2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

NuRisk: A Visual Question Answering Dataset for Agent-Level Risk Assessment in Autonomous Driving

Yuan Gao, Mattia Piccinini, Roberto Brusnicki, Yuchen Zhang, Johannes Betz

***Abstract*— Understanding risk in autonomous driving requires not only perception and prediction, but also high-level reasoning about agent behavior and context. Current Vision Language Model (VLM)-based methods primarily ground agents in static images and provide qualitative judgments, lacking the spatio–temporal reasoning needed to capture how risks evolve over time. To address this gap, we propose NuRisk, a comprehensive Visual Question Answering (VQA) dataset comprising 2.9K scenarios and 1.1M agent-level samples, built on real-world data from nuScenes and Waymo, completed with safety-critical scenarios from the CommonRoad simulator. The dataset provides Bird's-eye view (BEV) based sequential images with quantitative, agent-level risk annotations, enabling spatio–temporal reasoning. We benchmark well-known VLMs across different prompting techniques and find that they fail to perform explicit spatio-temporal reasoning, resulting in a peak accuracy of 33% at high latency. To address these shortcomings, our fine-tuned 7B VLM agent improves accuracy to 41% and reduces latency by 75%, demonstrating explicit spatio-temporal reasoning capabilities that proprietary models lacked. While this represents a significant step forward, the modest accuracy underscores the profound challenge of the task, establishing NuRisk as a critical benchmark for advancing spatio-temporal reasoning in autonomous driving. More information can be found at `https://github.com/TUM-AVS/NuRisk`.**

I. Introduction

Autonomous driving has progressed rapidly, with milestones like Waymo's fully autonomous robotaxi services demonstrating SAE Level 4 capabilities in defined urban environments [1], [2]. These achievements are built on either modular software stacks covering perception, prediction, planning, and control [3], or more recently, end-to-end learning-based approaches [4]. However, both paradigms face a fundamental limitation: they struggle to handle the variability of real-world driving, particularly the rare, safety-critical corner cases that fall outside their operational design domains (ODDs) [5]. Ensuring trustworthiness requires methods that can reason beyond hand-crafted rules or training data distributions. Recent advances in Vision Language Model (VLM) offer such potential: by combining scalable, cross-modal knowledge with flexible reasoning, they can improve the coverage of autonomous driving tasks, complementing traditional AD software stacks [6].

Despite this potential, the application of VLMs to autonomous driving has been focused on scenario risk assessment, such as hazard detection [7] or qualitative risk assessment [8]. While these models can identify a potential hazard, they typically lack the quantitative reasoning required for rigorous safety evaluation. For example, a VLM can identify a vehicle cutting into the ego-car's lane yet offer a qualitative assessment like, "This is a dangerous situation, please drive slowly", which encourages conservative behavior. This raises a question: can pre-trained VLMs perform agent-level quantitative risk assessment by leveraging spatio-temporal reasoning to interpret safety metrics, such as the temporal metric Time-to-Collision (TTC) and the spatial metric Distance to Collision (DTC)? Our work, illustrated in Figure 1, confronts this question by proposing a Visual Question Answering (VQA) dataset to evaluate this capability. Such quantitative insights are crucial for downstream motion planners to make informed, precise decisions rather than resorting to overly conservative maneuvers.

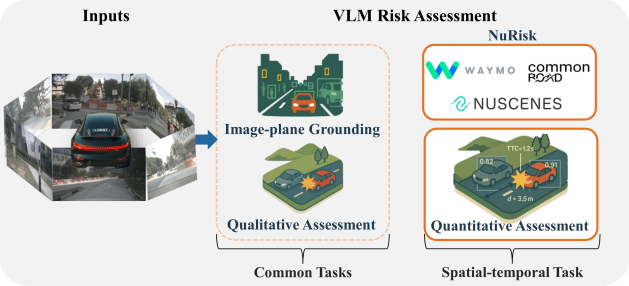

Fig. 1. Overview of NuRisk: Existing VLM-based risk assessment is typically limited to (i) image-plane grounding and (ii) qualitative assessment. NuRisk introduces an agent-level quantitative dataset that enables (iii) spatial–temporal quantitative assessment for risk reasoning.

A. Related Work

1) VLMs in Autonomous Driving: Recent studies have explored integrating VLMs into autonomous driving systems. Surveys such as [6], [9] provide comprehensive overviews of the role of VLMs in tasks including perception, motion planning, scene understanding, and visual reasoning. In the context of scenario analysis, VLMs have shown promising progress. Recently, a comprehensive survey [10] summarizes four main application areas of VLMs-based scenario analysis: VQA datasets, scene understanding, benchmarks, and risk assessment.

2) VLMs-based Risk Assessment in Autonomous Driving: Current approaches to VLM-based risk assessment can be categorized into two paradigms: prompting-based and fine-tuning-based methods. Prompting-based approaches adapt the reasoning capabilities of large pre-trained models via instructions. GPT-4V has been applied to structured scene

All authors are with the Professorship of Autonomous Vehicle Systems, TUM School of Engineering and Design, Technical University of Munich, 85748 Garching, Germany; Munich Institute of Robotics and Machine Intelligence (MIRMI)

representations for risk scoring with natural language justifications [11], demonstrating the potential for interpretable risk assessment. Similarly, frameworks like LATTE [7] and Ronecker et al. [12] integrate visual foundation models with contextual prompting for hazard detection and anomaly recognition.

Fine-tuning strategies like LoRA [13] have emerged to address domain-specific requirements and improve robustness by fine-tuning model parameters. Think-Driver [14] fine-tunes a VLM with chain-of-thought style VQA data with multi-view images for hazard reasoning and maneuver risk evaluation. By decomposing risk assessment into sequential reasoning steps, this approach provides more transparent, auditable decision-making than direct predictions. Lee et al. [15] adapt VLMs to occlusion-aware BEV representations for uncertainty prediction, while works like INSIGHT [16] enhance hazard localization and interpretability. And LKA-lert [17] focuses on failure anticipation by predicting failures in lane-keeping assist systems using multimodal cues. These methods are focused on qualitative agent risk assessment. However, these methods all lack the quantitative labels required for explicit spatio-temporal reasoning, a gap that our NuRisk dataset is precisely designed to fill.

