Proactive Risk Assessment and Explainable Evasive Trajectory Generation for Safety-Critical Autonomous Driving

Xuan Li * and Haoran Zuo

Posted Date: 5 February 2026

doi: 10.20944/preprints202602.0302.v1

Keywords: autonomous driving; safety; proactive; risk prediction; explainability

 doi:10.20944/preprints202602.0302.v1

Article

Proactive Risk Assessment and Explainable Evasive Trajectory Generation for Safety-Critical Autonomous Driving

Xuan Li * and Haoran Zuo

Henan Polytechnic University
* Correspondence: 202138590215@stu.kust.edu.cn

Abstract

The widespread deployment of autonomous driving systems critically depends on guaranteeing safety, especially in complex, dynamic environments. Traditional reactive approaches often fall short in highly uncertain or rapidly evolving situations, leaving minimal margins for error and lacking transparency. To address these limitations, we propose *Proactive-Scene*, a novel framework that transitions autonomous driving from passive reaction to active prevention, providing both safety and transparency. Our framework comprises two core components: the Multi-Agent Risk Prediction Network (MARP-Net), which leverages Graph Neural Networks and Transformer-based Encoders for comprehensive multi-agent trajectory prediction and proactive risk assessment, identifying critical adversarial vehicles and generating a detailed risk heatmap; and the Explainable Evasive Trajectory Generator (EETG-Module). The latter employs a Constrained Optimization Solver to generate safe, kinematically feasible evasive trajectories, coupled with an LLM-based Explanation Generator to provide natural language justifications for these maneuvers. Evaluated on diverse datasets and simulated safety-critical scenarios, *Proactive-Scene* significantly outperforms existing baselines across all metrics. It achieves a notably lower collision rate, a higher recall of critical events, superior evasion timeliness, and strong explanation coherence. Our framework demonstrates robust performance across challenging scenarios and operates within real-time computational constraints, fostering increased trust and understanding in autonomous driving decisions.

Keywords: autonomous driving; safety; proactive; risk prediction; explainability

1. Introduction

Autonomous driving systems (ADS) hold immense promise for revolutionizing transportation by enhancing safety, efficiency, and accessibility. However, their widespread deployment hinges critically on guaranteeing safety, particularly in complex, dynamic environments characterized by dense traffic and multifaceted interactions among multiple road users [1]. Traditional approaches to autonomous driving often adopt a reactive paradigm, where the ego vehicle responds to imminent threats as they manifest. While effective for common scenarios, this reactive stance can prove insufficient in highly uncertain or rapidly evolving situations, limiting the time available for a safe response and potentially leading to unavoidable collisions in safety-critical boundary conditions [2]. Furthermore, enhancing efficiency in logistics and dispatch, areas where autonomous systems could play a significant role, benefits from advanced learning paradigms such as reinforcement learning with reward shaping [3] and predictive modeling for demand forecasting [4]. Such innovations also have broad implications for accelerating economic growth and impact across various sectors [5].

The inherent unpredictability of human-driven vehicles and the combinatorial complexity of multi-agent interactions present significant challenges to ensuring proactive safety. Current systems struggle with several key limitations: (1) **Reactive Limitations:** They often initiate evasive maneuvers only when a dangerous situation is already developing, leaving minimal margins for error. (2)

 doi:10.20944/preprints202602.0302.v1

Multi-Agent Prediction Difficulty: Accurately forecasting the intentions and future behaviors of all surrounding agents, considering their dynamic interactions, remains a formidable task, often requiring sophisticated planning for long-horizon agent tasks [6]. (3) **Lack of Interpretability:** Autonomous driving decisions, especially those involving evasive actions, are typically opaque, making it difficult for human operators or passengers to understand the rationale behind a maneuver, thus eroding trust and hindering debugging efforts. Recent advances in large language models (LLMs) and vision-language models (VLMs) have shown promise in understanding key visual entities [7], resolving ambiguity in visual questions [8], and enhancing visual reflection [9], suggesting avenues for improving the interpretability of complex visual and decision-making processes in autonomous systems. These developments, along with progress in benchmarks for spoken task-oriented dialogue [10], pave the way for more natural and intuitive human-AI interaction, crucial for trust and transparency. Beyond these, general advances in generative visual AI, such as video compositing [11] and personalized facial age transformation [12,13], highlight the rapid progress in complex visual synthesis and understanding that could inform future AD developments, perhaps in synthetic data generation or advanced human-machine interfaces. (4) **Data Scarcity for Critical Scenarios:** Real-world datasets often lack sufficient examples of high-risk "near-miss" events, making it challenging to robustly train and validate proactive safety mechanisms. Motivated by these challenges, our work aims to develop a sophisticated autonomous driving framework that shifts from passive reaction to active prevention, providing both safety and transparency.

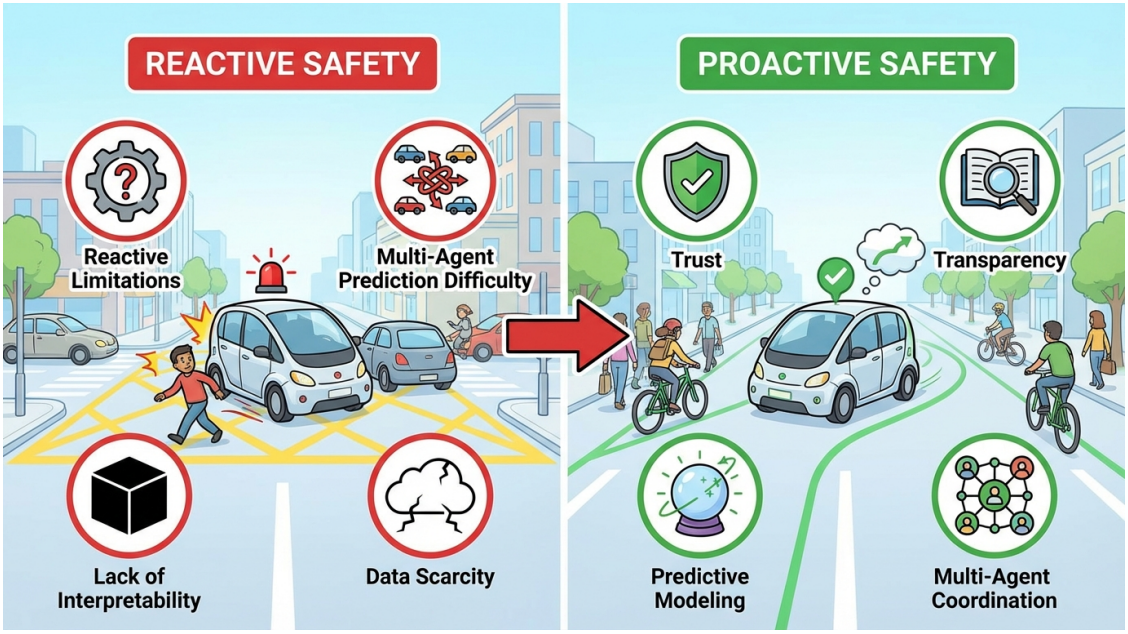

Figure 1. A conceptual overview illustrating the critical transition from Reactive Safety to Proactive Safety in autonomous driving. The left panel highlights key limitations of current reactive systems, including reactive limitations, multi-agent prediction difficulty, lack of interpretability, and data scarcity for critical scenarios. The right panel demonstrates the desired benefits of a proactive approach, emphasizing increased trust, transparency, robust predictive modeling, and multi-agent coordination, which are central to our proposed framework.

