VISION-RISK: Vision-Language Model for Risk Assessment in Autonomous Driving

1st Andrei-Bogdan Constantin
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
andrei.constantin042@gmail.com

2nd Vlad Negru
Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Vlad.Negru@cs.utcluj.ro

3rd Camelia Lemnaru
Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Camelia.Lemnaru@cs.utcluj.ro

4th Rodica Potolea
Department of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Rodica.Potolea@cs.utcluj.ro

***Abstract*—Autonomous driving involves the challenge of assessing and managing risks in complex environments. One approach for addressing this is the use of human-readable explanations of risk-related tasks to support justification of driving behavior. In this paper, we introduce VISION-RISK, a vision-language model (VLM) designed for risk assessment in autonomous driving using a lightweight architecture, optimized for deployment on edge devices. To train the model, we developed a custom dataset combining real-world driving scenarios from Honda Driving Dataset and extreme high-risk cases from Car Crash Dataset, augmented with synthetic annotations using Dolphins and refined via DeepSeek-V3. VISION-RISK stands out through three key characteristics: the integration of danger level classification with natural language explanation generation, a lightweight architecture optimized for deployment on resource-constrained devices, and a focus on safety through risk assessment to support trust in autonomous driving.**

I. Introduction

The pursuit of achieving full autonomy in driving has fueled innovation processes, combining knowledge from domains like artificial intelligence, robotics and automotive engineering. In essence, the ultimate goal is to develop autonomous driving vehicles, capable of navigating through complex real-world environments while demonstrating a level of understanding and adaptability comparable to that of a human driver.

In recent years, autonomous driving technologies have made significant progress, more specifically, with the help of deep learning models that are capable of interpreting complex visual scenarios and making real-time decisions. Nevertheless, a major challenge persists: the lack of transparency and limited interpretability of the outputs in these decision-making processes as most current autonomous driving systems are data-driven, featuring end-to-end architectures that transform inputs into control output directly [1] [2].

To address these challenges, a growing number of auxiliary systems are being developed, that are not necessarily focused on direct vehicle control, but rather on understanding the behaviors of traffic participants and reasoning about driving scenarios to support safer decision-making. [3] [4] [5] [6] [7] [8]. These solutions serve as auxiliary modules, complementing the core functionalities of autonomous systems through textual explanations and providing an additional layer of transparency.

In this context, a crucial first step toward enhanced safety and building trust in these models lies in the essential need to explicitly assess the danger level of traffic situations and provide a textual explanation [9]. Such a mechanism is intended to provide greater transparency into the system's decisions and to support both internal validation of the model's behavior and external understanding by human users.

In this paper, we propose VISION-RISK, a vision–language model (VLM) for scenario risk assessment in autonomous driving that processes video input and generates both a risk-level label (low, medium, or high) and a corresponding textual explanation. To support this approach, we introduce:

1) VISION-RISK model: a lightweight multimodal architecture capable of processing driving video data and generating both risk assessments and natural language explanations
2) Driving Risk Assessment Dataset (DRAD): built from the Honda Driving Dataset and Car Crash Dataset, containing driving video scenarios, corresponding risk labels and textual explanations.

II. Related Work

Recent vision-language models (VLMs) offer powerful new frameworks for representing and reasoning about driving scenes. For instance, Dolphins [10] is a visual-language model tailored for autonomous driving domain and presented as a conversational driving assistant. It demonstrates strong generalization and few-shot performance on benchmarks, though its computational demands limit suitability for resource-constrained deployment. Think-driver [11] is a vision-language model framework that processes multi-view camera images

979-8-3315-5896-3/25/$31.00 ©2025 IEEE

(front, rear, side views) from a vehicle and uses a large language model to generate driving decisions along with natural language explanations justifying those decisions. It specializes in creating interpretable justifications for driving actions (e.g., braking, accelerating); however, it remains primarily validated in simulation environments (CARLA) and lacks the efficiency for resource-constrained hardware. A notable area within this field addresses the challenge of automated hazard detection. HazardVLM [12] is a vision-language model developed to overcome the problem of hazard recognition and real-time short explanation in autonomous driving scenarios. While HazardVLM emphasizes generating short hazard descriptions, our model complements this with categorical risk quantification and more detailed textual explanations, providing a structured view of driving risk.