3) VQA datasets in Autonomous Driving: VQA datasets for autonomous driving pair visual inputs with natural language queries to evaluate scene understanding across perception, prediction, and planning. Early works enriched perception tasks with BEV maps datasets [18], while later efforts extended to reasoning tasks, including counterfactual reasoning [19] for trajectory generation and video question answering datasets covering perception and prediction [20]. More recent benchmarks push toward multimodal reasoning across the full pipeline, incorporating step-by-step reasoning with images and LiDAR [21] or standardized multiple-choice evaluations for VLMs [22]. In terms of spatio-temporal reasoning, datasets like NuPlanQA [23] focus on agents' spatial relations recognition, and TumTraffic-VideoQA [24] focus on spatio-temporal grounding in driving scenarios. Additionally, other VQA datasets incorporate qualitative agent risk-aware reasoning, such as NuInstruct [25], HiLM-D [26] and DVBench [27]. However, none of the existing VQA datasets systematically evaluate spatial-temporal reasoning for quantitative risk assessment.

B. Critical Summary

To the best of our knowledge, the existing literature is limited by at least one of the following aspects:

1) Insufficient evaluation of spatio-temporal risk assessment: While VLMs demonstrate strong performance in static scene risk analysis [7], [11], [12], [14], their capabilities in reasoning about temporal dynamics, agent trajectories, and evolving risk scenarios remain largely unexplored in safety-critical contexts. Existing approaches predominantly analyze risk at individual time instances rather than understanding how risks develop and propagate over time.
2) Limited scope of existing VQA datasets: Current VQA datasets show a clear gap in risk-oriented evaluation. While datasets like NuPlanQA [23] and TumTraffic-VideoQA [24] focus on perception and spatio-temporal understanding, and works like NuInstruct [25] and DVBench [27] incorporate risk-aware elements, no benchmark systematically evaluates agent-level risk assessment that combines spatio-temporal reasoning with quantitative, safety-critical metrics.
3) Lack of safety-critical scenario coverage: Existing real-world datasets, such as Waymo Open Motion [28] and nuScenes [29], predominantly capture normal driving scenarios, with limited representation of safety-critical situations. This limits the evaluation of VLM's risk reasoning capabilities in high-stakes scenarios, where accurate risk assessment is most crucial to autonomous driving safety.

C. Contribution

To address the previous limitations, the key contributions of this paper are the following:

1) We introduce **NuRisk**, a VQA dataset with 2.9K scenarios and 1.1M agent-level samples for agent-level quantitative risk reasoning, completed with synthetic safety-critical collision scenarios to ensure comprehensive coverage of rare events.
2) We provide a systematic evaluation of pre-trained off-the-shelf VLMs across prompting strategies, highlighting their strengths and limitations in quantitative risk assessment.
3) We propose a fine-tuning pipeline for open-source VLMs, leveraging a parameter-efficient (LoRA) approach to adapt a 7B model for specialized risk assessment tasks, capable of explicit spatio-temporal reasoning, significantly outperforming pre-trained baselines.

II. Methodology

This paper introduces NuRisk, a novel VQA dataset for agent-level risk assessment in autonomous driving. While high-quality datasets like nuScenes [29] and Waymo Open [28] are widely used by VLMs for their diverse scenarios and multimodal sensor data from Lidar, Camera, Radar, IMU, etc [10], a gap exists: none provide ground truth for quantitative, agent-level risk across spatio-temporal scene sequences. To address this, NuRisk utilizes scenarios from these two high-impact datasets and augments them with safety-critical collision scenarios from the CommonRoad simulator [30]. The resulting framework, shown in Figure 2, serves as a base for quantitative risk analysis for individual traffic participants.

A. NuRisk Dataset Construction

As illustrated in Figure 3 a), we construct NuRisk from 1000 Waymo scenarios, 850 NuScenes scenarios, and 1000 safety-critical scenarios from CommonRoad simulation, where Waymo and NuScenes primarily capture normal driving behavior, and CommonRoad emphasizes safety-

