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

NSF-HRPT: Neural Semantic Field meets Hierarchical Risk Perception Tree for Safety-Critical Scenario Assessment

Yu Zhao[1], Jiangyu Pan[1], Tao Hu[1], Ming Yin[1], Fan Yang[1], Jiangfan Liu[2], and Xiubo Liang[1,†]

***Abstract*— The ability to accurately assess and anticipate risks in safety-critical scenarios is crucial for autonomous driving systems. While existing research has made progress in collision prediction, accurately quantifying risk levels from monocular vision inputs remains challenging due to the complex dynamics of multi-agent interactions and the inherent uncertainty in real-world environments. To address these challenges, we present NSF-HRPT, a novel framework that combines learning-based perception with structured reasoning for quantitative risk assessment. Our approach features a Neural Semantic Field (NSF) that learns to model scene semantics, trajectory predictions, and probabilistic Time-to-Collision (TTC) distributions from simulation data. During inference, the pre-trained NSF serves as a prior for our Hierarchical Risk Perception Tree (HRPT), which enables efficient parallel computation and spatial reasoning about multi-agent risks. Additionally, we introduce a Sim2Real enhancement strategy that improves real-world applicability without retraining by incorporating priors from foundation models. Extensive evaluations demonstrate that our framework achieves state-of-the-art performance on synthetic benchmarks and delivers competitive, near-state-of-the-art results on real-world datasets for both TTC estimation accuracy and risk localization precision. The proposed method provides an effective solution for real-time risk awareness from monocular camera inputs.**

I. INTRODUCTION

The validation of autonomous driving systems is fundamentally benchmarked against their ability to navigate safety-critical scenarios—infrequent yet high-risk situations that define the operational boundaries of system robustness [1], [2]. This endeavor faces a significant data dichotomy. While large-scale datasets such as Waymo Open [3], [4] and nuScenes [5] offer extensive annotations for nominal driving conditions, they remain inherently sparse in critical incidents. Specialized accident datasets, including DAD [6] and CCD [7], provide valuable collision examples with Time-to-Collision (TTC) ground truth, yet their single-perspective, egocentric view lacks the comprehensive bird's-eye-view (BEV) trajectory information essential for precise multi-agent dynamic risk analysis.

To address these challenges, the research community has pursued two complementary directions: safety-critical scenario generation for offline validation [8]–[10], and safety-critical scenario perception for real-time risk identification. Within the latter, previous works have primarily focused on predicting scene-level metrics such as mean-Time-to-Accident (mTTA), which estimates the probability of future collisions [11], [12]. However, for actionable decision-making in autonomous systems, a more fundamental and valuable capability is the precise estimation of Time-to-Collision (TTC) for individual entities, providing a continuous, physically-grounded measure of collision urgency. This capability is particularly crucial for embodied intelligent systems, such as autonomous vehicles, that must make reliable safety decisions based on limited visual observations.

This task of quantitative risk assessment from monocular vision presents several fundamental challenges: First, obtaining the comprehensive ground truth data necessary for training deep dynamics models, particularly precise BEV trajectories of all traffic participants, remains difficult in real-world settings. Second, performing efficient risk assessment for numerous agents in complex scenes requires computational frameworks that can scale effectively for real-time operation. Third, bridging the visual domain gap between simulation-trained models and real-world application demands robust adaptation strategies.

To address these challenges, we propose **NSF-HRPT**, a novel framework for structured risk reasoning from sequential monocular frames. Our approach leverages the capability of high-fidelity simulation platforms like CARLA [13] to generate diverse critical scenarios with comprehensive ground truth, including precise BEV trajectories and TTC values. The framework operates through two coordinated stages: an offline training phase that learns a powerful multi-task **Neural Semantic Field (NSF)** representing scene semantics, motion, and risk distributions; and an online inference phase where the pre-trained NSF drives a **Hierarchical Risk Perception Tree (HRPT)** for parallel, structured risk reasoning. Additionally, we incorporate a **Sim2Real Enhancement** strategy that bridges the visual domain gap without retraining by integrating geometric and semantic priors from foundation models.