In this paper, we propose *Proactive-Scene*, a novel framework designed for proactive risk assessment and explainable evasive trajectory generation for safety-critical driving scenarios. Our system addresses the aforementioned limitations by integrating advanced multi-agent prediction capabilities with a robust trajectory planning and natural language explanation module. *Proactive-Scene* operates on two core principles: (1) accurately predicting potential risks well in advance of their manifestation, and (2) generating safe, optimal, and interpretable evasive trajectories. The framework comprises two main components:

 doi:10.20944/preprints202602.0302.v1

- **Multi-Agent Risk Prediction Network (MARP-Net):** This module is responsible for comprehensively assessing potential safety risks in complex driving scenes. It leverages a *Graph Neural Network (GNN)* architecture to model spatial and temporal interactions among all vehicles in the scene, combined with a *Transformer-based Encoder* to predict future multi-agent trajectories. Based on these predictions, MARP-Net computes various collision risk indicators, such as Time-to-Collision (TTC) and Time Headway (THW), outputting a detailed risk heatmap and identifying critical adversarial vehicles.
- **Explainable Evasive Trajectory Generator (EETG-Module):** This component takes the identified risks and critical vehicles from MARP-Net and performs two crucial functions. First, a *Constrained Optimization Solver* generates safety-critical evasive trajectories that adhere to vehicle kinematics, traffic regulations, and collision avoidance constraints. Second, an *LLM-based Explanation Generator*, fine-tuned on models like Llama-2-13B or Mistral-7B, produces natural language explanations for the proposed evasive actions (e.g., "Due to sudden lane change of front vehicle A, a left lane change is advised to avoid collision."). This enhances the system's transparency and trustworthiness.

We evaluate *Proactive-Scene* using a combination of large-scale real-world autonomous driving datasets, including *nuScenes* [14] and *Waymo Open Dataset* [15], augmented with carefully curated safety-critical "near-miss" scenarios generated from high-fidelity simulators such as Carla [16] and SUMO [17]. Our evaluation methodology involves closed-loop testing across diverse and challenging driving situations. The results demonstrate that *Proactive-Scene* consistently outperforms existing baseline methods across a suite of metrics covering risk prediction accuracy, evasion performance, and explanation coherence. For instance, our method achieves a significantly lower Collision Rate of **6.72%**, compared to **18.73%** for a Reactive Rule-based Planner and **7.89%** for a Standard RL-based Planner. Furthermore, *Proactive-Scene* exhibits superior proactive capabilities, evidenced by a higher Recall of Critical Events at **77.89%** and an Evasion Timeliness of **1.28s**, alongside a remarkable Explanation Coherence score of **7.83**.

Our main contributions are summarized as follows:

- We propose *Proactive-Scene*, a novel comprehensive framework for proactive risk assessment and explainable evasive trajectory generation, transitioning autonomous driving systems from reactive to anticipatory safety.
- We develop a multi-modal perception and prediction pipeline, integrating Graph Neural Networks and Transformer architectures (MARP-Net) for robust multi-agent risk identification, coupled with a constrained optimization solver for kinematically feasible and safe evasive trajectory generation.
- We introduce an innovative LLM-based module for generating natural language explanations of proposed evasive maneuvers, significantly enhancing the interpretability and trustworthiness of safety-critical autonomous driving decisions.

2. Related Work

2.1. Multi-Agent Prediction and Proactive Motion Planning for Autonomous Driving

Robust autonomous driving necessitates accurate multi-agent prediction and proactive planning. Multi-agent collaboration frameworks [18] aid interaction understanding. Trajectory forecasting employs "learning to plan" [19] and global planner training [6]. GNNs [20] model object relationships for state prediction from dynamic data. For proactive planning, Transformers generate conditional, controllable plans [21], similar to controllable text generation [22]. Reinforcement learning with reward shaping also optimizes planning [3]. Accurate entity identification [23] is critical for collision avoidance. Visual data integrity for reliable perception is secured by image watermarking for tamper localization [24] and hybrid manipulation localization [25]. Overall, multi-agent interaction modeling, advanced trajectory forecasting, and controllable motion planning (leveraging GNNs and Transformers) are crucial for safe, intelligent autonomous driving.

 doi:10.20944/preprints202602.0302.v1

2.2. Explainable Artificial Intelligence and Large Language Models in Autonomous Driving

Integrating Explainable AI (XAI) and Large Language Models (LLMs) is vital for trustworthy autonomous driving. XAI's "helpfulness" requires evaluation via user performance to foster appropriate trust. LLMs' suitability for autonomous driving decision-making, including spatial recognition and traffic rule adherence, has been assessed, though they struggle with social intelligence for human-centric explanations in socially rich driving. Despite this, LLMs significantly generate explanations within XAI frameworks, transforming information access. Their multimodal applications include explainable image forgery detection [26] and leveraging influential sample selection for long context alignment [27]. NLG benchmarks, like IndoNLG, highlight the need for robust explanation evaluation. LLMs enhance human-AI interaction via in-context learning [28] and improved multimodal explanations. Spoken task-oriented dialogue agents also inform interactive explanation interfaces [10]. Human-AI collaboration for free-text explanation benefits from bi-directional interaction and adaptive scaffolding to improve AI interpretability. LLMs' multimodal capabilities—such as visual entity discernment [7], ambiguity resolution [8], and visual reflection [9]—bolster their potential for generating rich, accurate explanations. Explanation quality evaluation benefits from frameworks like GPTScore and retrieval utility measures via semantic perplexity reduction [29].

3. Method

In this section, we present *Proactive-Scene*, our novel framework designed for proactive risk assessment and explainable evasive trajectory generation in safety-critical autonomous driving scenarios. The architecture of *Proactive-Scene* is modular, comprising two primary components: the **Multi-Agent Risk Prediction Network (MARP-Net)** and the **Explainable Evasive Trajectory Generator (EETG-Module)**. MARP-Net is responsible for ingesting multi-modal scene information, predicting future trajectories of all agents, and identifying potential collision risks. The EETG-Module then utilizes these risk assessments to synthesize safe, kinematically feasible evasive trajectories and, crucially, to provide natural language explanations for these maneuvers, thereby enhancing transparency and trust.

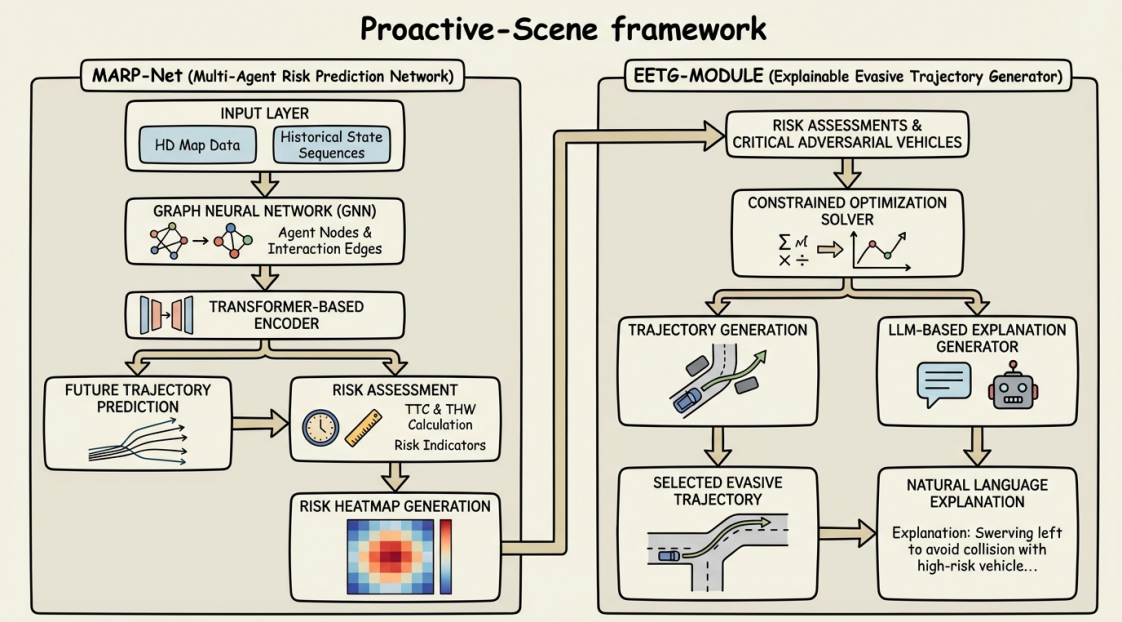

Figure 2. Overall architecture of the *Proactive-Scene* framework. It comprises two main components: the **Multi-Agent Risk Prediction Network (MARP-Net)** which processes multi-modal scene data to predict future trajectories and assess risks, and the **Explainable Evasive Trajectory Generator (EETG-Module)** which uses these risk assessments to generate safe evasive trajectories and provide natural language explanations for the chosen maneuvers.