Several datasets have been created to support research on explanation and driver behavior in driving scenarios. The Berkeley DeepDrive eXplanation Dataset (BDD-X) [13] is a large-scale multimodal dataset that has also been used in prior work, for example in the fine-tuning of the Dolphins [10], highlighting its role in explanation-oriented driving research. However, it primarily focuses on everyday driving situations and does not include high-risk scenarios, which are specifically represented in our DRAD. The Honda Research Institute Driving Dataset (HDD) [14] is a large-scale dataset designed to facilitate research in driving behavior understanding and causal reasoning. A key innovation of HDD is its 4-layer hierarchical annotation scheme (Goal-oriented action, Stimulus-driven action, Cause, Attention), which we have leveraged in our DRAD. There are datasets that explicitly incorporate danger, hazard, or risk factors into their design. Car Crash Dataset [15] consists of 1,500 trimmed high-risk traffic videos collected from YouTube (50 frames each at 10 fps) and 3,000 normal driving videos randomly sampled from BDD100K; the high-risk subset, Crash1500, is the portion we used in our DRAD. DoTA-HEC (Detection of Traffic Anomalies - Hazardous Event Caption) [12] is a novel dataset introduced and used to train and evaluate the HazardVLM model. The hazard detection annotations, which have a fixed structure, were automatically generated by combining post-processed labels from the original DoTA dataset [16] with heuristic-based techniques. However, these annotations are not very rich or detailed. The Risk Object Identification with Attention (ROI-A) dataset [17] was introduced to study driver-centric risk perception and contains 4,706 short driving video clips collected in the San Francisco Bay Area. The dataset includes annotations for driver intention, road topology, pedestrian attentiveness and driver response. In this work, the authors formulate the task of risk object identification, predicting which objects influence the driver using graph convolutional networks (GCNs). The DRAMA dataset [18] was introduced to study joint risk localization and captioning in driving, containing 17,785 short interactive scenarios collected in Tokyo. This dataset provides bounding boxes, object attributes, and natural language captions describing risks, supporting research on linking object-level localization with explanatory text. However, these approaches do not account for the severity of the risks. The RiskBench benchmark [19] was introduced to provide a systematic, scenario-based evaluation framework for risk identification in autonomous driving, built using the CARLA simulator. It contains 6,916 diverse driving scenarios covering different interaction types, weather, and traffic conditions, and supports tasks such as risk localization, risk anticipation, and planning awareness. Unlike RiskBench dataset, which is simulation-based, our Driving Risk Assessment Dataset is built from real driving videos and explicitly incorporates severity levels for different types of risk. Driving Hazard Prediction and Reasoning (DHPR) [20] is a novel dataset containing dashcam images from BDD100K [21] and EuroCity Persons (ECP) [22]. While it provides around 15,000 annotated samples with bounding boxes, entity descriptions, and natural language hazard explanations, it relies only on images and hypothetical hazards, whereas our DRAD is video-based and incorporates real hazardous events with explicit severity levels. Table I provides a comparative overview of existing risk-related driving datasets, highlighting differences in scope, modality, and annotation, and showing how our DRAD uniquely incorporates explicit severity levels alongside explanatory text.

III. Dataset

This section focuses on describing the method to generate the structured Driving Risk Assessment Dataset (DRAD), that will be used for training VISION-RISK model, containing diverse driving scenarios with annotated risk levels and explanations. The goal is to compile a comprehensive dataset, whose structure is shown in Figure 1, to support robust risk assessment and the generation of detailed explanations of driving scenarios.

video_path	start	end	description
<example_path>	<start_in _ms>	<end_in _ms>	**Danger level: <label>** <explanation>
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

Fig. 1. Dataset Structure

A. Generating the Explanations for Low and Medium Risk Scenarios

Understanding driver behavior in low and medium risk scenarios is critical for developing safer autonomous systems as they are very common in daily driving. We leveraged Honda Driving Dataset (HDD) [14] which provides a rich collection of real-world driving sequences as it contains the 4-layer annotation scheme (described in Section II). This in combination with Dolphins and Deepseek-V3 [23] models, allows us to generate human-interpretable explanations that highlight key factors and driver responses.

TABLE I
COMPARISON OF RISK-RELATED DRIVING DATASETS.

Dataset	#Scenarios	Severity	Explanation	Object Relation	Type	Input Modality	Annotation
ROI-A	4706	✗	✗	Driver–Object	Real	Video, LiDAR, GPS, CAN	Manual + driver response
DoTA-HEC	7901	✗	✓	Short Text	Real	Video	Auto (heuristics)
DHPR	15000	✗	✓	BBox + Text	Real	Image	Manual (crowdsourced)
RiskBench	6916	✗	✗	Scenario-level	Sim	CARLA Simulations	Scripted + augmentation
DRAMA	17785	✗	✓	BBox + Text	Real	Video + text	Manual (bbox + captions)
DRAD (ours)	**32310**	✓	✓	**Text Explanations**	Real	Video	Synthetic + manual

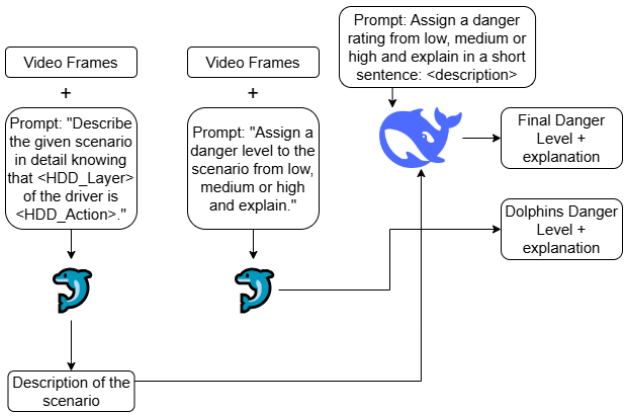

Fig. 2. Strategy for generation of Low/Medium Risk Explanations

TABLE II
DISTRIBUTION OF RISK LEVELS OBTAINED USING TWO ANNOTATION STRATEGIES.