The main contributions of our work are as follows:

1) We propose a **Neural Semantic Field (NSF)** that learns a unified representation of scene semantics, motion, and risk from simulation data, with explicit probabilistic modeling of TTC uncertainty.
2) We design a **Hierarchical Risk Perception Tree (HRPT)**, an efficient parallel inference algorithm that uses the pre-trained NSF for structured multi-agent risk reasoning in real time.
3) We introduce a **Sim2Real Enhancement** strategy that leverages foundation models to bridge the visual domain gap without retraining, improving real-world applicability.

[1]Yu Zhao, Jiangyu Pan, Tao Hu, Ming Yin, Fan Yang, and Xiubo Liang are with Zhejiang University, Ningbo, China. (†Corresponding author: Xiubo Liang, e-mail: xiubo@zju.edu.cn)
[2]Jiangfan Liu is with Beihang University, Beijing, China.

4) We conduct extensive experiments in simulated and real-world safety-critical scenarios, showing our framework achieves state-of-the-art or near-state-of-the-art performance.

II. RELATED WORK

Safety-critical Scenario Generation focuses on creating diverse test cases for offline evaluation, largely driven by generative models. High-fidelity simulators, such as CARLA [13], provide a critical execution environment for diverse generation algorithms, enabling evaluation in virtual worlds that approximate real-world physics and traffic rules. A prominent trend is the use of diffusion models to generate rare and realistic critical events [9], create entire scenes with static road layouts and dynamic actors [14], or synthesize training data from semantic descriptions [12]. Other strategies include autoregressive models for long-term traffic flows [15] and reinforcement learning to iteratively edit scenarios for maximal risk while maintaining plausibility [8], [16]. Additionally, LLM agents are increasingly used to interpret natural language into executable simulator instructions [10] or augment standard scenarios with adversarial conditions [17]. Retrieval-augmented LLMs can further discover and memorize adversarial behaviors online [18], while knowledge-grounded LLMs reason about core threats amplified through multi-agent trajectory optimization [19].

Safety-Critical Scenarios Prediction aims to identify hazards and anticipate accidents from sensor data, supported by datasets such as CCD [7], DAD [6], A3D [20], [21], and benchmarks for rare event perception [22], [23]. Methodologies increasingly emphasize handling open-world risks, including unsupervised approaches using generative world models to identify prediction anomalies [24] and frameworks leveraging synthetic data to detect unseen objects with minimal real labels [25]. While earlier work emphasized scene-level metrics using spatio-temporal attention or graph networks [11], [26], [27], recent efforts shift toward agent-specific risk understanding [28]. This requires accurate egocentric trajectory prediction, motivating diffusion-based frameworks and visual memory models [29]–[31]. Concurrently, new methods target direct agent-level risk assessment, such as vision-language models for motion risk prediction [32] and LLMs to identify agents involved in predicted collisions [33].

III. METHOD

We propose **NSF-HRPT**, a novel framework for structured risk reasoning from monocular video in safety-critical street scenes. It operates through two coordinated components: a trainable **Neural Semantic Field (NSF)** that learns a unified representation of scene semantics, motion, and risk, and a **Hierarchical Risk Perception Tree (HRPT)** that enables efficient parallel risk assessment at inference time using the pre-trained NSF. A **Sim2Real enhancement** strategy further bridges the sim-to-real visual domain gap without retraining.

A. Neural Semantic Field (NSF)

Inspired by NEAT [34], the Neural Semantic Field (NSF) is a continuous representation that maps queries from monocular multi-frame images to semantic logits, trajectory, and TTC risk predictions. The Neural Semantic Field (NSF) comprises three core components: a Spatio-Temporal Encoder, a Neural Attention Field, and a Multi-task Decoder, as illustrated in Figure 1.

Spatio-Temporal Encoder. The Spatio-Temporal Encoder $\mathbf{E}_\phi$ extracts spatio-temporal features from an input sequence $\mathcal{X} \in \mathbb{R}^{T\times H\times W\times 3}$ representing multi-frame images. A ResNet backbone [35] first extracts a feature tensor $\mathbf{F}_{\text{feat}} \in \mathbb{R}^{T\times P\times C}$, where $P = 64$ is the number of spatial patches and C is the feature dimensionality. Learnable spatial $\mathbf{E}_s \in \mathbb{R}^{P\times C}$ and temporal $\mathbf{E}_t \in \mathbb{R}^{T\times C}$ positional embeddings are broadcast and added element-wise .

$$\mathbf{F}'_{\text{feat}}[t,p,:] = \mathbf{F}_{\text{feat}}[t,p,:] + \mathbf{E}_s[p,:] + \mathbf{E}_t[t,:]. \tag{1}$$

The tensor $\mathbf{F}'_{\text{feat}}$ is then processed by a transformer [36], which globally integrates spatio-temporal features via self-attention, enriching each patch with contextual information across space and time. The resulting tensor is flattened into the final feature matrix $\mathbf{F} \in \mathbb{R}^{(T\cdot P)\times C}$.