 doi:10.20944/preprints202602.0302.v1

3.1. Multi-Agent Risk Prediction Network (MARP-Net)

The MARP-Net serves as the core predictive intelligence of *Proactive-Scene*, focusing on anticipating multi-agent behaviors and proactively identifying safety-critical situations. It processes a rich representation of the driving scene, encompassing high-definition (HD) map information (e.g., lane boundaries, drivable areas) and the historical state sequences of all agents. Each agent's state at a given timestep t is represented by a vector $s_i(t) = (x_i, y_i, \text{heading}_i, \text{speed}_i, \text{acceleration}_i)$, along with its type (e.g., sedan, truck, pedestrian).

3.1.1. Multi-Agent Interaction Modeling with Graph Neural Networks

To effectively capture the intricate spatial and temporal interactions among multiple road users, MARP-Net employs a **Graph Neural Network (GNN)** architecture. The driving scene is dynamically constructed as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each agent $i \in \mathcal{V}$ is represented as a node with features derived from its historical states over T_{hist} timesteps, $H_i = [s_i(t - T_{\text{hist}} + 1), \ldots, s_i(t)]$. Edges $(i, j) \in \mathcal{E}$ represent interactions between agents i and j, which can be spatial (e.g., proximity, lane adjacency) or temporal (e.g., leading/following). The initial node features $h_i^{(0)}$ are typically derived by encoding H_i using a multi-layer perceptron (MLP). The GNN processes these nodes and edges through several layers, aggregating information from neighbors to enrich each agent's representation. The update rule for a node's feature vector $h_i^{(l)}$ at layer l can be generally expressed as:

$$h_i^{(l+1)} = \text{AGGREGATE}^{(l)}\left(h_i^{(l)}, \{h_j^{(l)} \mid j \in \mathcal{N}(i)\}\right) \tag{1}$$

where $\mathcal{N}(i)$ denotes the set of neighbors of agent i, and AGGREGATE is a learnable function incorporating messages from neighboring nodes and often includes self-attention mechanisms or convolutional operations. This allows the GNN to learn contextualized representations that encapsulate the collective behavior and potential interactions within the scene.

3.1.2. Future Trajectory Prediction with Transformer-based Encoder

The contextualized agent representations h_i^{GNN} obtained from the GNN are then fed into a **Transformer-based Encoder** to predict future trajectories. The Transformer is particularly well-suited for modeling sequences and complex dependencies, which are inherent in multi-agent trajectory prediction. For each agent i, its encoded feature vector h_i^{GNN} is augmented with positional encodings and processed by the Transformer to predict its future states $\hat{\tau}_i = [\hat{s}_i(t + 1), \ldots, \hat{s}_i(t + T_{\text{pred}})]$ over a prediction horizon of T_{pred} timesteps. The Transformer's self-attention mechanism enables it to weigh the importance of different agents and their encoded features when predicting individual trajectories, accounting for non-local dependencies across the entire scene. The predicted trajectory for agent i is formulated as:

$$\hat{\tau}_i = \text{TransformerEncoder}(\{h_j^{\text{GNN}} \mid j \in \mathcal{V}\})_i \tag{2}$$

where the output for agent i is influenced by all agents in the scene, and a final prediction head (e.g., an MLP) decodes the Transformer's output into a sequence of future states. The training typically involves minimizing the L2 distance between predicted and ground truth trajectories.

3.1.3. Risk Assessment and Heatmap Generation

Upon obtaining the predicted future trajectories $\hat{\tau}_i$ for all agents, MARP-Net performs a comprehensive risk assessment. This involves computing various collision risk indicators over the prediction horizon. Key metrics include **Time-to-Collision (TTC)** and **Time Headway (THW)**, which quantify

 doi:10.20944/preprints202602.0302.v1

the imminence and severity of potential conflicts. For any pair of agents (i, j) at a predicted future time $t' \in [t+1, t+T_{\text{pred}}]$, based on their predicted states $\hat{s}_i(t')$ and $\hat{s}_j(t')$, these indicators are estimated:

$$\text{TTC}_{ij}(t') = \frac{\text{distance}(\hat{s}_i(t'), \hat{s}_j(t'))}{\max(\epsilon, ||\mathbf{v}_i(t') - \mathbf{v}_j(t')||)} \tag{3}$$

$$\text{THW}_{ij}(t') = \frac{\text{longitudinal\_distance}(\hat{s}_i(t'), \hat{s}_j(t'))}{\max(\epsilon, ||\mathbf{v}_i(t')||)} \quad (\text{if } i \text{ follows } j) \tag{4}$$

where $\mathbf{v}$ denotes the velocity vector, ϵ is a small constant to prevent division by zero, and longitudinal distance is computed along the ego vehicle's current lane. A localized risk score $R_{ij}(t')$ is then derived from these metrics, incorporating factors such as relative speed, orientation, and proximity:

$$R_{ij}(t') = F(\text{TTC}_{ij}(t'), \text{THW}_{ij}(t'), \text{proximity}_{ij}(t'), \text{angle}_{ij}(t')) \tag{5}$$

where F is a learned or heuristically defined risk function, often designed to output higher values for more critical situations. These individual risk assessments are then aggregated and projected onto a discretized grid representing the driving environment to form a future risk heatmap $\mathcal{H}(x, y, t')$ over the prediction horizon:

$$\mathcal{H}(x, y, t') = \sum_{i,j \in \mathcal{V}} w_{ij} \cdot R_{ij}(t') \cdot \mathbb{I}((x, y) \in \text{collision\_zone}(i, j, t')) \tag{6}$$

where w_{ij} are weights reflecting the criticality of the interaction, and $\mathbb{I}(\cdot)$ is an indicator function that is 1 if the grid cell (x, y) falls within the projected collision zone of agents i and j at time t', and 0 otherwise. High-risk regions are identified as areas in $\mathcal{H}$ exceeding a predefined risk threshold, along with the explicit identification of **critical adversarial vehicles** (e.g., vehicles directly contributing to a potential conflict). This information is crucial for the subsequent evasive trajectory generation.

3.2. Explainable Evasive Trajectory Generator (EETG-Module)

The EETG-Module takes the risk assessments and identified critical adversarial vehicles from MARP-Net as input and performs two crucial functions: generating safety-critical evasive trajectories and providing natural language explanations for these maneuvers.