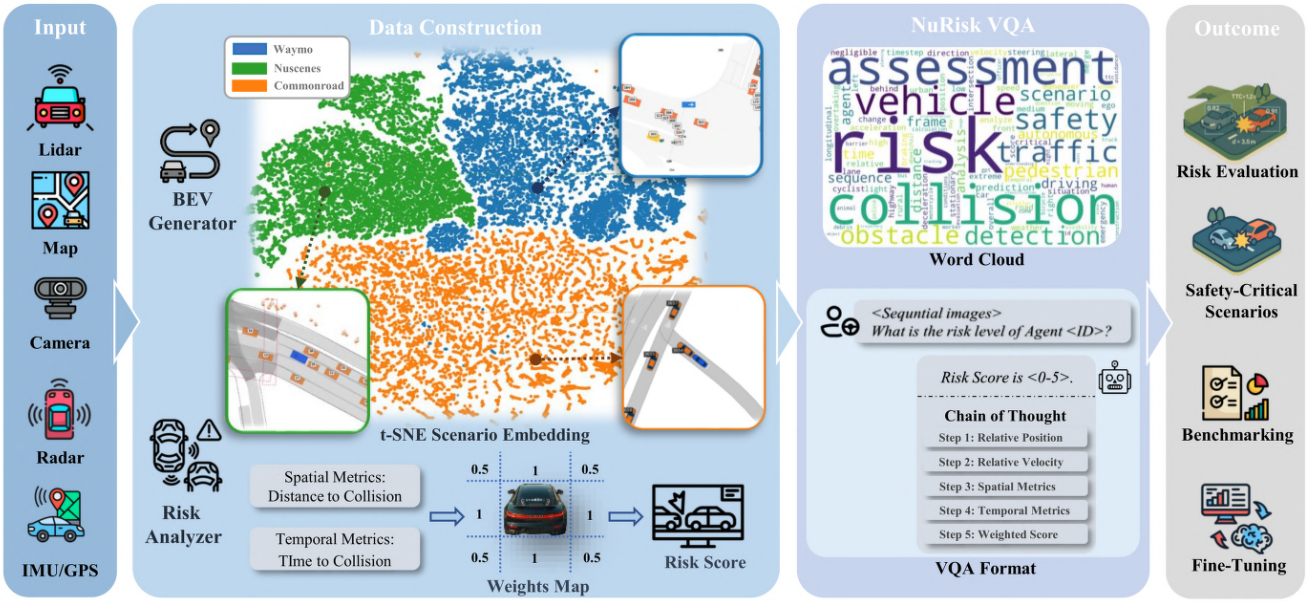

Fig. 2. Framework of NuRisk. Multi-modal inputs are processed into BEV scenes and risk metrics to enable conversation-based VQA with chain-of-thought reasoning, supporting risk evaluation, benchmarking, fine-tuning, and safety-critical scenario analysis.

critical corner cases. This multi-source composition provides coverage across scenarios characterized by minimum agent risk levels and agent categories. Synthetic scenarios are generated via motion planning in a Commonroad with road networks, traffic signs, and realistic traffic participants. All scenarios are organized into a structured set S that contains driving environment information. We process these structured scenarios through a multi-stage pipeline shown in Algorithm 1:

Stage 1: Dataset-Specific Data Extraction. We extract ego and traffic agent data using dataset-specific methods. For the Waymo Open dataset, we parse TFRecord files to extract ego-vehicle states and all tracked objects, including their trajectories, categories, and dimensional properties. For the nuScenes dataset, we use scene tokens to split the data into individual scenarios and extract ego poses and agent annotations from the structured files, preserving object categories, sizes, and temporal sequences. For CommonRoad scenarios, we obtain ego trajectories from motion planner outputs while extracting other traffic participants from recorded traffic data. Across all datasets, we algorithmically compute velocities and accelerations from position sequences.

Stage 2: BEV Image Generation. We adopt Bird's-eye view (BEV) representations to isolate spatio-temporal risk reasoning from perception uncertainty. RGB inputs introduce confounding factors such as occlusion and lighting variations, making it difficult to determine whether failures stem from perception limitations or deficiencies in the reasoning ability we aim to evaluate. We prepare dataset-specific maps and visualization components, then generate BEV images. For nuScenes scenarios, we utilize the map extension API to extract lane boundaries, crosswalks, and road surface information. For Waymo Open datasets, we parse map features from TFRecord files, including lane geometries and traffic infrastructure elements. For CommonRoad, we employ Lanelet2 visualization libraries to render road networks and lane structures. This stage ensures consistent BEV representation across all three data sources while preserving dataset-specific road topology and infrastructure details. For each timestep, the pipeline renders a BEV image focused on a 30-meter radius around the ego vehicle, visualizing all dynamic agents with distinct shapes and colors. Our BEV visualizations, as shown in Figure 2, match standard autonomous driving BEV representations, enabling seamless integration with real-world BEV generation algorithms.

Stage 3: Ground Truth Annotation. We compute physics-based risk annotations for each agent in each BEV image. To ensure temporal consistency in the unified dataset, all scenarios are resampled to a uniform 2Hz frequency, following nuScenes' sampling rate, which is the lowest among the constituent datasets. The pipeline transforms object coordinates into an ego-centric frame, calculates weighted risk scores based on longitudinal and lateral DTC and TTC, and stores the temporally aligned BEV images and risk annotations together in the final dataset $D_{unified}$.

B. VQA Dataset Creation

After generating the ground truth dataset, we prepare the data for VLM training and benchmarking through the following pipeline.

1) Preprocessing Pipeline: The proposed pipeline (Algorithm 1) constructs a unified BEV–risk dataset by jointly generating rasterized scene representations and structured risk annotations for all traffic participants.

For each scenario $s \in S$ and each resampled timestep t, the algorithm produces a BEV image I_t and a corresponding

Algorithm 1 BEV Generation and Risk Annotation

Input: Scenario Set S, Target Frequency H, DTC Threshold τ_{dtc}, TTC Threshold τ_{ttc}
Output: Unified dataset D
$D \leftarrow \emptyset$
for each scenario $s \in S$ **do**
  $(X_e(t), V_e(t)) \leftarrow$ GetEgoTrajectory(s)
  $\{(X_i(t), V_i(t))\} \leftarrow$ GetAgentTrajectories(s)
  $T_s \leftarrow$ GetTimestamps(s)
  $T_r \leftarrow$ ResampleIndices(T_s, H)
  for each $t \in T_r$ **do**
    ▷ **— BEV Generation —**
    $ego\_state \leftarrow X_e(t)$
    $agent\_states \leftarrow \{X_i(t) \mid$ agent i exists at $t\}$
    $agent\_states^{ego} \leftarrow$ ToEgo($agent\_states$, ego_state)
    $map_t \leftarrow$ GetMapFeatures(s, ego_state)
    $I_t \leftarrow$ Rasterize(map_t, $agent\_states^{ego}$)
    ▷ **— Risk Calculation —**
    $R_t \leftarrow \emptyset$
    for each agent $agent_i$ existing at time t **do**
      $\Delta x_i \leftarrow X_i(t) - X_e(t)$
      $\Delta v_i \leftarrow V_i(t) - V_e(t)$
      $dtc_i \leftarrow \|\Delta x_i\|$
      $ttc_i \leftarrow$ ComputeTTC(Δx_i, Δv_i)
      $r_i^{dtc} \leftarrow$ RiskFromDTC($\Delta x_i, \tau_{dtc}$)
      $r_i^{ttc} \leftarrow$ RiskFromTTC($\Delta x_i, \Delta v_i, \tau_{ttc}$)
      $r_i \leftarrow$ CombineRisk(r_i^{dtc}, r_i^{ttc})
      $R_t^i \leftarrow (agent_i,\ dtc_i,\ ttc_i,\ \Delta v_i,\ r_i^{dtc},\ r_i^{ttc},\ r_i)$
      $R_t \leftarrow R_t \cup \{R_t^i\}$
    end for
    $D \leftarrow D \cup \{(I_t,\ R_t)\}$
  end for
end for
return D