Neural Attention Field. The Neural Attention Field $\mathbf{a}_\psi$ implements iterative, query-dependent attention. It operates on the feature matrix $\mathbf{F}$ and a query vector $\mathbf{q}$, which can be either $\mathbf{q}_1 = (x, y, \Delta t)$ for semantic occupancy prediction or $\mathbf{q}_2 = (x, y, \Delta t, \boldsymbol{\tau})$ for behavior prediction, where (x, y) is a BEV coordinate, Δt is a future time offset, and $\boldsymbol{\tau} \in \mathbb{R}^{D_\tau}$ is a learnable entity-type embedding which represents distinct behavioral priors for different entity categories (e.g., vehicles, pedestrians). The module initializes a context vector $\mathbf{c}_0$ to the mean of $\mathbf{F}$. For N steps, an MLP a_ψ computes an attention weight vector:

$$\mathbf{a}_i = a_\psi([\mathbf{q}, \mathbf{c}_i]), \tag{2}$$

used to produce the next context vector:

$$\mathbf{c}_{i+1} = \text{softmax}(\mathbf{a}_i)^\top \mathbf{F}. \tag{3}$$

The final output is the refined context $\mathbf{c}_N \in \mathbb{R}^C$.

Multi-task Decoder. The Multi-task Decoder $\mathbf{D}_\theta$ is implemented as a MLP that processes the concatenated input $[\mathbf{q}, \mathbf{c}_N] \in \mathbb{R}^{(D_q+C)}$, where D_q represents the dimensionality of the query vector and varies depending on the query type (3 for $\mathbf{q}_1$, $3 + D_\tau$ for $\mathbf{q}_2$). The decoder produces specialized outputs through two distinct mapping functions:

Semantic Occupancy Prediction:

$$\mathbf{q}_1 : (x, y, \Delta t) \mapsto \mathbf{s}, \tag{4}$$

predicts the semantic class distribution $\mathbf{s} \in \mathbb{R}^6$ (vehicle, pedestrian, road, red light, green light, background) at the fixed coordinate (x, y) after time interval Δt.

Entity Behavior Prediction:

$$\mathbf{q}_2 : (x, y, \Delta t, \boldsymbol{\tau}) \mapsto \mathbf{o},\ (\mu, \log \sigma^2), \tag{5}$$

predicts the future state of the entity currently located at (x, y). It outputs both the trajectory offset vector $\mathbf{o} \in \mathbb{R}^2$ and

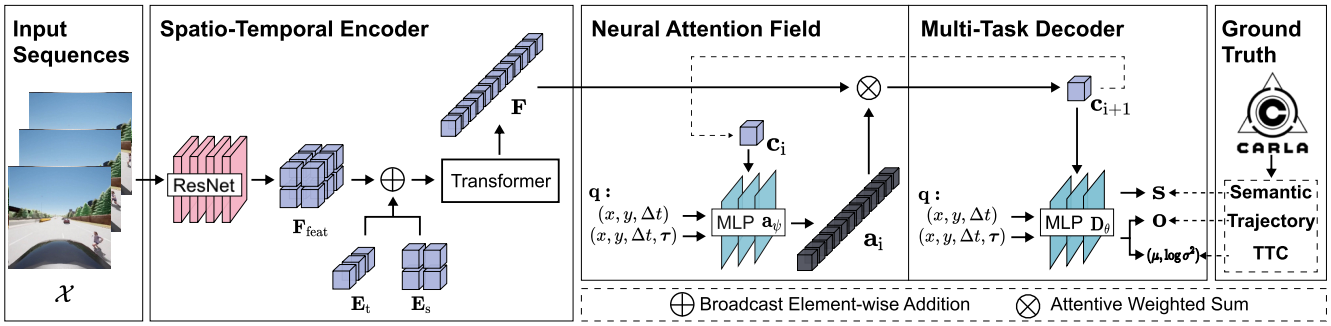

Fig. 1. **Neural Semantic Field (NSF) Architecture.** The NSF model comprises three core modules: a Spatio-Temporal Encoder for feature extraction, a Neural Attention Field for iterative query-conditioned context refinement, and a Multi-task Decoder that outputs semantic occupancy, trajectory offsets, and probabilistic TTC distributions. The illustration uses $T = 3$ input frames and $P = 4$ spatial patches for demonstration.

the TTC distribution parameters, comprising the mean $\mu \in \mathbb{R}$ and log variance $\log \sigma^2 \in \mathbb{R}$ which quantifies prediction uncertainty, after time interval Δt.