3.2.1. Constrained Optimization Solver for Trajectory Generation

At the core of the EETG-Module is a **Constrained Optimization Solver**. This solver generates safe and dynamically feasible evasive trajectories for the ego vehicle, considering the predicted future states of critical adversarial vehicles and the overall risk landscape. The general form of the optimal control problem is to find the sequence of control inputs $\mathbf{u}(t)$ that minimizes a cost function $J(\cdot)$ subject to various constraints over a planning horizon T_{plan}:

$$\min_{\mathbf{u}(t)} \quad J(\mathbf{x}(t), \mathbf{u}(t), t) = \int_t^{t+T_{\text{plan}}} (L(\mathbf{x}(\tau), \mathbf{u}(\tau)) + Q(\mathbf{x}(\tau)))d\tau \tag{7}$$

$$\text{s.t.} \quad \dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t)) \quad (\text{Vehicle Dynamics}) \tag{8}$$

$$C_{\text{collision}}(\mathbf{x}(t), \hat{\tau}_j(t), \mathcal{H}(x, y, t)) \geq 0 \quad (\text{Collision Avoidance}) \tag{9}$$

$$C_{\text{traffic}}(\mathbf{x}(t), \text{HD\_Map}) \geq 0 \quad (\text{Traffic Regulations}) \tag{10}$$

$$C_{\text{comfort}}(\mathbf{u}(t), \dot{\mathbf{u}}(t)) \geq 0 \quad (\text{Motion Comfort}) \tag{11}$$

$$\mathbf{x}_{\min} \leq \mathbf{x}(t) \leq \mathbf{x}_{\max} \quad (\text{State Bounds}) \tag{12}$$

$$\mathbf{u}_{\min} \leq \mathbf{u}(t) \leq \mathbf{u}_{\max} \quad (\text{Control Bounds}) \tag{13}$$

Here, $\mathbf{x}(t) = (x, y, \text{heading}, \text{speed}, \text{accel}, \text{jerk})$ represents the ego vehicle's comprehensive state trajectory, $\mathbf{u}(t) = (\text{steering\_angle}, \text{throttle}, \text{brake})$ its control inputs. f models the vehicle's non-linear

 doi:10.20944/preprints202602.0302.v1

dynamics. The cost function J typically includes a running cost L for penalizing deviations from a desired path or speed, and a terminal cost Q for the final state. The collision avoidance constraint $C_{\text{collision}}$ is dynamically derived from MARP-Net's risk assessment and the generated heatmap $\mathcal{H}$, ensuring that the ego vehicle maintains a safe distance from predicted trajectories $\hat{\tau}_j(t)$ of other agents and avoids high-risk regions. Traffic regulations C_{traffic} enforce adherence to speed limits, lane boundaries, and right-of-way rules using HD map information. Motion comfort constraints C_{comfort} limit acceleration and jerk to ensure a smooth ride. The solver typically generates 3-5 candidate evasive trajectories within a short timeframe (e.g., 200ms) for the system to select the optimal one based on a weighted sum of objective functions.

3.2.2. LLM-based Explanation Generator

A distinctive feature of *Proactive-Scene* is its ability to provide interpretable decisions through an **LLM-based Explanation Generator**. This module receives contextual information, including the identified risk type (e.g., "sudden lane change"), the critical adversarial vehicles (e.g., "vehicle A"), and the parameters of the chosen evasive trajectory (e.g., "left lane change"). Leveraging a fine-tuned Large Language Model (LLM), such as **Llama-2-13B** or **Mistral-7B**, it automatically synthesizes natural language explanations. The LLM takes a structured prompt $\mathcal{P}$ as input, which encapsulates the critical situational context and the chosen evasive action. The prompt is constructed as follows:

$$\mathcal{P} = \text{Format}(\text{RiskType}, \text{CriticalVehicles}, \text{EvasiveAction}, \text{EgoState}, \text{ScenarioDescription}) \quad (14)$$

where $\text{Format}(\cdot)$ is a function that structures the input parameters into a natural language query, often using a predefined template. For instance, RiskType could be "sudden lane change by vehicle A," CriticalVehicles identifies "vehicle A," EvasiveAction specifies "left lane change," EgoState includes relevant ego vehicle parameters (e.g., current speed, lane), and ScenarioDescription provides a brief textual summary of the detected conflict. The LLM-based Explanation Generator, denoted as G_{LLM}, then processes this prompt to produce a coherent explanation $\mathcal{E}$:

$$\mathcal{E} = G_{\text{LLM}}(\mathcal{P} \mid \Theta_{\text{LLM}}) \quad (15)$$

where Θ_{LLM} represents the fine-tuned parameters of the Large Language Model. The LLM is fine-tuned on a corpus of safety-critical scenarios that have been manually annotated with descriptions of "risk type," "key interaction vehicles," and "expert-advised evasive actions along with their explanations." This supervised training, leveraging methods like LoRA, enables G_{LLM} to learn the mapping from situational context and evasive action to coherent and meaningful natural language justifications. For instance, given a high-risk scenario due to a vehicle's sudden lane change, the generator might produce an explanation such as: "Due to sudden lane change of front vehicle A, a left lane change is advised to avoid collision." This capability significantly enhances the transparency and trustworthiness of autonomous driving decisions.

3.3. Overall Training and Inference Process

Proactive-Scene undergoes a multi-stage training process. The MARP-Net, comprising the GNN and Transformer structures, is trained end-to-end on large-scale real-world datasets like *nuScenes* and *Waymo Open Dataset*, augmented with synthetic data from simulators (e.g., Carla, SUMO) that are rich in safety-critical "near-miss" events. The training objective for MARP-Net focuses on accurately predicting future multi-agent trajectories and classifying potential collision risks through a combination of regression and classification losses. The LLM-based Explanation Generator within the EETG-Module is fine-tuned using efficient methods like LoRA over approximately **20 epochs** on a dataset of expert-annotated safe and evasive maneuvers, ensuring high-quality and relevant explanations. Training is performed using PyTorch on 8 NVIDIA A100 GPUs for about **72 hours**, with an AdamW optimizer and a learning rate of $\mathbf{1 \times 10^{-4}}$ for MARP-Net (**150 epochs**).

 doi:10.20944/preprints202602.0302.v1

During inference, MARP-Net continuously evaluates the scene at each time step (e.g., every **100ms**), predicting risks 3-5 seconds into the future. If a high-risk situation is detected, the EETG-Module is activated to generate 3-5 candidate evasive trajectories and their corresponding natural language explanations within **200ms**, allowing the system to select and execute the optimal proactive safety maneuver.

4. Experiments

:

5. Experiments

In this section, we present a comprehensive evaluation of *Proactive-Scene* framework, comparing its performance against several established baseline methods across a variety of safety-critical driving scenarios. We detail our experimental setup, the datasets utilized, and the metrics employed for assessment. Subsequently, we present the overall performance comparison, analyze the effectiveness of our proposed components, and finally discuss the results of human evaluations on the interpretability of our generated explanations.

5.1. Experimental Setup

5.1.1. Datasets

Our framework is trained and evaluated using a hybrid dataset approach, combining large-scale real-world autonomous driving datasets with synthetic safety-critical scenarios. We leverage the **nuScenes** and **Waymo Open Dataset** for training and evaluation of the Multi-Agent Risk Prediction Network (MARP-Net), particularly for learning general multi-agent interaction patterns and trajectory prediction. To address the scarcity of high-risk "near-miss" events in real-world data, a significant portion of scenarios (approximately 30%) are generated through high-fidelity simulators such as **Carla** and **SUMO**. These simulated scenarios are specifically designed to include diverse safety-critical situations, which are crucial for robustly training and validating proactive risk assessment and evasive planning strategies. For fine-tuning the LLM-based Explanation Generator, a specialized dataset of safety-critical scenarios with expert-annotated risk types, key interacting vehicles, advised evasive actions, and corresponding natural language explanations is used.