Strategy	Low	Medium	High
Dolphins only	649	30,182	927
Dolphins + DeepSeek	22,510	8,375	873

We employed two different strategies (see Figure 3). In the first approach, we used Dolphins directly to generate both a danger label and an explanation, using the prompt illustrated in Figure 2; however, the resulting distribution was clearly imbalanced for danger classification (Table II), and a manual inspection revealed frequent hallucinations, with cases where the predicted danger label did not match the video scenario, and explanations that were sometimes incorrect or forced to align with the assigned label. In the second approach, we instead used Dolphins to generate contextual scene descriptions, while providing some additional context with the annotations from HDD in the prompt, and then applied DeepSeek to assign a danger level conditioned on those descriptions, using the prompts illustrated in Figure 2. While this strategy still occasionally produced high-risk labels, a manual inspection showed that a considerable amount of these cases corresponded more closely to medium risk. Overall, the distribution of danger levels was more balanced and the explanations were much more accurate compared to the previous method.

B. Generating the Explanations for High-Risk Scenarios

The Car Crash Dataset [15] contains the Crash1500 subset, which focuses on high-risk scenarios. Integrating these events

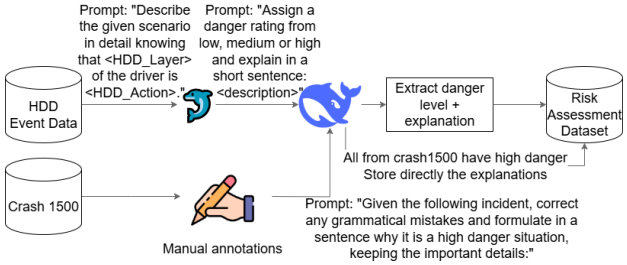

Fig. 3. Pipeline for generating DRAD

into our DRAD is essential for modeling dangerous situations and learning robust representations. A major challenge in the risk assessment domain is the limited availability of labeled data. We initially attempted to use Dolphins for automatic annotation, but it failed to recognize crash and collision scenarios, as its training data did not include such cases. To overcome this, we manually annotated 714 samples from Crash1500 (Figure 3). A subsequent manual inspection revealed minor grammatical errors and redundant expressions that compromised the quality and introduced noise. Therefore, to refine these explanations and obtain a clean, professionally written dataset, we employed DeepSeek-V3 again, using the prompt illustrated in Figure 3 after we manually annotated the scenarios.

TABLE III
DRAD STATISTICS AND COMMON LINGUISTIC PATTERNS.

Category	Statistic / Example
Dataset Size	32,310 valid scenarios
Avg. Lgth.	Scenario: 7.99 s Explanation: 25.37 words
Vocab. Size	2,260 unique words
Top Unigrams	*the* (73,971), *a* (36,898), *and* (25,252), *is* (24,935), *risk* (23,767)
Top Bigrams	*the driver* (11,592), *the risk* (9,380), *risk of* (8,192), *the car* (7,704), *stop sign* (7,187)
Top Trigrams	*the risk of* (7,168), *the driver is* (5,249), *the car is* (5,201), *the stop sign* (4,887)
Top Fivegrams	*the risk of a collision* (2,521), *reducing the risk of a* (1,432), *come to a complete stop* (1,424), *with a green light and* (1,389), *the scenario describes a normal* (1,100), *a green light and no* (1,081), *risk of rear end collisions* (1,075), *to a complete stop as* (1,058)

C. Dataset Overview

Our Driving Risk Assessment Dataset contains 32310 (see Table III) scenarios with a natural distribution as seen in Table IV. As shown in Table III, the videos are relatively

TABLE IV
DANGER LEVEL BREAKDOWN WITH TOTAL SCENARIO COUNTS.

	DRAD	Balanced Subset
Total Scenarios	32,310	11,350
Low	69.46%	52.9%
Medium	25.72%	33.0%
High	4.82%	14.1%

short on average, and the vocabulary is highly domain-specific, reflecting frequent use of traffic- and risk-related terminology. To enable effective training of our VISION-RISK model, we needed a more balanced distribution of danger levels, as high-risk scenarios were initially underrepresented. We therefore selected samples from each danger level to construct a balanced subset of 11,350 scenarios, as shown in Table IV. This dataset was split into training (80%), validation (10%), and test (10%) sets. During training, we further applied oversampling of high-risk scenarios to ensure the model was sufficiently exposed to these critical but less frequent cases. A manual inspection of the dataset shows that the risk factors were captured reliably in low- and high-danger cases, while medium cases often involved a degree of subjectivity, such as scenarios where intersections were assumed to be risky. For more details and concrete examples from each danger category, see Appendix A.