The decoder employs Conditional Batch Normalization (CBN), with parameters dynamically modulated based on the input query features to specialize the processing for each task.

Multi-task Loss. The model is trained end-to-end by minimizing the objective:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sem}}\mathcal{L}_{\text{CE}} + \lambda_{\text{traj}}\mathcal{L}_{\text{SmoothL1}} + \lambda_{\text{ttc}}\mathcal{L}_{\text{NLL}}. \tag{6}$$

The constituent losses are defined as follows. The semantic loss is the cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{K} s_k^{\text{gt}} \log \left(\frac{\exp(s_k)}{\sum_{j=1}^{K} \exp(s_j)} \right). \tag{7}$$

The offset loss uses the Smooth L1 norm:

$$\mathcal{L}_{\text{SmoothL1}} = \begin{cases} 0.5(\mathbf{o} - \mathbf{o}^{\text{gt}})^2/\beta, & \text{if } |\mathbf{o} - \mathbf{o}^{\text{gt}}| < \beta, \\ |\mathbf{o} - \mathbf{o}^{\text{gt}}| - 0.5\beta, & \text{otherwise}, \end{cases} \tag{8}$$

where $\beta = 1.0$ is a predefined threshold parameter.

The TTC loss is the negative log-likelihood under a Gaussian distribution:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{2}(TTC^{\text{gt}} - \mu)^2 \exp(-\log \sigma^2) + \frac{1}{2}\log \sigma^2. \tag{9}$$

Hyperparameters $\lambda_{\text{sem}}, \lambda_{\text{traj}}, \lambda_{\text{ttc}}$ balance the task weights.

The training process utilizes ground truth data from the CARLA simulation environment [13] to optimize the parameters $\Theta = \{\phi, \psi, \theta\}$ and learnable embeddings $\mathbf{E}_s$, $\mathbf{E}_t$, and $\boldsymbol{\tau}$.

NSF for Inference. During inference, the pre-trained NSF operates in a query-driven manner to parse scenes and predict future states. The process consists of two sequential steps:

Scene Parsing for Implicit BEV Map Initialization. A 2D query grid $\mathcal{G}$ is defined in the ego-centric coordinate frame. For each point (x, y) in $\mathcal{G}$, the semantic query $\mathbf{q}_1 = (x, y, \Delta t = 0)$ is processed by NSF to output a semantic probability distribution $\mathbf{s} \in \mathbb{R}^6$. Evaluating these queries across $\mathcal{G}$ constructs an implicit BEV semantic map $\mathcal{M}_{\text{BEV}}$, providing the foundational scene representation.

Entity-centric Behavior and Risk Prediction. Dynamic entities are identified from $\mathcal{M}_{\text{BEV}}$. For each entity i at position (x_i, y_i) with type embedding $\boldsymbol{\tau}^{(i)}$, a behavior query $\mathbf{q}_2 = (x_i, y_i, \Delta t, \boldsymbol{\tau}^{(i)})$ is formulated. The decoder outputs a trajectory offset $\mathbf{o} \in \mathbb{R}^2$ and TTC distribution parameters $(\mu, \log \sigma^2)$. The RiskValue R is computed as:

$$R = p \cdot \frac{1}{\mu + \delta}, \quad \text{where} \quad p = \frac{1}{1 + \sigma^2}, \tag{10}$$

and $\delta = 0.1$ ensures numerical stability.