multi-agent risk set R_t. The resulting sequential dataset is defined as

$$\mathcal{D}_{seq} = \{(I_{1:T}, R_{1:T})\}, \tag{1}$$

where I_t denotes the BEV representation at timestep t, and

$$R_t = \left\{ R_t^{(a)} \mid a \in \mathcal{A}_t \right\} \tag{2}$$

is the set of per-agent risk annotations at time t. Each agent-level record is defined as

$$R_t^{(a)} = \left(\mathrm{id}^{(a)}, \mathrm{dtc}^{(a)}, \mathrm{ttc}^{(a)}, \Delta v^{(a)}, r_{dtc}^{(a)}, r_{ttc}^{(a)}, r^{(a)} \right), \tag{3}$$

For agent-centric modeling, the dataset can be reorganized into per-agent sequences:

$$\mathcal{D}_{agent} = \left\{ \left(I_{1:T}, R_{1:T}^{(a)} \right) \right\}_{a \in \mathcal{A}}, \tag{4}$$

Subsequently, we perform image optimization to ensure compatibility with VLMs input limits, resizing images to meet token constraints while maintaining aspect ratios to preserve visual quality. Finally, we convert the processed data into the following conversation format.

2) Conversation Format Conversion: To ensure broad compatibility and leverage a proven instruction-following structure, we adopt the widely used LLaVA conversation format [31]. Each Visual Question Answering sample contains an *image* field specifying the sequential image path and a *conversations* array with alternating *human* and *gpt* messages. *Human* queries include the `<image>` token and request agent-specific risk analysis, while *gpt* responses provide structured JSON output containing risk assessments with numerical scores, spatio-temporal distance tracking, and chain of thought explanations. This format enables the model to explicitly articulate the reasoning process for predicting agent behavior and assessing risk across temporal sequences, as shown in Figure 2. We also include related word clouds for conversation analysis in the framework visualization.

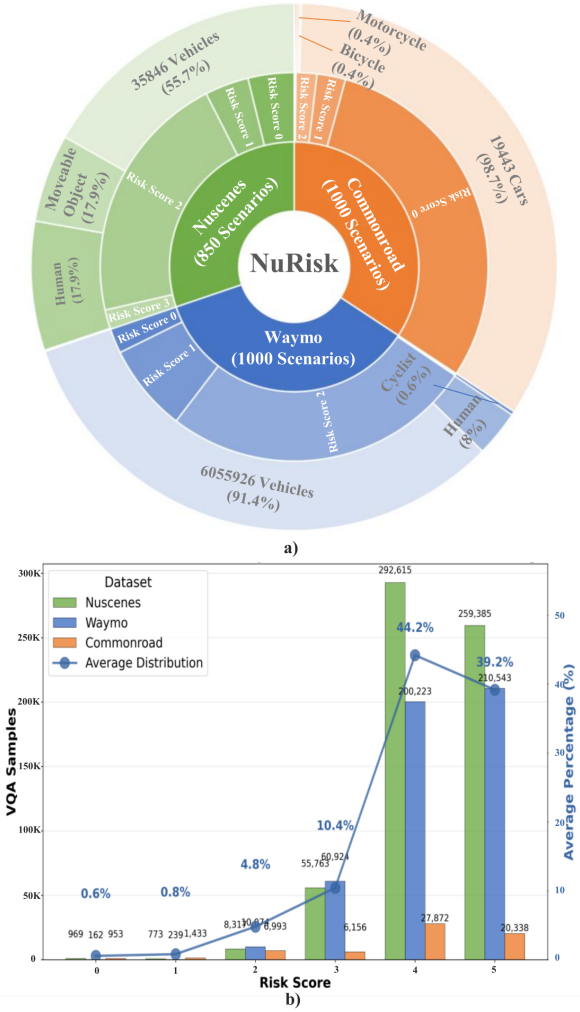

Fig. 3. Dataset statistics and risk distribution of NuRisk. Risk scores range from 0 (highest risk/collision) to 5 (lowest risk). (a) Scenario-level composition across data sources, where each scenario's risk level is defined by the minimum agent risk score within that scenario. Due to its safety-critical design, CommonRoad contains a large proportion of risk-0 scenarios. (b) Agent-level risk distribution in the final NuRisk dataset. High-risk interactions typically involve only a small subset of agents. Consequently, at the agent level, most agents exhibit lower-risk levels (4–5).

This format enables direct compatibility with existing VLMs training frameworks and ensures consistent, structured outputs for quantitative risk assessment.

3) Dataset Quality Assurance: We implement validation throughout the preprocessing pipeline to ensure dataset reliability. Automated quality checks include risk score validation within established ranges, image file integrity verification,

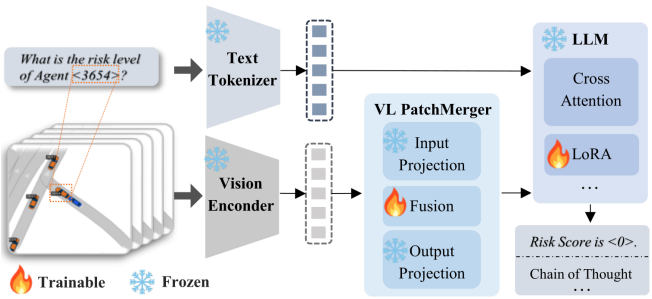

Fig. 4. NuRisk VLM Agent Fine-tuning Architecture.

conversation pair completeness assessment, JSON structure validation, agent data field verification, and timestep alignment confirmation. Additionally, we conduct human validation of sampled data to verify that sequential images properly display target agents and to cross-validate risk-level assessments against human annotations. The pipeline automatically filters invalid entries and provides detailed processing statistics while maintaining strict quality standards for VLM training effectiveness.