IV. PROPOSED SOLUTION

The proposed multimodal architecture processes video input and generates a textual output with the assessed risk level and explanation. It consists of six core components (see Figure 4):

- Frame Extractor
- Vision encoder (CNN)
- Temporal Encoder (RNN)
- Projection Layer
- Transformer
- Token Decoder

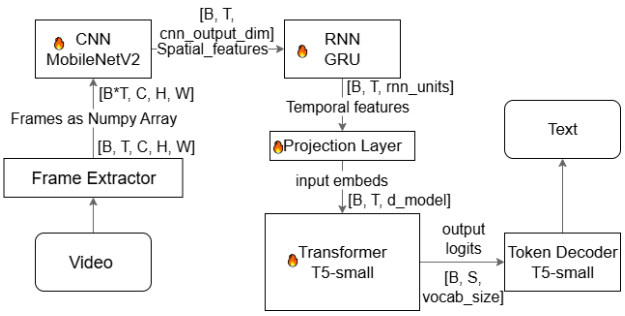

Fig. 4. Model's Architecture

The Frame Extractor component is designed to sample frames from a video file at regular time intervals. The frames are grouped in batches and they get outputted in the following format [B, T, C, H, W], where B is batch size (for multiple videos, when training), T is timesteps (number of frames), C is channels (3 for RGB) and H,W are the height and width of the input frames, which are resized to 224 × 224.

Before entering the CNN, the frames data is flattened to [B*T, C, H, W] as CNNs expects input of shape size 4 (batch, channels, height, width). By flattening the batch and temporal dimensions the model treats each frame independently during CNN feature extraction and leverages GPU parallelism (the CNN processes all frames in one forward pass, avoiding slow per-frame loops). For the CNN backbone, we chose MobileNetV2 [24], as it provides an efficient implementation that is computationally lightweight while maintaining strong feature extraction capability. In our implementation, the MobileNetV2 feature extractor is combined with an adaptive average pooling layer inside a Sequential module, producing feature vectors of size 1280 (cnn_output_dim in Fig. 4), which serve as the input to the temporal encoder.

For the temporal encoder, we chose a GRU (Gated Recurrent Unit) [25] as an implementation of a recurrent neural network (RNN). The GRU had a hidden size of 1024 (rnn_units in Fig. 4), meaning that each time step in the sequence is represented by a 1024-dimensional hidden state vector that captures the temporal information. The GRU was configured with a hidden size of 1024, after preliminary trials with 512, as this setting provides greater representational capacity while maintaining computational efficiency. The GRU processes the CNN features and transforms them into this higher-dimensional representation that captures temporal dynamics.

The projection layer is implemented as a fully connected linear layer, whose main purpose is to map the output of the temporal encoder to the dimensionality expected by the transformer.

We employed T5-small [26], a transformer-based sequence-to-sequence model with approximately 60 million parameters. T5 is built on the standard encoder–decoder transformer architecture and includes mechanisms such as multi-head attention, layer normalization, and residual connections, which help capture long-range dependencies and support stable training. In our setup, the transformer receives features directly as input embeddings, bypassing the tokenizer and vocabulary projection layer, which allows seamless integration of learned video features into the language model. T5-small was chosen as it offers a good trade-off between generation quality and computational efficiency. The hidden dimensionality of the model is d_model = 512 for T5-small (see Fig. 4), corresponding to the size of the internal representations used throughout the transformer.

TABLE V
TRAINING STAGES WITH ACTIVE (✓) AND INACTIVE (✗) COMPONENTS.

Stage	CNN	GRU + Projection	T5 Encoder	T5 Decoder
1	✗	✓	✗	✓
2	✗	✓	✓	✓
3	✓	✓	✓	✓

A. Multi-Stage Training Strategy

Training deep multimodal architectures end-to-end from the beginning can lead to catastrophic results, such as instability,

vanishing gradients or even the model "forgetting" previously learned knowledge—especially when combining multiple pretrained components [27]. To address these challenges, we adopted a multi-stage training strategy, where modules are gradually introduced into the learning process (see Table V). This approach allows the model to first learn stable representations and output formats before fine-tuning the entire architecture in a unified manner. In Stage 1, we trained the projection layer and GRU, which were randomly initialized, together with the T5 decoder to help it adapt early to visual features. In Stage 2, we added the T5 encoder, allowing its pretrained language representations to adjust once the decoder and recurrent layers had stabilized. In Stage 3, we fine-tuned MobileNet, which was already pretrained and only required minor adjustments for the driving domain.

V. Experimental Work

A. Training

In the context of our model, VISION-RISK, which is designed to predict risk levels and generate natural language explanations, it is important to evaluate both outputs: to measure the accuracy and reliability of the model's classifications, and to assess the clarity, relevance, and coherence of the generated explanations. A comprehensive evaluation must incorporate both quantitative and qualitative perspectives to fully capture the model's effectiveness and trustworthiness.

Accuracy, precision, recall, and F1-score were used as the evaluation metrics for the danger level classification task. For the generated explanations, text generation metrics such as BLEU [28], ROUGE [29], and BERTScore [30] were employed to measure lexical and semantic similarity between the model's outputs and the reference texts. The model's output followed the structure **Danger level: <level>** <explanation>, with the danger level evaluated using classification metrics and the explanation using text generation metrics.