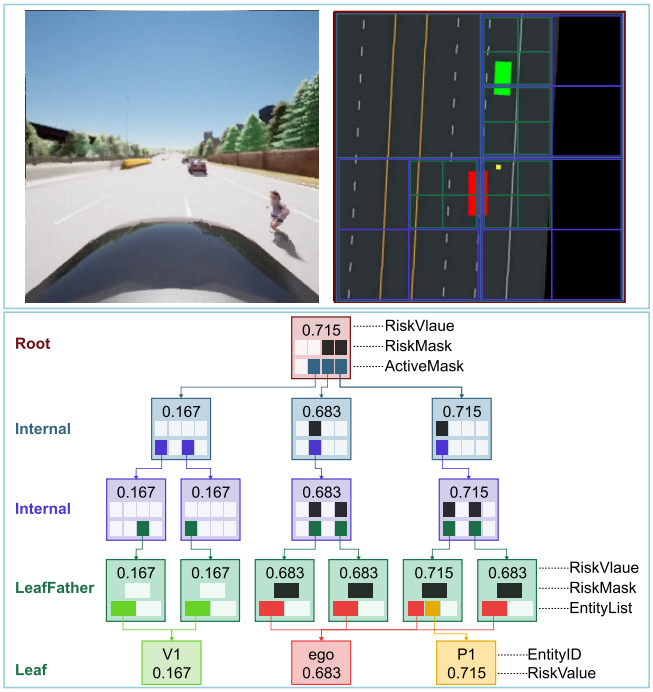

Fig. 2. **The Structure of HRPT.** The quadtree organizes spatial risk information across five hierarchical layers (Root, two Internal, LeafFather, Leaf). The BEV view illustrates the spatial partitioning, while the tree diagram details the data (RiskValue, RiskMask, ActiveMask, EntityList, EntityID) stored at each node type, facilitating efficient parallel risk aggregation.

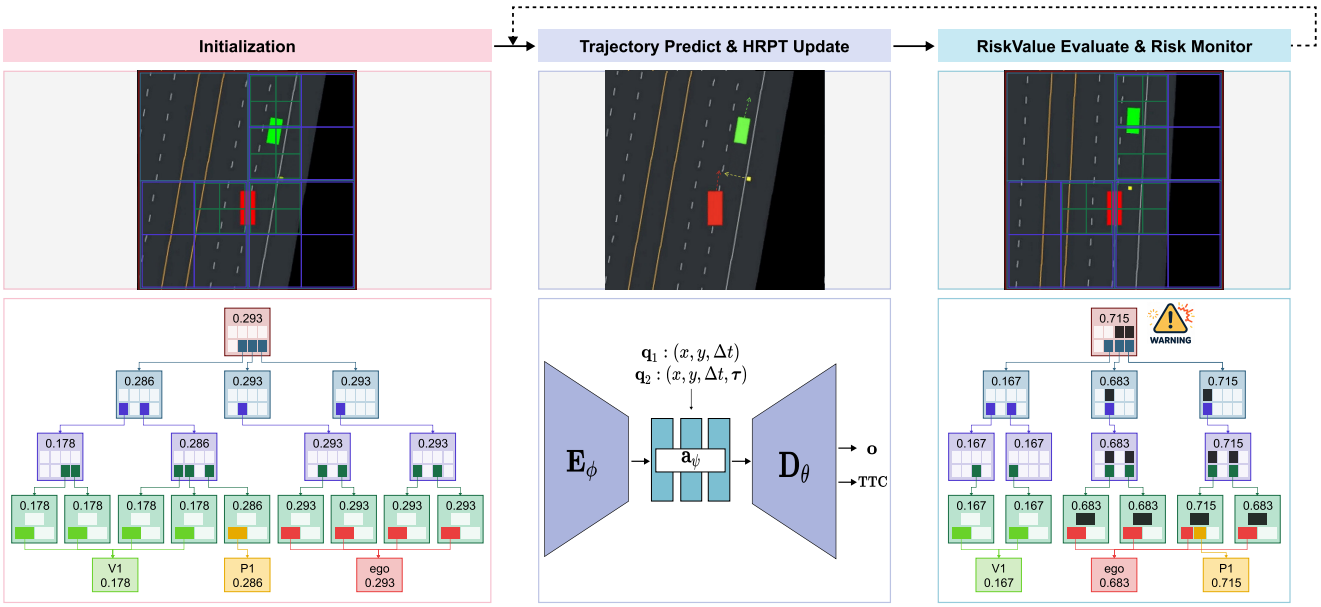

Fig. 3. **The Inference Process of HRPT.** The diagram contrasts the tree's state during initial safe conditions versus upon detecting a high-risk entity. The BEV views are for conceptual illustration; the framework operates on a logical BEV representation from the NSF without requiring explicit rendering. **NOTE:** The BEV visualization is a schematic for clarity; the actual algorithm uses an implicit, query-based spatial representation.