The final NuRisk VQA dataset comprises 1.1M agent-level VQA samples (617K from nuScenes, 482K from Waymo, 64K from CommonRoad). As shown in Figure 3 b), risk scores are balanced across levels 0-5, ensuring safety-critical coverage for robust VLM training,

C. Outcome Analysis

1) Benchmarking for pre-trained VLMs: To evaluate off-the-shelf VLMs on risk assessment tasks using our proposed NuRisk dataset, we employ a comprehensive evaluation framework that combines zero-shot inference with prompting techniques to adapt pre-trained models for autonomous driving risk analysis.

Our evaluation strategy incorporates three key adaptation techniques: *Contextual Prompting (CP)*, which augments input prompts with task instruction and related driving safety metrics information and collision patterns like safety metrics with explicit thresholds; *Chain-of-Thought reasoning (CoT)*, which enables step-by-step analysis of agent behavior and quantitative risk assessment across temporal sequences; and *In-Context Learning (ICL)*, which leverages selected risk scenario exemplars to establish clear reasoning patterns for understanding different risk driving situations. Results from this evaluation framework, also known as prompting strategies or ablation studies, will be presented in III.

2) Fine-tuning NuRisk VLM Agent: To enhance the spatio-temporal reasoning capability of VLMs for autonomous driving agent risk analysis, we fine-tune a VLM agent specifically designed for the NuRisk dataset.

We adopt Qwen2.5-VL-7B-Instruct as the base model for its strong multimodal reasoning and moderate parameter count, making it well-suited for parameter-efficient fine-tuning. Its 7B size also remains practical for real-world deployment on edge platforms [10]. As shown in Figure 4, the model consists of three main components: (1) a *Vision Encoder*, kept frozen to preserve robust visual feature extraction, (2) a *Language Model*, where base weights are frozen but LoRA adapters are inserted into the attention and feed-forward layers for domain-specific adaptation, and (3) a *Vision Language Merger*, where the input and output projections remain frozen while the Fusion layer is LoRA-adapted to improve cross-modal alignment for risk reasoning.

We employ causal language modeling for risk assessment conversation completion, optimizing the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i|x_i, I_i; \theta) \tag{5}$$

where y_i, x_i, I_i, and θ denote the target response, conversation context, sequential images, and model parameters, respectively. The supervision is provided by structured JSON annotations from the NuRisk VQA dataset. Model selection utilizes validation every 150 training steps, with the optimal checkpoint chosen by $\theta^* = \arg\min_\theta \mathcal{L}_{eval}(\theta)$ on held-out validation data, ensuring robust generalization to unseen driving scenarios. This design enables efficient domain adaptation of large VLMss, strengthening agent-level spatio-temporal risk reasoning without retraining the entire model.

Furthermore, this fine-tuning pipeline is designed to be transferable, as Qwen's architecture shares similarities with the LLaVA framework. The LLaVA conversation format, therefore, provides a standardized training interface, while the LoRA-based parameter-efficient fine-tuning approach enables efficient adaptation to other LLaVA-like VLM architectures without requiring full model retraining.

III. Results & Discussion

This section presents our three-stage evaluation. First, we benchmark pre-trained VLMs on NuRisk across different prompting techniques (Experiment 1). Second, we incorporate physics-based information to enhance risk assessment (Experiment 2). Finally, we fine-tune a 7B VLM on NuRisk to examine whether domain-specific training can close the performance gap (Experiment 3).

A. Experimental Setup

All experiments were conducted on a workstation equipped with an AMD Ryzen 7 9800X3D CPU, 96 GB of RAM, and an NVIDIA GeForce RTX 5090 GPU.

1) Evaluated Vision Language Models: Our evaluation includes a suite of leading proprietary and open-source VLMs. For proprietary models, we accessed Gemini-2.5-Pro, Gemini-2.5-Flash, GPT-5-Mini, Qwen-VL-Max, and Qwen-VL-Plus via their respective APIs. For open-source models, we conducted local inference using Qwen2.5-VL-7B-Instruct and InternVL3-8B, using the LMDeploy[1] framework.

2) Evaluation Metrics: VLM performance is assessed across three categories. For risk assessment accuracy, we use *Mean Absolute Error (MAE)* to quantify the average magnitude of error in the ordinal risk predictions, *Quadratic Weighted Kappa (QWK)* to evaluate the agreement between

[1]https://github.com/InternLM/lmdeploy

predicted and ground-truth risk levels, standard *Accuracy (Acc)*, and *Precision, Recall, F1-Score* for agent risk score analysis. For spatio-temporal reasoning, we measure *Spatial Accuracy* (percentage of longitudinal and lateral predictions within a 0.5-meter tolerance) and *Temporal Accuracy (TTC)* (percentage of longitudinal and lateral predictions within a 0.5-second tolerance). For computational efficiency, we measure the average *Response Time* in seconds.

B. Experiment 1: Benchmarking for pre-trained VLMs

We conduct our first experiment to evaluate the spatio-temporal reasoning capabilities of pre-trained VLMs across different prompting techniques using the NuRisk dataset. This experiment focuses on assessing whether current VLMs can perform quantitative risk assessment using mainly image sequences.

Each model is assessed using a baseline model (zero-shot) and with advanced prompting strategies: contextual prompting (CP), chain-of-thought (CoT), and in-context learning (ICL) to establish comprehensive baseline performance across different adaptation techniques, as emphasized in II-C.1. To provide deeper insights into model capabilities, we assess performance using the metrics defined in our evaluation framework. An analysis examining the effectiveness of different prompting strategies across multiple VLMs is presented in Table I.

TABLE I
PERFORMANCE OF PRE-TRAINED VLMS WITH VISION-ONLY INPUT ACROSS DIFFERENT PROMPTING STRATEGIES.