The training metrics show that the most significant loss reduction occurred in Stage 1, where the model rapidly converged on the high-risk class, reaching values close to 1 and indicating overfitting. In contrast, medium-risk classification showed little improvement. In Stage 2, medium-risk classification rose slightly, approaching an F1-score of 50%, while overall classification gains remained limited. Text generation metrics, however, improved as the model produced more coherent explanations (Figure 5). By Stage 3, metric gains plateaued and clear signs of overfitting became evident, with little further enhancement in performance (Figure 6). Training was stopped after epoch 6, as no improvements had been observed for five consecutive epochs.

Table VI illustrates the evaluation metrics for our best model (Stage 3 Epoch 6). F1-score for High-risk scenarios is nearly 100%, whereas performance on Medium-risk remains poor. The confusion matrix in Figure 7 shows strong performance on Low-risk cases with some confusion toward Medium, weaker results on Medium where misclassification as Low is frequent, and near-perfect accuracy on High. We think that the poor performance on Medium can be due to the subjectivity in the

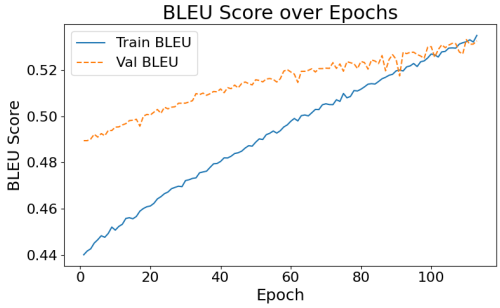

Fig. 5. BLEU Score Stage 2

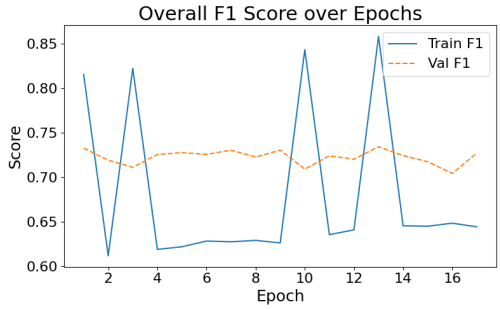

Fig. 6. Overall F1 score Stage 3

potential dangers that may be involved. High-risk scenarios are identified with near-perfect accuracy, which at first glance appears highly desirable, but it also suggests a high degree of overfitting. Several factors may contribute to this effect: targeted oversampling of High-risk samples during training, camera biases across different setups, and geographical variations in the dataset may have led the model to rely on dataset-specific patterns rather than fully generalizable cues. Future work should focus on balancing the classes more effectively and introducing greater diversity across Low- and Medium-risk video scenarios. By doing so, the dataset bias can be reduced, and the model can be encouraged to generate more reliable and generalizable outputs, minimizing overfitting and improving the quality of both the predicted labels and the accompanying explanations. In terms of semantic fidelity, the metrics illustrate a very high BERTScore, while BLEU, ROUGE scores demonstrate that the model produces high-quality text, very similar to the reference text. Overall, the model performs reliably, but further refinement is necessary to enhance its ability to detect medium-risk situations and avoid overfitting on high-risk scenarios.

A manual inspection of the generated outputs confirms that VISION-RISK produces explanations that are domain-specific and contextually appropriate. The wording often differs from the references, but the generated text consistently captures the main risk factors in each scenario, such as lack of reaction time, parked cars, stop sign, or lawful traffic light responses. For detailed examples, consult Appendix B, which provides concrete samples.

TABLE VI
TRAINING AND VALIDATION METRICS FOR BEST MODEL

Metric	Validation	Train
Global Metrics		
Loss	0.1443	0.1355
BLEU	53.2	53.68
ROUGE-1	74.15	75.01
ROUGE-2	53.39	53.92
ROUGE-L	72.98	73.57
BERTScore (Precision)	92.71	92.70
BERTScore (Recall)	93.98	93.96
BERTScore (F1)	93.34	93.32
Danger Level Classification (Overall)		
Accuracy	64.16%	80.78%
Precision	72.96%	84.46%
Recall	72.37%	83.29%
F1-score	72.56%	83.75%
Per-Class Classification Metrics		
Low		
Precision	67.9%	79.61%
Recall	73.32%	85.76%
F1	70.51%	82.58%
Medium		
Precision	50.97%	73.91%
Recall	44.5%	64.95%
F1	47.52%	69.14%
High		
Precision	100%	99.88%
Recall	99.29%	99.18%
F1	99.65%	99.53%

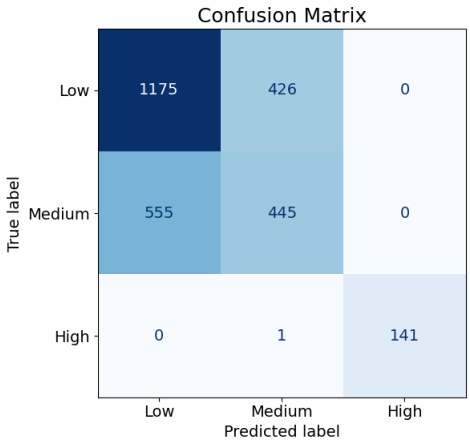

Fig. 7. Confusion matrix of the test set.