B. Hierarchical Risk Perception Tree (HRPT)

The **Hierarchical Risk Perception Tree (HRPT)** is a parallel quadtree structure that organizes dynamic entities and propagates risk information hierarchically. At inference, it leverages the pre-trained NSF to transform multi-agent risk evaluation into an efficient, parallel reasoning process for real-time safety assessment.

1) HRPT Structure and Node Definitions: The HRPT organizes spatial information across five hierarchical layers within the egocentric world coordinate system: (1) Root node, (2) two Internal node layers, (3) LeafFather node layer, and (4) Leaf node layer. Each node type employs specialized encoding schemes to efficiently capture and propagate risk information, with the complete structure and a detailed example of node data presented in Figure 2.

Leaf Nodes represent individual dynamic entities and store:

- EntityID: A unique identifier for each dynamic entity, which primarily encompasses vehicles and pedestrians. Vehicle entities are further categorized into the *ego vehicle* (assigned the fixed identifier `ego`) and *other vehicles* (assigned identifiers starting from `V1`). Pedestrian entities are assigned identifiers starting from `P1`.
- RiskValue $\in \mathbb{R}^+$: Risk value computed from NSF queries.
- EntityInfo: State information including coordinates (x, y) and type embedding $\boldsymbol{\tau}$.

LeafFather Nodes aggregate spatial information within fixed-size grid cells and maintain:

- RiskValue: Maximum risk among child nodes, $R_{\text{parent}} = \max(\{R_e\})$, where $\{R_e\}$ represents the set of RiskValues from all its children (Leaf nodes).
- RiskMask $\in \{0, 1\}$: Binary indicator of significant risk, computed as:

$$\text{RiskMask} = \begin{cases} 1 & \text{if } R_{\text{parent}} > R_{\text{Th}} \wedge p_{\max} > p_{\text{Th}} \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

  where R_{parent} is the aggregated risk value, $p_{\max}$ is the maximum TTC confidence probability p (as defined in Eq. (10)) among child nodes, $R_{\text{Th}} = 0.3$ is the RiskValue threshold, and $p_{\text{Th}} = 0.7$ is the confidence probability threshold.
- EntityList: Catalog of contained entities

The RiskValue of entities(leaf nodes) under each LeafFather node can be computed in parallel.

Internal & Root Nodes facilitate hierarchical risk propagation through:

- RiskValue: Maximum risk among children
- RiskMask $\in \{0, 1\}^4$: 4-bit risk indicator per child quadrant
- ActiveMask $\in \{0, 1\}^4$: 4-bit entity presence indicator

All four child quadrants at each Internal and Root level can be processed concurrently for efficient risk aggregation.

2) Initialization: The initialization process establishes the foundational state for iterative prediction. The procedure commences with BEV semantic map construction through NSF-based scene parsing of historical image sequences $\mathcal{X}$. Dynamic entities are detected from $\mathcal{M}_{\text{BEV}}$, and corresponding Leaf nodes are instantiated with initial state information.

Each entity's initial risk assessment is obtained through NSF behavior queries $\mathbf{q}_2 = (x, y, \Delta t = 0, \boldsymbol{\tau})$, yielding TTC parameters $(\mu, \log \sigma^2)$ from which the initial risk value R and confidence probability p are derived. The spatial registration phase assigns Leaf nodes to appropriate LeafFather

nodes based on coordinate information, establishing initial parent-child relationships.

The initialization culminates with bottom-up aggregation where risk values and mask encodings propagate upward through the hierarchy. LeafFather nodes compute aggregated RiskValues and binary RiskMasks, while Internal and Root nodes determine their states through maximum risk aggregation and bitwise mask operations.

3) Iterative Prediction Protocol: The HRPT executes $T = 50$ iterative prediction steps ($\delta t = 0.1s$), each comprising five distinct phases, with the complete workflow illustrated in Figure 3.

Trajectory Predict: Entity trajectories are updated through NSF queries:

$$\mathbf{p}_{\text{new}} = \mathbf{p}_{\text{current}} + \mathbf{o}, \tag{12}$$

where $\mathbf{p}_{\text{current}}$ denotes the entity's current position and $\mathbf{p}_{\text{new}}$ represents its predicted future position.