Model + Technique	MAE↓	QWK↑	Acc↑	Time↓
Proprietary Models				
Gemini-2.5-Flash (Baseline)	1.91	0.49	0.15	38.29
Baseline + CP	**1.23**	**0.89**	**0.33**	**45.88**
Baseline + CP + CoT	1.20	0.87	0.30	96.67
Baseline + CP + CoT + ICL	1.20	0.88	0.32	107.32
Gemini-2.5-Pro (Baseline)	2.00	0.49	0.15	40.13
Baseline + CP	**1.25**	**0.88**	**0.33**	**45.24**
Baseline + CP + CoT	1.15	0.88	0.31	95.43
Baseline + CP + CoT + ICL	1.22	0.86	0.30	106.55
Qwen-VL-Plus (Baseline)	0.63	0.55	0.13	8.81
Baseline + CP	0.58	0.62	0.17	35.81
Baseline + CP + CoT	0.62	0.59	0.16	37.40
Baseline + CP + CoT + ICL	**0.62**	**0.69**	**0.22**	**48.22**
Qwen-VL-Max **(Baseline)**	**1.26**	**0.04**	**0.22**	**13.54**
Baseline + CP	0.96	0.05	0.16	14.20
Baseline + CP + CoT	1.20	-0.05	0.12	26.54
Baseline + CP + CoT + ICL	1.02	0.07	0.18	22.18
GPT-5-Mini **(Baseline)**	**1.16**	**0.14**	**0.30**	**27.67**
Baseline + CP	1.23	0.07	0.25	40.49
Baseline + CP + CoT	1.14	0.06	0.25	47.65
Baseline + CP + CoT + ICL	1.24	0.05	0.25	53.34
Open-Source Models				
InternVL3-8B (Baseline)	0.62	0.56	0.14	6.55
Baseline + CP	0.55	0.70	0.20	7.15
Baseline + CP + CoT	**0.54**	**0.70**	**0.21**	**11.88**
Baseline + CP + CoT + ICL	0.33	0.72	0.19	11.68
Qwen2.5-VL-7B (Baseline)	1.87	0.51	0.14	11.98
Baseline + CP	1.88	0.46	0.12	11.60
Baseline + CP + CoT	1.76	0.58	0.18	16.66
Baseline + CP + CoT + ICL	**1.53**	**0.68**	**0.22**	**19.64**

Performance comparison of prompting strategies. Arrows indicate if higher (↑) or lower (↓) values are better. CP: Contextual Prompting, CoT: Chain-of-Thought, ICL: In-Context Learning.

1) Proprietary vs. Open-Source Performance: Proprietary models significantly outperform open-source alternatives, with Gemini variants achieving the highest QWK scores (0.88-0.89) and peak accuracy (0.33). Open-source models reach a maximum QWK of 0.72 (InternVL3-8B) and an accuracy of 0.22 (Qwen2.5-VL-7B). However, response time is significantly higher for proprietary models than for open-source models, with proprietary models taking 45.88-107.32 s, while open-source models take 6.55-19.64 s.

2) Prompting Strategy Effectiveness: Proprietary models demonstrate a clear preference for informational context over procedural guidance. Contextual prompting can enhance performance by providing task-specific details. Conversely, advanced strategies like chain-of-thought and in-context learning, which dictate reasoning steps, yield negligible improvements while increasing latency. This suggests that these models' internal reasoning mechanisms are not easily augmented by external templates and may even be hindered by them. Furthermore, even for the best proprietary models like Gemini-2.5-Flash or Gemini-2.5-Pro, the moderate accuracy rates ($\leq$0.33) indicate visual information alone is insufficient for reliable autonomous driving risk assessment. While our analysis in Table I is based on the final risk score, a crucial next step is to examine the full reasoning outputs of these models to better evaluate the nuances of their spatio-temporal reasoning capabilities.

C. Experiment 2: Physics-Enhanced Input Configuration Analysis

To address the performance limitations observed in Experiment 1, we conduct an evaluation incorporating physics-based information like position, velocity, and acceleration from the ego vehicle and nearby vehicles into the input modalities. From the paper [32], LLMs with physical information from sensor data can already reach more than 0.8 accuracy for the risk assessment task with a hand-crafted prompt. We evaluate the performance of both proprietary and open-source models across different input configurations, including single-image and multi-image sequences. We also assess their performance across various prompting strategies, including contextual, chain-of-thought, and in-context prompting.

1) Impact of Input Modality: As shown in Table II, incorporating textual physics information yields substantial performance gains, particularly for proprietary models. Gemini-2.5-Flash, for instance, achieves a peak accuracy of 0.92 and an MAE of 0.19, significantly outperforming both its vision-only counterpart in Table I and the text-only Gemini-1.5-Pro baseline from the paper [32]. In contrast, open-source models fail to realize similar improvements. This is largely due to their limited context windows, which cannot effectively process token-intensive textual data. For InternVL3-8B, this results in a dramatic increase in response time from 1.36 s to over 32 s.

The addition of sequential-image data provides a more modest benefit. While Gemini-2.5-Flash sees a slight accuracy increase from 0.91 to 0.92, open-source models

TABLE II

PERFORMANCE ANALYSIS OF VLMS WITH PHYSICS-ENHANCED AND SEQUENTIAL-IMAGE INPUTS.