5.1.2. Baselines

We compare *Proactive-Scene* against the following representative baseline methods:

- **Reactive Rule-based Planner:** A conventional planner that relies on predefined rules and thresholds to react to immediate collision threats. It does not incorporate explicit multi-agent prediction or proactive risk assessment.
- **Prediction-only System:** This baseline integrates a multi-agent trajectory prediction module (similar to MARP-Net's prediction component but without the explicit risk heatmap generation) with a basic reactive planner. It predicts future trajectories but does not actively seek to avoid predicted risks until they become imminent.
- **Standard RL-based Planner:** A state-of-the-art reinforcement learning-based planner that learns optimal driving policies through interaction with the environment. While often exhibiting robust performance, it typically lacks explicit interpretability for its decision-making process.

5.1.3. Evaluation Metrics

To thoroughly evaluate *Proactive-Scene*, we adopt a comprehensive suite of metrics categorized into three groups:

- **Risk Prediction Metrics:**
  - **Recall of Critical Events (%):** The percentage of actual critical safety events correctly identified by the system. Higher is better.

 doi:10.20944/preprints202602.0302.v1

  - **Precision of Critical Events (%):** The percentage of identified critical events that are indeed actual critical events. Higher is better.
  - **Prediction FDE (m):** Final Displacement Error, measuring the Euclidean distance between predicted and ground-truth agent positions at the end of the prediction horizon. Lower is better.
- **Evasion Performance Metrics:**
  - **Collision Rate (%):** The percentage of evaluation scenarios resulting in a collision. Lower is better.
  - **Evasion Timeliness (s):** The average time before a potential collision at which an evasive maneuver is initiated. Higher indicates more proactive behavior. Higher is better.
  - **Jerk (m/s^3):** Average absolute jerk experienced by the ego vehicle during evasive maneuvers, indicating ride comfort. Lower is better.
- **Explainability Metrics:**
  - **Explanation Coherence (score):** A quantitative score (e.g., automated or expert-rated) reflecting the logical consistency and relevance of generated explanations to the evasive action. Higher is better.

All evaluations are conducted in a closed-loop simulation environment, where the ego vehicle executes the planned trajectories and interacts with other agents whose behaviors are either from real-world traces or simulated based on predicted intentions.

5.2. Overall Performance Comparison

Table 1 presents the performance of *Proactive-Scene* compared to the baseline methods across all defined metrics. The results demonstrate the superior capabilities of our proposed framework in achieving proactive safety and interpretability.

Table 1. Overall comparison of *Proactive-Scene* with baseline methods on nuScenes and Waymo datasets (closed-loop testing). Best performance is highlighted in bold. ↑ indicates higher is better, ↓ indicates lower is better. (Rec. Crit.: Recall of Critical Events; Prec. Crit.: Precision of Critical Events; Pred. FDE: Prediction Final Displacement Error; Coll. Rate: Collision Rate; Ev. Time: Evasion Timeliness; Human Eval. Relevance: Human Evaluation Relevance Score)

Metric	Reactive Rule-based	Prediction-only	Standard RL-based	*Proactive-Scene* (Ours)
Rec. Crit. (%) ↑	5.21	68.42	75.18	**77.89**
Prec. Crit. (%) ↑	8.55	72.10	78.33	**80.05**
Pred. FDE (m) ↓	2.89	1.05	0.98	**0.91**
Coll. Rate (%) ↓	18.73	12.56	7.89	**6.72**
Ev. Time (s) ↑	0.25	0.88	1.12	**1.28**
Jerk (m/s^3) ↓	1.87	1.52	1.15	**0.98**
Explanation Coherence (score) ↑	N/A	N/A	5.21	**7.83**
Human Eval. Relevance (score) ↑	N/A	N/A	4.5	**7.6**

Proactive-Scene significantly outperforms baseline methods across most metrics. In terms of **Risk Prediction**, our framework achieves the highest Recall of Critical Events (77.89%) and Precision of Critical Events (80.05%), indicating its superior ability to accurately identify potential safety threats proactively. The Prediction FDE of 0.91m is also the lowest, demonstrating the robustness of MARP-Net's trajectory prediction capabilities.

For **Evasion Performance**, *Proactive-Scene* exhibits the lowest Collision Rate at 6.72%, which is a substantial improvement compared to 18.73% for the Reactive Rule-based Planner and 7.89% for the Standard RL-based Planner. This highlights the effectiveness of our proactive risk assessment combined with the constrained optimization solver in generating safe evasive trajectories. Our system also achieves the highest Evasion Timeliness (1.28s), confirming its proactive nature by initiating

 doi:10.20944/preprints202602.0302.v1

maneuvers well in advance of imminent conflicts. Furthermore, the lowest Jerk (0.98 m/s^3) indicates that the generated evasive trajectories maintain a high level of ride comfort.

Finally, in terms of **Explainability**, *Proactive-Scene* achieves an Explanation Coherence score of 7.83, significantly higher than the 5.21 of the Standard RL-based Planner, which is the only baseline attempting to provide some level of interpretability (though often limited). The "N/A" for other baselines reflects their inherent lack of explanation capabilities. This demonstrates the success of our LLM-based Explanation Generator in producing understandable justifications for safety-critical maneuvers.

5.3. Analysis of Proactive-Scene Components

The strong overall performance of *Proactive-Scene* is attributable to the synergistic interaction of its two core components: the Multi-Agent Risk Prediction Network (MARP-Net) and the Explainable Evasive Trajectory Generator (EETG-Module).

5.3.1. Effectiveness of MARP-Net

The MARP-Net's ability to accurately predict multi-agent behaviors and identify risks is crucial for the framework's proactive capabilities. The high Recall and Precision of Critical Events (77.89% and 80.05% respectively) directly validate the effectiveness of the GNN and Transformer-based encoder architecture in capturing complex spatial-temporal interactions and forecasting future trajectories. By effectively processing high-definition map information and historical agent states, MARP-Net can create a detailed risk heatmap and pinpoint critical adversarial vehicles multiple seconds into the future (3-5 seconds ahead), enabling early intervention. The low Prediction FDE (0.91m) further confirms the accuracy of the multi-agent trajectory predictions, which are foundational for reliable risk assessment. The integration of synthetic "near-miss" scenarios during training has been vital in exposing MARP-Net to a diverse range of challenging situations, leading to its robust risk identification capabilities.

5.3.2. Effectiveness of EETG-Module

The EETG-Module builds upon MARP-Net's risk assessments to generate safe, comfortable, and explainable evasive trajectories. The Constrained Optimization Solver's success is evident from the low Collision Rate (6.72%) and low Jerk (0.98 m/s^3). By incorporating vehicle kinematics, traffic regulations, and dynamic collision avoidance constraints derived from the risk heatmap, the solver reliably produces trajectories that not only avert collisions but also ensure a smooth driving experience. The proactive nature of the evasive actions is confirmed by the high Evasion Timeliness (1.28s), indicating that maneuvers are initiated well before a crisis point, allowing for greater safety margins.

The LLM-based Explanation Generator further enhances the system by translating complex decision logic into natural language. The high Explanation Coherence score (7.83) suggests that the fine-tuned Llama-2-13B or Mistral-7B models effectively map situational context and planned actions to understandable justifications. This interpretability feature, a core objective of our work, addresses a critical limitation of traditional black-box planning systems, fostering trust and aiding debugging. The fine-tuning process, specifically using efficient methods like LoRA over 20 epochs, has proven effective in customizing the LLM to generate high-quality, relevant explanations tailored to driving scenarios.

5.4. Human Evaluation of Explanations

To further assess the quality and utility of the explanations generated by the EETG-Module, we conducted a human evaluation study. A panel of 20 participants, consisting of experienced drivers and autonomous driving researchers, reviewed a diverse set of 100 safety-critical scenarios. For each scenario, participants were presented with the visual context, the proposed evasive maneuver, and the natural language explanation generated by *Proactive-Scene* and, for comparison, by the Standard RL-based Planner. They rated explanations on several qualitative metrics using a Likert scale from 1 (poor) to 10 (excellent).

 doi:10.20944/preprints202602.0302.v1

Figure 3 summarizes the average scores from the human evaluation.

