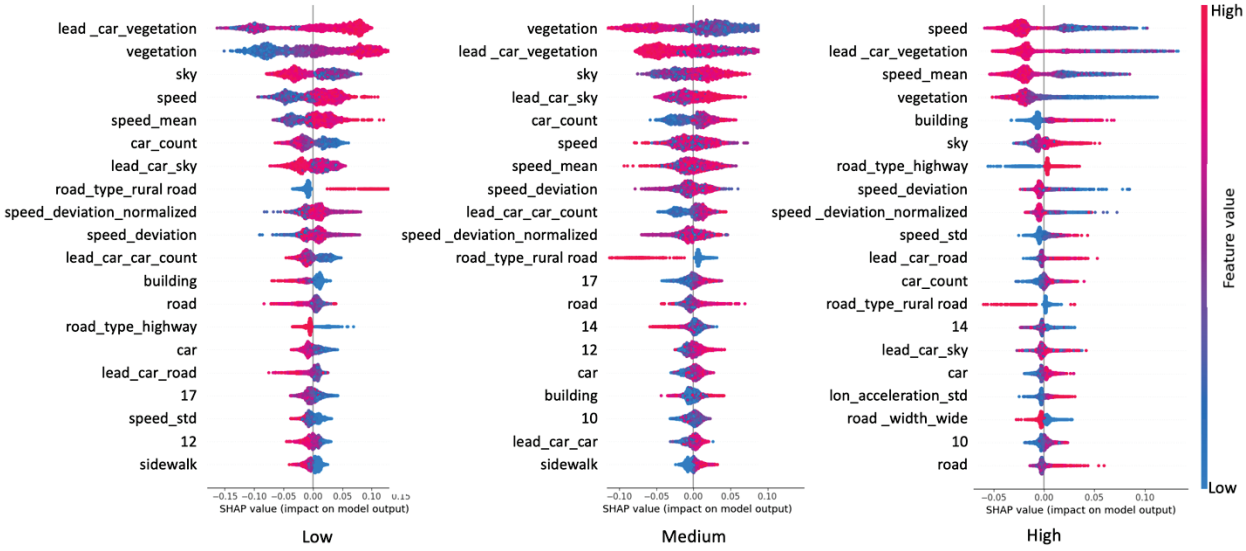

Fig. 4. The SHAP values of the 20 most influential features for each class of the best-performing model. Features represented by numerical codes correspond to latent complexity features.

For the high-density class, speed-related features were the most influential. The results revealed that lower speeds are associated with higher crash-density areas.

Regarding contextual features, rural roads were associated with medium crash-density areas compared to low-density and high-density areas, while highways were more frequently associated with high crash-density areas.

Among the latent complexity features, four (numbers 17, 14, 12, and 10) were identified as influential in distinguishing between different pairwise crash-density areas.

Using the latent complexity features alone or adding the 1-dimensional complexity index to the baseline models did not outperform the baseline models, except for the neural network model on semantic, kinematic, and contextual features. This suggests that simply adding the complexity index does not substantially improve the model's prediction capability. However, the latent complexity features were found to effectively enhance model performance. This indicates that the complexity index is closely associated with semantic representations, vehicle kinematics, and contextual characteristics, and that hidden information can be effectively extracted using neural network structures.

The model using latent complexity features derived from MLLM data consistently yielded higher accuracy than the one using MTurk data. This suggests that the MLLM-generated complexity indexes may capture relevant information more effectively for crash rate prediction.

Conclusions

In this paper, we presented a two-stage framework for extracting hidden context from semantic, kinematic, and contextual features and for predicting crash density by incorporating these hidden context features into the original feature sets. This approach addresses the challenges of understanding the key factors associated with crash prevalence in real-world naturalistic driving environments. To our knowledge, this is the first model to integrate all scene-related and driving-related features and link them to a complexity index obtained from the MLLM. Our experiments revealed that the framework can accurately predict crash density, achieving an accuracy of 90.46%. The data generated by MLLM provided better predictive capability than that obtained from human annotation via MTurk. Different encoder structures consistently outperformed models without latent complexity features.

Outcomes

This solution has the potential to inform the development of advanced driver assistance systems and driver monitoring systems to improve safety in manually driven and automated vehicles. The work can lead to enhanced human-AI collaboration that more effectively supports drivers through complex environmental changes. Furthermore, insights from this framework have the potential to support roadway design by empowering highway engineering departments with the data needed to identify and mitigate risk factors in crash hotspots. If embraced through

infrastructure changes, these results would thus contribute to a key proactive traffic safety intervention that aligns with US DOT's Safe-Systems Approach.

Outcomes *- Any changes made to the transportation system or its regulatory, legislative, or policy framework, resulting from research outputs. Examples include the full-scale adoption of a new technology technique, or practice, or the passing of a new policy, regulation, rulemaking, or legislation.*

Impacts

Beyond vehicle-level safety, insights from this framework can directly contribute to roadway infrastructure improvements. By providing highway engineering departments with data-driven methods to identify and mitigate risk factors in high-crash areas, this research supports proactive safety interventions that align with the U.S. Department of Transportation's Safe System Approach. If incorporated into transportation planning, these findings could help reduce roadway fatalities and injuries while optimizing infrastructure investments to improve overall traffic flow and efficiency.