Model + Technique + Modality	MAE↓	QWK↑	Acc↑	Time↓
Proprietary Models				
Gemini-1.5-Pro [32] (Text)	-	-	0.83	25.00
Gemini-2.5-Flash (Baseline) (Single)	1.08	0.44	0.11	7.66
Baseline + CP (Single+Text)	0.21	0.83	0.91	24.51
Baseline + CP + CoT (Single+Text)	0.20	0.84	0.91	32.63
Baseline + CP + CoT + ICL (Single+Text)	0.22	0.82	0.90	31.57
Gemini-2.5-Flash (Baseline) (Multi)	1.91	0.49	0.15	38.29
Baseline + CP (Multi+Text)	0.20	0.84	0.92	26.76
Baseline + CP + CoT (Multi+Text)	0.20	0.84	0.92	34.72
Baseline + CP + CoT + ICL (Multi+Text)	**0.19**	**0.85**	**0.92**	**40.55**
Open-Source Models				
InternVL3-8B (Baseline) (Single)	1.65	0.14	0.13	1.36
Baseline + CP (Single+Text)	1.55	0.20	0.20	1.54
Baseline + CP + CoT (Single+Text)	**0.95**	**0.44**	**0.39**	**21.14**
Baseline + CP + CoT + ICL (Single+Text)	1.23	0.39	0.30	32.84
InternVL3-8B (Baseline) (Multi)	0.62	0.56	0.14	6.55
Baseline + CP (Multi+Text)	1.71	0.18	0.16	2.01
Baseline + CP + CoT (Multi+Text)	1.13	0.30	0.33	23.17
Baseline + CP + CoT + ICL (Multi+Text)	1.18	0.40	0.38	33.18
Qwen2.5-VL-7B (Baseline) (Single)	1.95	0.12	0.12	2.63
Baseline+ CP (Single+Text)	1.70	0.18	0.18	2.96
Baseline + CP + CoT (Single+Text)	1.71	0.18	0.18	13.02
Baseline + CP + CoT + ICL (Single+Text)	1.43	0.27	0.18	44.91
Qwen2.5-VL-7B (Baseline) (Multi)	1.87	0.51	0.14	11.98
Baseline + CP (Multi+Text)	1.78	0.16	0.16	6.50
Baseline + CP + CoT (Multi+Text)	**1.55**	**0.20**	**0.25**	**22.95**
Baseline + CP + CoT + ICL (Multi+Text)	1.58	0.20	0.23	33.18

Performance with physics-enhanced inputs. CP: Contextual Prompting, CoT: Chain-of-Thought, ICL: In-Context Learning. Multi/Single refers to Sequential/single-image input.

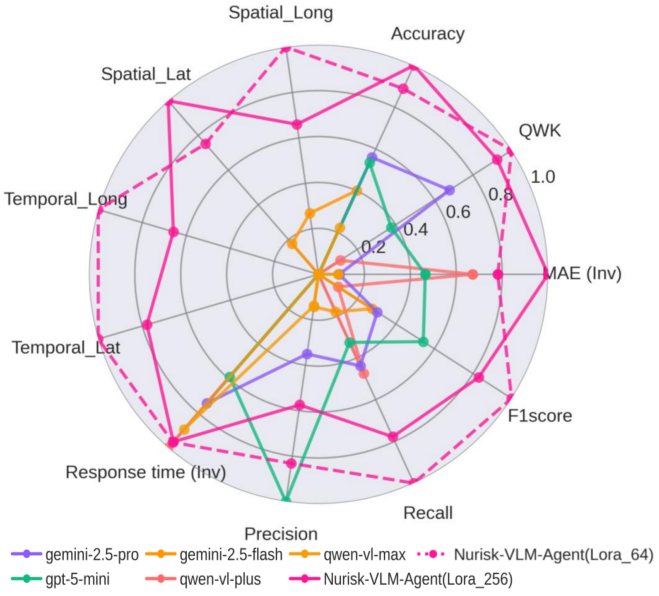

Fig. 5. Performance comparison between the best proprietary VLMs and our two fine-tuned NuRisk VLM Agent configurations, with the inverse scale of the response time and MAE to make it more interpretable.

derive no clear advantage, as the increased token load from multiple images further strains their processing capabilities and increases latency. This suggests that for risk assessment at the current timestep, explicit physics data is the most critical input, with sequential visual context offering only marginal gains.

2) Impact of Prompting Strategy: From the ablation study, we reach the same conclusion as shown in Table II: for pre-trained reasoning models, performance with contextual prompting is better than zero-shot performance. However, other advanced prompting strategies do not improve performance well while substantially increasing response time. This occurs because the proposed prompting strategies do not enhance the internal reasoning of models that already have access to physical information and primarily rely on visual-spatio-temporal capabilities for task reasoning. So Multi-modal Large Language Models (MLLMs) look promising for the risk assessment task, since they can reason about the physical information from the other sensors and the visual information.

D. Experiment 3: Fine-tuning the NuRisk VLM Agent

Experiments 1 and 2 revealed a critical gap: while proprietary VLMs like gemini-2.5-pro with 0.33 accuracy cannot achieve reasonable quantitative risk assessment accuracy and operate as high-latency ”black boxes.” To address these shortcomings, we developed the NuRisk VLM Agent by fine-tuning a Qwen2.5-VL-7B-Instruct as introduced in Section II-C.2. We fine-tuned the agent on 50,000 samples from the NuRisk dataset for 2 epochs, creating 2 variants with LoRA ranks of 64 and 256.

The comparative performance is visualized in Figure 5, which is normalized to our NuRisk VLM Agent, and we also invert the scale of the response time and MAE to make it more interpretable. Our fine-tuned agents significantly outperform all proprietary models across nearly every metric, demonstrating the effectiveness of specialized training.

Our fine-tuned agents demonstrate a superior balance of accuracy and efficiency. The LoRA-256 agent achieves a peak accuracy of 41.1% (MAE: 1.01, QWK: 0.28), representing a significant improvement over its pre-trained baseline accuracy of 14% and outperforming the best proprietary model (Gemini-2.5-Pro at 33%). Furthermore, it maintains an average response time of 10.2 seconds, which is 4 times faster than the leading proprietary model, making it more viable for real-world applications.

The most critical distinction lies in spatio-temporal reasoning. Proprietary models completely fail in this domain, even with the provided collision patterns in the prompt. In contrast, our NuRisk LoRA-256 agent exhibits strong performance, achieving longitudinal spatial accuracies of 34.1% and lateral spatial accuracies of 26.0%, and temporal longitudinal accuracies of 27.0% and lateral accuracies of 26.4%. This indicates it has learned the underlying causal relationship between vehicle dynamics and risk, proving our agents are not merely classifying risk but are reasoning about spatio-temporal information. The two fine-tuning configurations offer distinct advantages. The LoRA-256 agent is optimized for maximum classification accuracy. The LoRA-64 agent, while slightly less accurate, provides

a more balanced performance profile, achieving the highest QWK score (0.304) and showing stronger performance on temporal reasoning metrics. Both configurations confirm the effectiveness of domain-specific fine-tuning.