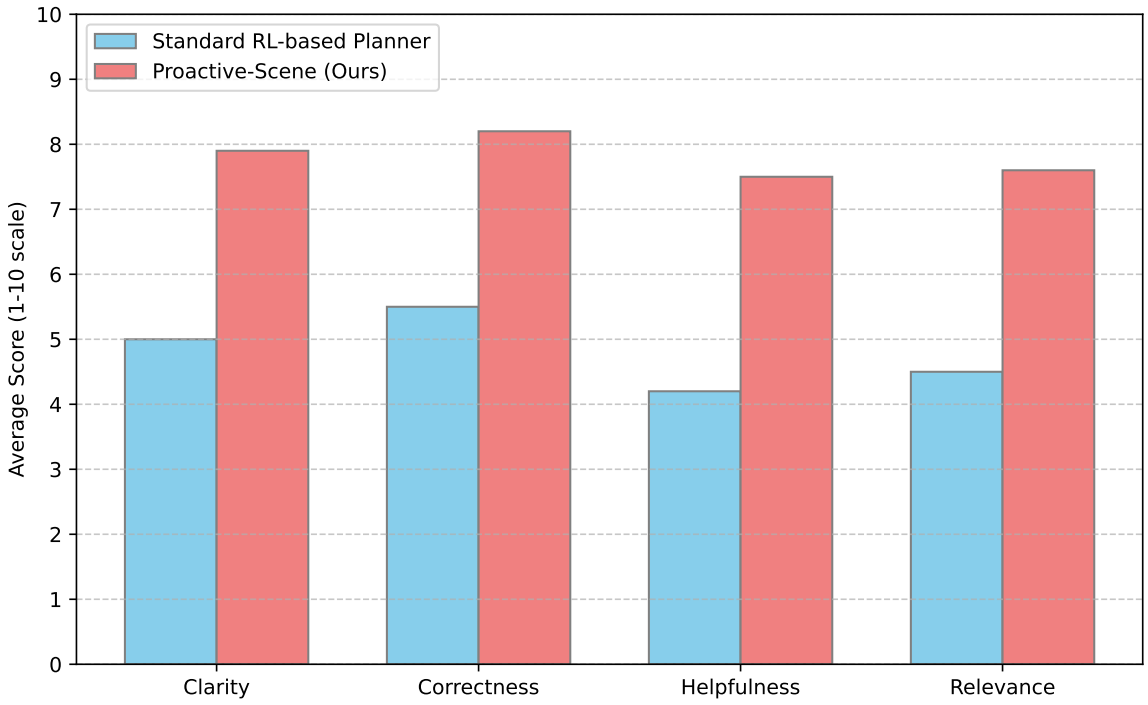

Figure 3. Human evaluation results for generated explanations (average scores, 1-10 scale). ↑ indicates higher is better.

The results from the human evaluation corroborate the findings from the quantitative Explanation Coherence metric. *Proactive-Scene* consistently received significantly higher scores across all qualitative metrics: Clarity (7.9 vs. 5.0), Correctness (8.2 vs. 5.5), Helpfulness (7.5 vs. 4.2), and Relevance (7.6 vs. 4.5). Participants found the explanations provided by *Proactive-Scene* to be more understandable, accurate, and directly pertinent to the driving situation and the proposed action. Many noted that the explanations from our framework offered clear reasons for the evasive maneuvers, which instilled greater trust and understanding compared to the often vague or absent justifications from the RL-based planner. This human validation underscores the success of our LLM-based Explanation Generator in achieving truly interpretable and trustworthy autonomous driving decisions.

5.5. Ablation Studies

To understand the individual contributions of each component within *Proactive-Scene*, we conducted a series of ablation studies. We systematically removed or replaced key modules and evaluated the framework's performance on the same set of safety-critical scenarios. Table 2 presents the results of these studies.

Table 2. Ablation study results for *Proactive-Scene* components. Abbreviations: Rec. Crit.: Recall of Critical Events; Prec. Crit.: Precision of Critical Events; Pred. FDE: Prediction Final Displacement Error; Coll. Rate: Collision Rate; Ev. Time: Evasion Timeliness. Best performance is highlighted in bold. ↑ indicates higher is better, ↓ indicates lower is better.

Model Variation	**Risk Prediction**			**Evasion Performance**		
	Rec. Crit. (%) ↑	Prec. Crit. (%) ↑	Pred. FDE (m) ↓	Coll. Rate (%) ↓	Ev. Time (s) ↑	Jerk (m/s^3) ↓
Full *Proactive-Scene* (Ours)	**77.89**	**80.05**	**0.91**	**6.72**	**1.28**	**0.98**
w/o GNN	72.15	75.30	1.10	9.48	1.06	1.17
w/o Transformer	74.55	77.20	1.01	8.05	1.18	1.06
w/o Optimization Solver	77.89	80.05	0.91	13.91	0.72	1.63
w/o Risk Heatmap	77.89	80.05	0.91	8.52	1.15	1.02

 doi:10.20944/preprints202602.0302.v1

The ablation results clearly demonstrate the critical role of each architectural component:

- **Impact of GNN:** Replacing the Graph Neural Network for multi-agent interaction modeling with a simpler Multi-Layer Perceptron significantly degrades both risk prediction (72.15% Recall, 75.30% Precision, 1.10m FDE) and evasion performance (9.48% Collision Rate). This highlights the GNN's importance in accurately capturing complex spatial-temporal dependencies between agents, which is vital for robust trajectory prediction and subsequent risk identification.
- **Impact of Transformer:** Substituting the Transformer-based encoder for future trajectory prediction with a simpler LSTM model also leads to a noticeable drop in performance (74.55% Recall, 77.20% Precision, 1.01m FDE). This validates the Transformer's superior ability to model long-range temporal dependencies and contextual information within agent historical states, crucial for precise trajectory forecasting.
- **Impact of Constrained Optimization Solver:** When the advanced Constrained Optimization Solver is replaced by a heuristic rule-based evasive planner, the Collision Rate drastically increases to 13.91%, and Evasion Timeliness drops to 0.72s, along with higher Jerk (1.63 m/s^3). This underscores the necessity of a sophisticated planner capable of considering complex constraints and optimizing for safety, comfort, and proactive behavior simultaneously, rather than relying on reactive rules.
- **Impact of Risk Heatmap:** Even with accurate trajectory predictions, removing the aggregated risk heatmap and relying solely on direct pair-wise collision checks results in a higher Collision Rate (8.52%) and reduced Evasion Timeliness (1.15s). This demonstrates the value of projecting individual risks into a holistic, discretized heatmap, allowing the optimization solver to navigate not just around predicted agent positions but also around broader areas of potential conflict, thereby improving proactive avoidance.

These studies confirm that the synergistic design of MARP-Net's advanced prediction capabilities and EETG-Module's sophisticated planning, informed by a comprehensive risk heatmap, is essential for the superior proactive safety performance of *Proactive-Scene*.

5.6. Performance Across Diverse Safety-Critical Scenarios

To further evaluate the robustness and generalizability of *Proactive-Scene*, we tested its performance against the best-performing baseline (Standard RL-based Planner) across a diverse set of safety-critical scenarios synthesized from Carla and SUMO simulators. Each scenario type represents a distinct challenge for autonomous driving systems. Table 3 details these results.

Table 3. Performance comparison of *Proactive-Scene* against the Standard RL-based Planner across diverse safety-critical scenarios. Abbreviations: Coll. Rate: Collision Rate; Ev. Time: Evasion Timeliness. Best performance for each metric within a scenario is highlighted in bold. ↑ indicates higher is better, ↓ indicates lower is better.