B. Comparison with Dolphins

All comparisons reported in Table VII were conducted on the test subset. The prompt used for Dolphins was: “Assign a danger level to the scenario from low, medium, or high and explain”. VISION-RISK shows better results than the Dolphins outputs across the evaluated metrics, which can be explained by the fact that VISION-RISK was trained specifically for the risk assessment task, whereas Dolphins was not. During manual inspection of Dolphins’ generated explanations, we identified instances of semantic hallucinations where the model’s predicted risk level contradicted the textual description and video scenario (issue mentioned in Subsection III-A). A recurring pattern emerged where Dolphins would accurately describe the environment (e.g., "the car makes a right turn at an intersection") but draw incorrect risk conclusions (labeling it as high risk despite the intersection being empty).

TABLE VII
COMPARISON BETWEEN DOLPHINS AND VISION-RISK ON THE TEST SET.

Metric	Dolphins	VISION-RISK
Global Metrics		
BLEU	1.4	53.14
ROUGE-1	23.56	74.01
ROUGE-2	3.23	53.18
ROUGE-L	17.38	72.81
BERTScore (Precision)	88.06	92.75
BERTScore (Recall)	86.39	94.00
BERTScore (F1)	87.21	93.37
Danger Level Classification (Overall)		
Accuracy	39.65%	63.48%
Precision	36.13%	72.53%
Recall	43.00%	71.85%
F1-score	35.87%	72.13%
Per-Class Classification Metrics		
Low		
Precision (Low)	59.50%	67.72%
Recall (Low)	38.70%	71.78%
F1 (Low)	46.89%	69.69%
Medium		
Precision (Medium)	32.09%	49.88%
Recall (Medium)	39.6%	45.2%
F1 (Medium)	35.45%	47.42%
High		
Precision (High)	16.82%	100%
Recall (High)	50.70%	98.59%
F1 (High)	25.26%	99.29%

C. Deployment and benchmarking

The VISION-RISK model was deployed on a Raspberry Pi 5 along with a simple Web UI to evaluate inference performance in a low-power edge computing environment. The model has a size of 280MB, with 70,322,432 parameters and a computational cost of 3,462,764,544 FLOPs for an input of 8 frames, each with a spatial resolution of 224x224 and 3 RGB channels. We created a benchmark containing 100 videos from our DRAD, with an average video length of 6.64 seconds and an average of 8.93 sampled frames per clip. The first 10 videos were used for warmup, and the following 90 videos were used for computing the reported metrics. As shown in Table VIII, where the RTX 3070 laptop results serve as a baseline, the Raspberry Pi 5 achieved an average computing latency of 6.054 seconds per video and a throughput of 0.159 videos/s, compared to 0.605 seconds per video and 1.335 videos/s on the RTX 3070. Peak RAM usage on the Raspberry Pi reached 19.1% of the 8GB available (aprox. 1.53GB). Overall, the results indicate that the model is lightweight and can run in constrained environments, with expected differences in execution time and resource utilization.

We evaluated the model in several zero-shot generalization scenarios using driving videos from the Internet. The model consistently identified high-risk situations, though some explanations were inaccurate (e.g., mislabeling an overtaking

TABLE VIII
PERFORMANCE COMPARISON BETWEEN LAPTOP (RTX 3070) AND RASPBERRY PI 5 (CPU-ONLY).

Metric	Laptop (RTX 3070)	Raspberry Pi 5
Throughput (videos/s)	1.335	0.159
Computing latency mean (s)	0.605	6.054

maneuver as a U-turn or ambiguous references to the responsible vehicle). Still, it captured key risk factors such as failure to brake, right-of-way violations, delayed reactions, and lack of situational awareness. For low- and medium-risk cases, performance was less consistent across varied environments, and these cases were often mistaken for high-risk, which can be attributed to the overfitting issues discussed in Subsection V-A.

D. Implementation Details

The model was implemented in Python using PyTorch framework. It was trained using three NVIDIA L40S GPUs, with the complete training process spanning approximately 2.5 days. An early stopping criterion was employed to improve training efficiency and keep overfitting as low as possible: training was halted if the validation loss did not improve for five consecutive epochs. In total, we went through eight iterations of training, and the learning rates were determined through manual adjustment over multiple iterations of the training; this approach is not optimal and could be improved in the future by adopting more systematic hyperparameter optimization methods. The learning rates used were as follows:

- CNN: 1×10^{-5}
- GRU: 1×10^{-3}
- Projection Layer: 1×10^{-3}
- T5: 2×10^{-5}

In our case, we used AdamW [31] as optimizer, that decouples weight decay from the gradient update, which improves performance for transformer-based models like T5. For our experiments, the weight decay parameter was set to 1×10^{-4} for each component. Training was performed with a batch size of 32, using a maximum of 16 frames per video sampled at 0.66-second intervals. The loss function was Cross-Entropy, and the model was trained using teacher-forced sequence training. Mixed precision was enabled with GradScaler for efficient AMP training, and Distributed Data Parallel (DDP) was employed to utilize multiple GPUs. The maximum sequence length for the text output was fixed at 256 tokens.