From an operational and economic perspective, this solution can lead to cost savings by reducing crash-related expenses, including emergency response, vehicle repair, and insurance claims. Additionally, by integrating real-time risk assessment, this approach has the potential to support dynamic traffic management strategies, mitigating congestion and improving road network efficiency.

Impacts *- The impact of research outcomes on the transportation system, or society in general, such as reduced fatalities, decreased capital or operating costs, community impacts, or environmental benefits.*

References

[1] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," U.S. Department of Transportation Technical Report (DOT HS 812 115), Tech. Rep., 2015.

[2] C. Wang, M. A. Quddus, and S. G. Ison, "The effect of traffic and road characteristics on road safety: A review and future research direction," Safety science, vol. 57, pp. 264–275, 2013.

[3] M. H. Rashidi, S. Keshavarz, P. Pazari, N. Safahieh, and A. Samimi, "Modeling the accuracy of traffic crash prediction models," IATSS research, vol. 46, no. 3, pp. 345–352, 2022.

[4] C. Xu, J. Bao, C. Wang, and P. Liu, "Association rule analysis of factors contributing to extraordinarily severe traffic crashes in china," Journal of safety research, vol. 67, pp. 65–75, 2018.

[5] H. M. Hammad, M. Ashraf, F. Abbas, H. F. Bakhat, S. A. Qaisrani, M. Mubeen, S. Fahad, and M. Awais, "Environmental factors affecting the frequency of road traffic accidents: a case study of sub-urban area of pakistan," Environmental Science and Pollution Research, vol. 26, pp. 11 674–11 685, 2019.

[6] C. Xu, W. Wang, and P. Liu, "Identifying crash-prone traffic conditions under different weather on freeways," Journal of safety research, vol. 46, pp. 135–144, 2013.

[7] M. M. Karim, Y. Li, R. Qin, and Z. Yin, “A system of vision sensor based deep neural networks for complex driving scene analysis in support of crash risk assessment and prevention,” arXiv preprint arXiv:2106.10319, 2021.

[8] C. Hu, W. Yang, C. Liu, R. Fang, Z. Guo, and B. Tian, “An image-based crash risk prediction model using visual attention mapping and a deep convolutional neural network,” Journal of Transportation Safety & Security, vol. 15, no. 1, pp. 1–23, 2023.

[9] S. Park, Y. Xing, K. Akash, T. Misu, and L. N. Boyle, “The impact of environmental complexity on drivers’ situation awareness,” in Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2022, pp. 131–138.

[10] P. M. M. P. Asteriou, H. C. Kotsios, and P. Wintersberger, “What characterizes” situations” in situation awareness? findings from a human-centered investigation,” in Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2024, pp. 216–226.

[11] D. Kaber, Y. Zhang, S. Jin, P. Mosaly, and M. Garner, “Effects of hazard exposure and roadway complexity on young and older driver situation awareness and performance,” Transportation research part F: traffic psychology and behaviour, vol. 15, no. 5, pp. 600–611, 2012.

[12] X. Hao, Z. Wang, F. Yang, Y. Wang, Y. Guo, and K. Zhang, “The effect of traffic on situation awareness and mental workload: Simulator-based study,” in Engineering Psychology and Cognitive Ergonomics: 7th International Conference, EPCE 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 7. Springer, 2007, pp. 288–296.

[13] D. Mohan, S. I. Bangdiwala, and A. Villaveces, “Urban street structure and traffic safety,” Journal of safety research, vol. 62, pp. 63–71, 2017.

[14] D. Mukherjee and S. Mitra, “Impact of road infrastructure land use and traffic operational characteristics on pedestrian fatality risk: A case study of kolkata, india,” Transportation in Developing Economies, vol. 5, no. 2, p. 6, 2019.

[15] K. Khan, S. B. Zaidi, and A. Ali, “Evaluating the nature of distractive driving factors towards road traffic accident,” Civil Engineering Journal, vol. 6, no. 8, pp. 1555–1580, 2020.

[16] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, “Oneformer: One transformer to rule universal image segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2989–2998.

[17] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, “Panoptic segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[18] S. Sural, R. Rajkumar et al., “Contextvlm: Zero-shot and few-shot context understanding for autonomous driving using vision language models,” arXiv preprint arXiv:2409.00301, 2024.

[19] L. Fridman, D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, A. Patsekin, J. Kindelsberger, L. Ding et al., “Mit advanced vehicle technology study: Large-scale naturalistic driving study of driver behavior and interaction with automation,” IEEE Access, vol. 7, pp. 102 021–102 038, 2019.

[20] L. Ding, M. Glazer, M. Wang, B. Mehler, B. Reimer, and L. Fridman, “Mit-avt clustered driving scene dataset: Evaluating perception systems in real-world naturalistic driving scenarios,” in 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020, pp. 232–237.

[21] S. Yang, A. McKerral, M. D. Mulhall, M. G. Lenn´e, B. Reimer, and P. Gershon, “Takeover context matters: Characterising context of takeovers in naturalistic driving using super cruise and autopilot,” in Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2023, pp. 112–122.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., “Gpt-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.

[23] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, “From local explanations to global understanding with explainable ai for trees,” Nature Machine Intelligence, vol. 2, no. 1, pp. 2522–5839, 2020.

[24] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.