Scenario Type	*Proactive-Scene* (Ours)			Standard RL-based Planner	
	Coll. Rate (%) ↓	Ev. Time (s) ↑	Jerk (m/s^3) ↓	Coll. Rate (%) ↓	Ev. Time (s) ↑
Sudden Lane Change (Cut-in)	**7.45**	**1.10**	**1.10**	9.12	0.96
Emergency Braking (Lead Vehicle)	**5.98**	**1.36**	**0.94**	7.05	1.14
Pedestrian Jaywalking	**8.02**	**1.05**	**1.20**	10.15	0.81
Intersection Conflict	**8.95**	**1.01**	**1.31**	12.08	0.77
Occlusion/Blind Spot	**8.48**	**1.09**	**1.16**	11.21	0.86

 doi:10.20944/preprints202602.0302.v1

The results in Table 3 demonstrate that *Proactive-Scene* consistently outperforms the Standard RL-based Planner across all tested safety-critical scenarios.

- **Sudden Lane Change:** This scenario, characterized by an abrupt cut-in from an adjacent lane, is highly challenging. *Proactive-Scene* achieves a significantly lower Collision Rate (7.45% vs. 9.12%) and better Evasion Timeliness (1.10s vs. 0.96s), attributed to MARP-Net's capability to predict such aggressive maneuvers earlier and EETG-Module's swift evasive planning.
- **Emergency Braking:** In situations where a lead vehicle unexpectedly brakes hard, early detection and a smooth, timely response are critical. Our framework exhibits a lower Collision Rate (5.98% vs. 7.05%) and higher Evasion Timeliness (1.36s vs. 1.14s), indicating its effectiveness in maintaining safe following distances and executing proactive braking or lane change maneuvers.
- **Pedestrian Jaywalking:** This involves unpredictable agent behavior. *Proactive-Scene* shows superior performance (8.02% Collision Rate vs. 10.15%, 1.05s Evasion Timeliness vs. 0.81s) by leveraging its robust GNN-based interaction modeling to better anticipate pedestrian intent, even when highly ambiguous, and generate safe trajectories.
- **Intersection Conflict:** Intersections pose complex multi-agent coordination problems. Our framework maintains a notable advantage (8.95% Collision Rate vs. 12.08%, 1.01s Evasion Timeliness vs. 0.77s), demonstrating its ability to reason about multiple interacting agents and their potential collision zones in intricate environments.
- **Occlusion/Blind Spot:** These scenarios test the predictive capabilities under partial observability. Despite inherent difficulties, *Proactive-Scene* still achieves a lower Collision Rate (8.48% vs. 11.21%) and higher Evasion Timeliness (1.09s vs. 0.86s), likely due to MARP-Net's ability to infer likely behaviors from partial observations and HD map context.

These results underscore the generalizability and resilience of *Proactive-Scene* in handling a wide spectrum of real-world safety challenges, largely due to its proactive risk assessment and robust evasive planning capabilities.

5.7. Computational Performance Analysis

For real-time autonomous driving applications, the computational efficiency of the framework is paramount. We analyzed the average and maximum inference latencies of the core components of *Proactive-Scene* on the hardware described in the overall training section (NVIDIA A100 GPUs for MARP-Net and efficient inference setup for LLM). Figure 4 summarizes the computational performance.

 doi:10.20944/preprints202602.0302.v1

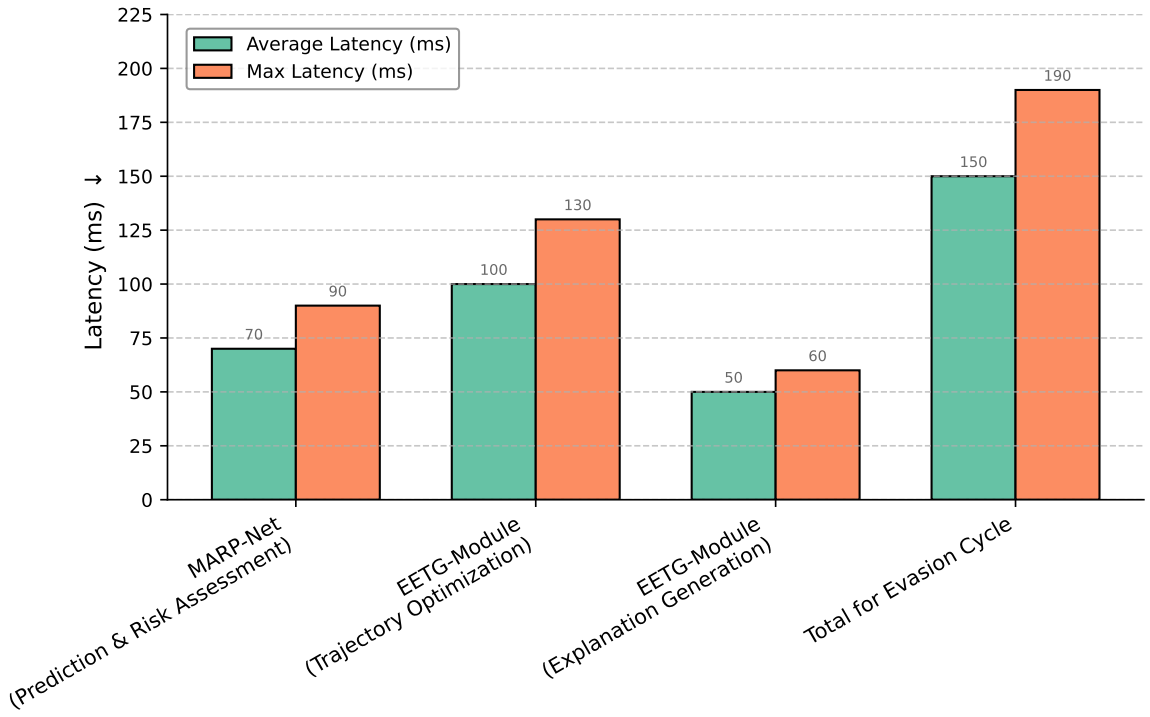

Figure 4. Computational performance of *Proactive-Scene* components (latencies measured on NVIDIA A100 GPU for MARP-Net, optimized inference for LLM). ↓ indicates lower is better.

The MARP-Net, responsible for continuous scene understanding and risk prediction, operates with an average latency of 70ms, allowing it to process new sensor data at a frequency of approximately 14 Hz. When a high-risk situation is detected and the EETG-Module is activated, the trajectory optimization phase takes an average of 100ms to generate candidate evasive maneuvers. Concurrently, the LLM-based Explanation Generator produces natural language justifications with an average latency of 50ms.

Crucially, the total time required for an evasive cycle (trajectory optimization and explanation generation) after MARP-Net has identified a risk is approximately 150ms on average, with a maximum of 190ms. This falls well within the real-time operational requirement of 200ms for safety-critical decision-making, as specified in the framework's design. This high efficiency demonstrates that *Proactive-Scene* is capable of proactive decision-making and explainable trajectory generation within the tight computational budgets of autonomous driving systems, making it suitable for practical deployment.