VI. CONCLUSION AND FUTURE DIRECTIONS

In the pursuit of improved risk awareness in autonomous driving, we proposed VISION-RISK, a VLM that assesses driving risk while generating human-readable textual explanations. By addressing both risk severity assessment and the generation of detailed explanations, VISION-RISK contributes to the exploration of approaches that aim to improve risk-awareness in autonomous driving. To enable this, our Driving Risk Assessment Dataset was created by annotating real driving data with synthetic labels and supplementing it with manually annotated examples, allowing the model to learn both risk severity assessment and explanatory reasoning. To explore its applicability in real-world edge scenarios, VISION-RISK was deployed on a Raspberry Pi 5, where it achieved reasonable performance even without hardware acceleration or optimizations, suggesting potential for lightweight, low-cost deployment in constrained environments.

These findings lay the groundwork for future research aimed at enhancing both the efficiency and capability of VISION-RISK. Promising directions include the application of model optimization techniques—such as quantization, pruning, or ONNX conversion—to significantly reduce inference time, alongside the integration of hardware accelerators to support real-time deployment. Moreover, expanding the Driving Risk Assessment Dataset with increasingly complex and diverse driving scenarios, as well as incorporating additional sensory modalities such as LiDAR and CAN parameters (e.g., speed, steering angle), holds the potential to further strengthen the model's robustness and explanatory depth.

ACKNOWLEDGMENT

This work is supported by the grant number PN-IV-P7-7.1-PED-2024-0952 from "Proiect experimental demonstrativ (PED) 2024".

REFERENCES

[1] D. Coelho and M. Oliveira, "A review of end-to-end autonomous driving in urban environments," *IEEE Access*, vol. 10, pp. 75296–75311, 2022.

[2] K. Charoenpitaks, Q. Nguyen Van, M. Suganuma, M. Takahashi, R. Niihara, and T. Okatani, "Exploring the potential of multi-modal ai for driving hazard prediction," *IEEE Transactions on Intelligent Vehicles*, vol. PP, pp. 1–11, 01 2024.

[3] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, " Drive Like a Human: Rethinking Autonomous Driving with Large Language Models ," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, (Los Alamitos, CA, USA), pp. 910–919, IEEE Computer Society, Jan. 2024.

[4] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," 2023.

[5] Y. Jin, R. Yang, Z. Yi, X. Shen, H. Peng, X. Liu, J. Qin, J. Li, J. Xie, P. Gao, G. Zhou, and J. Gong, "Surrealdriver: Designing llm-powered generative driver agent framework based on human drivers' driving-thinking data," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 966–971, 2024.

[6] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving withnbsp;graph visual question answering," in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LII*, (Berlin, Heidelberg), p. 256–274, Springer-Verlag, 2024.

[7] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8186–8193, 2024.

[8] Wayve, "Lingo-1: Exploring natural language for autonomous driving," 2025. [Online]. Available: https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/.

[9] L. Petersen, H. Zhao, D. Tilbury, X. J. Yang, and L. Robert, "The influence of risk on driver trust in autonomous driving systems," 11 2024.

[10] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, *Dolphins: Multimodal Language Model for Driving*, pp. 403–420. 11 2024.

[11] Q. Zhang, M. Zhu, and H. Yang, "Think-driver: From driving-scene understanding to decision-making with vision language models," in *Proc. Autonomous Vehicles Meet Multimodal Foundation Models Workshop (ECCV)*, 2024.

[12] D. Xiao, M. Dianati, P. Jennings, and R. Woodman, "Hazardvlm: A video language model for real-time hazard description in automated driving systems," *IEEE Transactions on Intelligent Vehicles*, pp. 1–13, 2024.

[13] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[14] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, " Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning ," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 7699–7707, IEEE Computer Society, June 2018.

[15] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *ACM Multimedia Conference*, May 2020.

[16] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins, and D. Crandall, "Dota: unsupervised detection of traffic anomaly in driving videos," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[17] N. Agarwal and Y.-T. Chen, "Risk perception in driving scenes," in *Proc. Machine Learning for Autonomous Driving Workshop (NeurIPS)*, 2022.

[18] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1043–1052, 2023.

[19] C.-H. Kung, C.-C. Yang, P.-Y. Pao, S.-W. Lu, P.-L. Chen, H.-C. Lu, and Y.-T. Chen, "Riskbench: A scenario-based benchmark for risk identification," 05 2024.

[20] K. Charoenpitaks, V.-Q. Nguyen, M. Suganuma, M. Takahashi, R. Niihara, and T. Okatani, "Exploring the potential of multi-modal ai for driving hazard prediction," 2024.

[21] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2633–2642, 2020.

[22] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, p. 1844–1861, Aug. 2019.