IV. Conclusion and Future Work

We introduced a framework for advancing spatio-temporal VLM reasoning in autonomous driving. We presented the NuRisk dataset with 2.9K scenarios and 1.1M agent-level samples, a diverse collection from nuScenes, Waymo, and CommonRoad. Based on the NuRisk dataset, we showed that even top proprietary models like Gemini-2.5-Pro (33% accuracy) fail at spatio-temporal reasoning (scoring zero on all spatio-temporal metrics). We address this gap by showing that physics-enhanced inputs can boost accuracy to 92%. Building on this, we introduced the NuRisk VLM Agent, fine-tuned on a small dataset and achieving 41.1% accuracy while running four times faster than the best proprietary model. Crucially, our agent is the only model to demonstrate the causal spatio-temporal reasoning necessary for true quantitative risk understanding. For future work, we plan to expand the NuRisk dataset with more challenging scenarios and further optimize the NuRisk VLM Agent for real-world deployment by lightweight VLM models with low latency.

ACKNOWLEDGMENT

The manuscript was initially drafted by the authors, with AI tools used to improve grammar and clarity.

References

[1] Waymo, "Waymo one: The next step on our self-driving journey," 2018. [Online]. Available: https://waymo.com/blog/2018/12/waymo-one-next-step-on-our-self-driving

[2] S. International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE J3016*, 2021.

[3] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani *et al.*, "Perception, planning, control, and coordination for autonomous vehicles," *Machines*, vol. 5, no. 1, p. 6, 2017.

[4] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger *et al.*, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] J. Betz, M. Lutwitzi, and S. Peters, "A new taxonomy for automated driving: Structuring applications based on their operational design domain, level of automation and automation readiness," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 1–7.

[6] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer *et al.*, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[7] J. Zhang, Y. Guan, C. Wang, H. Liao, G. Zhang *et al.*, "Latte: Lightweight attention-based traffic accident anticipation engine," *arXiv preprint arXiv:2504.04103*, 2025.

[8] M. Abu Tami, H. I. Ashqar, M. Elhenawy, S. Glaser, and A. Rakotonirainy, "Using multimodal large language models (mllms) for automated detection of traffic safety-critical events," *Vehicles*, vol. 6, no. 3, pp. 1571–1590, 2024.

[9] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris *et al.*, "Large (vision) language models for autonomous vehicles: Current trends and future directions," *Authorea Preprints*, 2024.

[10] Y. Gao, M. Piccinini, Y. Zhang, D. Wang, K. Moller *et al.*, "Foundation models in autonomous driving: A survey on scenario generation and scenario analysis," *IEEE Open Journal of Intelligent Transportation Systems*, pp. 1–1, 2026.

[11] H. Hwang, S. Kwon, Y. Kim, and D. Kim, "Is it safe to cross? interpretable risk assessment with gpt-4v for safety-aware street crossing," in *2024 21st International Conference on Ubiquitous Robots (UR)*. IEEE, 2024, pp. 281–288.

[12] M. P. Ronecker, M. Foutter, A. Elhafsi, D. Gammelli, I. Barakaiev *et al.*, "Vision foundation model embedding-based semantic anomaly detection," *arXiv preprint arXiv:2505.07998*, 2025.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[14] Q. Zhang, M. Zhu, and H. F. Yang, "Think-driver: From driving-scene understanding to decision-making with vision language models," in *European Conference on Computer Vision Workshop*, 2024.

[15] J. Lee, J. Cho, H. Suk, and S. Kim, "SFF rendering-based uncertainty prediction using visionLLM," in *AAAI 2025 Workshop LM4Plan*, 2025.

[16] D. Chen, Z. Zhang, Y. Liu, and X. T. Yang, "Insight: Enhancing autonomous driving safety through vision-language models on context-aware hazard detection and edge case evaluation," 2025.

[17] Y. Wang and H. Zhou, "Bridging human oversight and black-box driver assistance: Vision-language models for predictive alerting in lane keeping assist systems," *arXiv preprint arXiv:2505.11535*, 2025.

[18] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain *et al.*, "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 345–16 352.

[19] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi *et al.*, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.

[20] M. Nie, R. Peng, C. Wang, X. Cai, J. Han *et al.*, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 292–308.

[21] A. Ishaq, J. Lahoud, K. More, O. Thawakar, R. Thawkar *et al.*, "Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding," *arXiv preprint arXiv:2503.10621*, 2025.

[22] B. Khalili and A. W. Smyth, "Autodrive-qa-automated generation of multiple-choice questions for autonomous driving datasets using large vision-language models," *arXiv preprint arXiv:2503.15778*, 2025.

[23] S.-Y. Park, C. Cui, Y. Ma, A. Moradipari, R. Gupta *et al.*, "Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models," *arXiv preprint arXiv:2503.12772*, 2025.

[24] X. Zhou, K. Larintzakis, H. Guo, W. Zimmer, M. Liu *et al.*, "Tumtraffic-videoqa: A benchmark for unified spatio-temporal video understanding in traffic scenes," *arXiv preprint arXiv:2502.02449*, 2025.

[25] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 668–13 677.

[26] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving," *arXiv preprint arXiv:2309.05186*, 2023.

[27] T. Zeng, L. Wu, L. Shi, D. Zhou, and F. Guo, "Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding," *arXiv preprint arXiv:2504.14526*, 2025.

[28] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[29] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[30] M. Althoff, M. Koschi, and S. Manzinger, "Commonroad: Composable benchmarks for motion planning on roads," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 719–726.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[32] Y. Gao, M. Piccinini, K. Moller, A. Alanwar, and J. Betz, "From words to collisions: Llm-guided evaluation and adversarial generation of safety-critical driving scenarios," in *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2025, pp. 1–8.