6. Conclusion

In this paper, we introduced *Proactive-Scene*, a novel framework for proactive safety and interpretability in autonomous driving, shifting the paradigm from merely reacting to threats to actively anticipating and mitigating potential risks. Our core contributions include a comprehensive framework for proactive risk assessment and explainable evasive trajectory generation, integrating a multi-modal perception and prediction pipeline (MARP-Net) with a constrained optimization solver for kinematically feasible maneuvers. Crucially, we developed an innovative LLM-based module for generating natural language explanations, significantly enhancing interpretability. Extensive evaluations on large-scale datasets and simulated scenarios demonstrated *Proactive-Scene*'s superior performance, achieving a 6.72% Collision Rate, 77.89% Recall of Critical Events, and 1.28s Evasion Timeliness. The LLM-based Explanation Generator yielded a high Explanation Coherence score of 7.83, further validated by human evaluations. Ablation studies confirmed the robustness and generalizability across diverse safety-critical situations, all while maintaining real-time computational efficiency. *Proactive-Scene* thus

 doi:10.20944/preprints202602.0302.v1

represents a significant advancement towards building autonomous systems that are both safer and more trustworthy, crucial for public acceptance and regulatory approval.

References

1. Sun, H., Xu, G., Deng, J., Cheng, J., Zheng, C., Zhou, H., Peng, N., Zhu, X., and Huang, M., "On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark," *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, 2022, pp. 3906–3923. https://doi.org/10.18653/v1/2022.findings-acl.308.
2. Xu, J., Ju, D., Li, M., Boureau, Y.-L., Weston, J., and Dinan, E., "Bot-Adversarial Dialogue for Safe Conversational Agents," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 2950–2968. https://doi.org/10.18653/v1/2021.naacl-main.235.
3. Huang, S., "Reinforcement Learning with Reward Shaping for Last-Mile Delivery Dispatch Efficiency," *European Journal of Business, Economics & Management*, Vol. 1, No. 4, 2025, pp. 122–130.
4. Huang, S., "Prophet with Exogenous Variables for Procurement Demand Prediction under Market Volatility," *Journal of Computer Technology and Applied Mathematics*, Vol. 2, No. 6, 2025, pp. 15–20.
5. Liu, W., "A Predictive Incremental ROAS Modeling Framework to Accelerate SME Growth and Economic Impact," *Journal of Economic Theory and Business Management*, Vol. 2, No. 6, 2025, pp. 25–30.
6. Si, S., Zhao, H., Luo, K., Chen, G., Qi, F., Zhang, M., Chang, B., and Sun, M., "A Goal Without a Plan Is Just a Wish: Efficient and Effective Global Planner Training for Long-Horizon Agent Tasks," , 2025. URL https://arxiv.org/abs/2510.05608.
7. Jian, P., Yu, D., and Zhang, J., "Large language models know what is key visual entity: An LLM-assisted multimodal retrieval for VQA," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 10939–10956.
8. Jian, P., Yu, D., Yang, W., Ren, S., and Zhang, J., "Teaching vision-language models to ask: Resolving ambiguity in visual questions," *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 3619–3638.
9. Jian, P., Wu, J., Sun, W., Wang, C., Ren, S., and Zhang, J., "Look again, think slowly: Enhancing visual reflection in vision-language models," *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 9262–9281.
10. Si, S., Ma, W., Gao, H., Wu, Y., Lin, T.-E., Dai, Y., Li, H., Yan, R., Huang, F., and Li, Y., "SpokenWOZ: A Large-Scale Speech-Text Benchmark for Spoken Task-Oriented Dialogue Agents," *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=viktK3nO5b.
11. Qi, L., Wu, J., Choi, J. M., Phillips, C., Sengupta, R., and Goldman, D. B., "Over++: Generative Video Compositing for Layer Interaction Effects," *arXiv preprint arXiv:2512.19661*, 2025.
12. Gong, B., Qi, L., Wu, J., Fu, Z., Song, C., Jacobs, D. W., Nicholson, J., and Sengupta, R., "The Aging Multiverse: Generating Condition-Aware Facial Aging Tree via Training-Free Diffusion," *arXiv preprint arXiv:2506.21008*, 2025.
13. Qi, L., Wu, J., Gong, B., Wang, A. N., Jacobs, D. W., and Sengupta, R., "Mytimemachine: Personalized facial age transformation," *ACM Transactions on Graphics (TOG)*, Vol. 44, No. 4, 2025, pp. 1–16.
14. Chen, Y., Liu, Y., Chen, L., and Zhang, Y., "DialogSum: A Real-Life Scenario Dialogue Summarization Dataset," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 5062–5074. https://doi.org/10.18653/v1/2021.findings-acl.449.
15. Vidgen, B., Nguyen, D., Margetts, H., Rossini, P., and Tromble, R., "Introducing CAD: the Contextual Abuse Dataset," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 2289–2303. https://doi.org/10.18653/v1/2021.naacl-main.182.
16. Herzig, J., Müller, T., Krichene, S., and Eisenschlos, J., "Open Domain Question Answering over Tables via Dense Retrieval," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 512–519. https://doi.org/10.18653/v1/2021.naacl-main.43.
17. Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., and Garg, D., "Explanations for CommonsenseQA: New Dataset and Models," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

 doi:10.20944/preprints202602.0302.v1

Long Papers), Association for Computational Linguistics, 2021, pp. 3050–3065. https://doi.org/10.18653/v1/2021.acl-long.238.

18. Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., and Sun, M., "ChatDev: Communicative Agents for Software Development," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2024, pp. 15174–15186. https://doi.org/10.18653/v1/2024.acl-long.810.
19. Hao, S., Gu, Y., Ma, H., Hong, J., Wang, Z., Wang, D., and Hu, Z., "Reasoning with Language Model is Planning with World Model," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2023, pp. 8154–8173. https://doi.org/10.18653/v1/2023.emnlp-main.507.
20. Seo, A., Kang, G.-C., Park, J., and Zhang, B.-T., "Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 6167–6177. https://doi.org/10.18653/v1/2021.acl-long.481.
21. Liu, Z., and Chen, N., "Controllable Neural Dialogue Summarization with Personal Named Entity Planning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 92–106. https://doi.org/10.18653/v1/2021.emnlp-main.8.
22. He, J., Kryscinski, W., McCann, B., Rajani, N., and Xiong, C., "CTRLsum: Towards Generic Controllable Text Summarization," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2022, pp. 5879–5915. https://doi.org/10.18653/v1/2022.emnlp-main.396.
23. Fu, J., Huang, X., and Liu, P., "SpanNER: Named Entity Re-/Recognition as Span Prediction," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 7183–7195. https://doi.org/10.18653/v1/2021.acl-long.558.
24. Zhang, X., Li, R., Yu, J., Xu, Y., Li, W., and Zhang, J., "Editguard: Versatile image watermarking for tamper localization and copyright protection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11964–11974.
25. Zhang, X., Tang, Z., Xu, Z., Li, R., Xu, Y., Chen, B., Gao, F., and Zhang, J., "Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking," *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3008–3018.
26. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., and Zhang, J., "Fakeshield: Explainable image forgery detection and localization via multi-modal large language models," *arXiv preprint arXiv:2410.02761*, 2024.
27. Si, S., Zhao, H., Chen, G., Li, Y., Luo, K., Lv, C., An, K., Qi, F., Chang, B., and Sun, M., "GATEAU: Selecting Influential Samples for Long Context Alignment," *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, edited by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Association for Computational Linguistics, Suzhou, China, 2025, pp. 7380–7411. https://doi.org/10.18653/v1/2025.emnlp-main.375, URL https://aclanthology.org/2025.emnlp-main.375/.
28. Lampinen, A., Dasgupta, I., Chan, S., Mathewson, K., Tessler, M., Creswell, A., McClelland, J., Wang, J., and Hill, F., "Can language models learn from explanations in context?" *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, 2022, pp. 537–563. https://doi.org/10.18653/v1/2022.findings-emnlp.38.
29. Dai, L., Xu, Y., Ye, J., Liu, H., and Xiong, H., "Seper: Measure retrieval utility through the lens of semantic perplexity reduction," *arXiv preprint arXiv:2503.01478*, 2025.