[23] DeepSeek, "Deepseek-v3: Advanced language model," 2025. [Online]. Available: https://deepseek.com.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, " MobileNetV2: Inverted Residuals and Linear Bottlenecks ," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 4510–4520, IEEE Computer Society, June 2018.

[25] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, Jan. 2020.

[27] D. Zhu, Z. Sun, Z. Li, T. Shen, K. Yan, S. Ding, C. Wu, and K. Kuang, "Model tailor: mitigating catastrophic forgetting in multi-modal large language models," in *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, JMLR.org, 2024.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[29] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004.

[30] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

APPENDIX A
MANUALLY ANALYZED SAMPLES FROM DRAD

Table IX presents representative DRAD samples that were manually analyzed to evaluate the quality of the constructed dataset. In the low- and high-danger cases, we observed that the risk factors were captured very well (e.g., "inattentive driver failed to notice pedestrians crossing," "hard brake," "stop sign," "making a safe right turn when clear"), while in medium cases there is some subjectivity. For example, the sample where intersections were assumed to be dangerous, leading to a medium label.

TABLE IX
EXAMPLES OF MANUALLY ANALYZED SAMPLES FROM DRAD.

Video	Start (ms)	End (ms)	Danger Level	Explanation	Evaluation
000426.mp4	0	5000	High	This is a high-danger situation because an inattentive driver failed to notice pedestrians crossing, leading to a hard brake and a collision, posing serious risk to vulnerable road users despite the ego driver's non-involvement.	Risk factors recognized, no ambiguity
2017-06-14-18-19-51_new_0.75.mp4	2603852	2610321	Low	The driver is following traffic rules, stopping at the stop sign, checking for hazards, and making a safe right turn when clear.	Risk factors recognized, no ambiguity
2017-09-21-15-47-21_new_0.75.mp4	2230096	2235422	Medium	The stop-and-go traffic and high vehicle volume increase the risk of rear-end collisions and frustration-driven maneuvers, but the low speeds reduce the severity potential.	Risk factors recognized, no ambiguity
2017-03-01-10-16-57_new_0.75.mp4	1036035	1038495	Medium	While the green light indicates it is safe to proceed, intersections remain high-risk areas due to potential red-light runners, pedestrians, or unexpected actions by other drivers.	Assumed that intersections are high risk which can be subjective, a bit ambiguous

APPENDIX B
EXAMPLES OF VISION-RISK OUTPUTS FROM THE TEST SET

Table X presents examples from the test set comparing the reference annotations with the outputs generated by VISION-RISK. In high-risk scenarios, it identifies elements such as failure to maintain distance, lack of reaction time, or stopped traffic. For medium-risk cases, it refers to conditions like parked vehicles, stop signs, or wet roads that introduce potential hazards but can be mitigated through cautious driving. In low-risk examples, it emphasizes lawful responses to green lights and normal traffic flow. While the wording often differs from the references (e.g., "indicating a lack of sufficient reaction time" vs. "ego driver's failure to notice and react to stopped traffic"), the generated explanations remain aligned with the intended risk assessment.

TABLE X
EXAMPLES FROM THE TEST SET WITH REFERENCE AND VISION-RISK GENERATED OUTPUTS.

Video	Start (ms)	End (ms)	Reference Output	VISION-RISK Output
000628.mp4	0	5000	****Danger level: High**** This is a high-danger situation because the ego driver's failure to maintain a safe following distance led to a rear-end collision with the vehicle ahead, which had braked, indicating a lack of sufficient reaction time and increased risk of severe impact.	****Danger level: High**** This is a high-danger situation because the ego driver's failure to notice and react to stopped traffic led to a rear-end collision, risking severe injuries or multi-vehicle damage due to the unexpected maneuver.
2017-04-11-10-41-38_new_0.75.mp4	1689400	1692000	****Danger level: Medium**** The scenario involves a common maneuver, but the parked vehicle limits visibility, requiring extra caution to check for oncoming traffic and pedestrians.	****Danger level: Medium**** The scenario involves a potential risk if the driver fails to stop at the intersection, but the presence of a stop sign and parked cars suggests a controlled environment where proper attention and speed adjustment can mitigate the danger.
2017-06-08-14-45-09_new_0.75.mp4	2446687	2448906	****Danger level: Medium**** Wet roads reduce traction, increasing the risk of skidding or hydroplaning, and intersections add complexity due to potential cross-traffic or sudden stops. However, if the driver is maintaining appropriate speed and caution, the immediate danger is moderated.	****Danger level: Medium**** The wet road reduces traction, increasing stopping distance, and the parked cars reduce visibility and increase stopping distance, increasing the risk of sliding or delayed reaction times.
2017-06-13-11-27-56_new_0.75.mp4	3485080	3495960	****Danger level: Low**** The car is crossing the intersection legally with a green light, and other vehicles are also proceeding, indicating normal traffic flow with no immediate hazards.	****Danger level: Low**** The scenario describes a normal and lawful response to a traffic light change, with the driver proceeding through a crosswalk only when the light turns green, indicating it is safe to do so.