HRPT Update: Leaf node coordinates are adjusted with spatial reassignment to maintain accurate registration.

RiskValue Evaluate: Updated TTC parameters $(\mu, \log \sigma^2)$ are obtained through NSF behavior queries at new positions. Bottom-up risk aggregation recalculates all nodal RiskValues and masks according to equation (10), where R represents the RiskValue, p denotes the confidence probability of the predicted Time-to-Collision.

Risk Monitor: Root node's RiskValue and RiskMask are monitored for threshold violations, with risk source identification through hierarchical mask traversal.

This structured iterative process enables predictive, hierarchical risk assessment that anticipates potential collision scenarios and supports verifiable navigation decisions.

C. Sim2Real Enhancement

While trained exclusively on synthetic CARLA data, the NSF framework requires robust evaluation on real-world datasets. To address the visual domain gap at inference without retraining, we integrate geometric and semantic priors from pre-trained foundation models into the frozen NSF architecture.

Input Enhancement via Early-Fusion. For a real-world image sequence $\mathcal{X} \in \mathbb{R}^{T\times H\times W\times 3}$, we obtain a depth map $\mathcal{D} \in \mathbb{R}^{T\times H\times W\times 1}$ and semantic features $\mathcal{S} \in \mathbb{R}^{T\times H\times W\times C_s}$ from pre-trained models. These are concatenated and projected via a fixed linear transformation $\mathcal{P}$ to maintain original input dimensions:

$$\mathcal{X}_{\text{enhanced}} = \mathcal{P}([\mathcal{X}, \mathcal{D}, \mathcal{S}]) \in \mathbb{R}^{T\times H\times W\times 3}. \tag{13}$$

This enhanced input is fed to the frozen encoder $\mathbf{E}_\phi$.

Feature Enhancement via Mid-Fusion. The original image sequence is processed by $\mathbf{E}_\phi$ to extract initial features $\mathbf{F}_{\text{feat}} \in \mathbb{R}^{T\times P\times C}$. External models provide depth and semantic features $\mathbf{F}_d, \mathbf{F}_s \in \mathbb{R}^{T\times P\times C}$, which are aligned spatially and temporally. These are fused through a weighted summation:

$$\mathbf{F}_{\text{enhanced}} = \mathbf{F}_{\text{feat}} + \alpha_d \mathbf{F}_d + \alpha_s \mathbf{F}_s, \tag{14}$$

where $\alpha_d = 0.5, \alpha_s = 0.3$ are fixed scaling coefficients. The enhanced features $\mathbf{F}_{\text{enhanced}} \in \mathbb{R}^{T\times P\times C}$ proceed through the remaining NSF components.

Both methods require no additional training and enhance the NSF's robustness to real-world visual variations while preserving its pre-trained parameters and architectural integrity.

IV. EXPERIMENT

A. Experimental Setup

1) Simulation environment: We employ the CARLA simulator [13] with the SafeBench framework [37] for closed-loop autonomous vehicle evaluation. Our evaluation encompasses 8 critical traffic scenarios from the NHTSA Pre-Crash Typology Report [38] (including Straight Obstacle, Lane Changing, etc.), selected following the methodology in [10], [19], totaling 800 challenging scenarios (10 routes × 10 variations per scenario) for comprehensive benchmarking [19].

The BEV representation is configured with 512×512 pixel images covering a 32m×32m physical area, providing 12m rearward and 20m forward field of view from the ego vehicle through a 64-pixel viewpoint offset. We generate precise ground truth data including semantic segmentation masks, entity trajectories, and Time-to-Collision (TTC) values representing the duration to actual collision. This comprehensive dataset enables reproducible training and evaluation of safety-critical scenario detection in our risk assessment framework.

2) Real-World Datasets: Our model is validated on two standard benchmarks for accident anticipation.

DAD. The Dashcam Accident Dataset (DAD) [6] consists of 1,750 five-second, 20-FPS video clips of urban traffic, where accidents in its 620 positive samples are confined to the final 10 frames.

CCD. The larger Car Crash Dataset (CCD) [7] offers greater environmental diversity across its 4,500 videos, supported by rich metadata. Its 1,500 accident instances occur within the last two seconds and are augmented by 3,000 normal driving clips from BDD100K [39].

3) Evaluation Metrics: We evaluate our model using three key metrics: **Average Precision (AP)** , **Mean Time-to-Accident (mTTA)** , and **Accident Object Localization Accuracy (AOLA)** [33] for spatial precision.

AP measures accident detection accuracy under natural class imbalance, calculated as the area under the Precision-Recall curve.

mTTA measures the average time between the first valid warning and accident occurrence, we employ a rigorous criterion to determine the first valid warning moment t_θ. At each timestamp t_θ, our model performs efficient iterative reasoning to predict the RiskValue for future time $t_\theta + \delta t$. While the first moment the predicted RiskValue exceeds a threshold could be used, we define t_θ as a valid warning moment only when: (1) the RiskValue predicted for $t_\theta + \delta t$ first crosses the threshold, and (2) the absolute error between the predicted TTC ($\hat{T}_{t_\theta+\delta t}$) and ground-truth TTC

($T_{GT,t_\theta+\delta t}$) satisfies $|\hat{T}_{t_\theta+\delta t} - T_{GT,t_\theta+\delta t}| \leq \epsilon$ (with $\epsilon = 0.3$s based on autonomous driving requirements). The mTTA is then computed as $\frac{1}{N}\sum_{i=1}^{N}(t_{\text{collision},i} - t_{\theta,i})$ across all positive samples.

AOLA evaluates the model's precision in identifying accident-related objects, computed as the fraction of correctly predicted objects among all detected objects across frames.

4) Baselines: We compare our model against key state-of-the-art methods representing major technical trends. This includes attention-based models (DSA [6], adaLEA [40], DSTA [11]), graph-based frameworks (Ustring [7], GG [26]), and the recent LLM-based approach of Liao et al. [33] that predicts "when, where, and what".

5) Implementations: All experiments are conducted on a server equipped with an Intel(R) Core(TM) i9-14900K CPU and two NVIDIA GeForce RTX 4090 GPUs, each with 24GB of memory.

Our transformer architecture employs $L = 2$ layers with 4 parallel attention heads. For encoder selection, we evaluate ResNet-18, ResNet-34, and ResNet-50, with ResNet-34 delivering the optimal performance.

To address the absence of BEV trajectory and semantic annotations in real-world datasets (DAD and CCD), we utilize the simulated scenarios from Section IV-A.1, processed into video frames for model training. Each scenario is partitioned into training and testing subsets with a 4:1 ratio. This framework facilitates cross-domain evaluation, where models are trained on synthetic data and evaluated on real-world benchmarks. We maintain the official training/testing partitions for DAD (1,284/466) and CCD (3,600/900), using these datasets exclusively for testing purposes.

B. Comparison to State-of-the-art (SOTA)

We conduct comprehensive evaluations to benchmark our proposed framework against SOTA methods across both simulated and real-world datasets. The comparative analysis validates the effectiveness of our approach in accurate risk anticipation and demonstrates its potential for sim-to-real transfer.

Evaluation on CARLA Simulation. We first evaluate our method on the CARLA simulation scenarios described in Section IV-A.1. To ensure a fair comparison with baseline methods **DSA** [6] and **Liao et al.** [33], which require first-person view object annotations, we generate corresponding ground truth labels from the simulated scenarios and adhere to identical training/testing splits. As summarized in Table I, our method achieves superior performance across all three metrics (AP, mTTA, AOLA) in all eight base traffic scenarios, establishing a new state-of-the-art in the simulated environment. The consistent improvements demonstrate the strong scene understanding and predictive capability of our NSF-HRPT framework.

Evaluation on Real-World Datasets. To assess the sim-to-real generalization ability of our model, we directly evaluate our NSF-HRPT framework, trained exclusively on synthetic CARLA data, on the real-world DAD and CCD datasets without any fine-tuning. The results in Table II show that our base model already surpasses several early baseline methods (e.g., DSA, adaLEA), indicating a promising sim-to-real transfer capability. However, a performance gap remains compared to the current SOTA methods, which we attribute to the domain shift in visual appearance and increased scene complexity (e.g., vehicle types, background variations) in real-world data.

To bridge this domain gap, we employ the Sim2Real enhancement strategies described in Section III-C, utilizing frozen pre-trained models for semantic segmentation [41] and depth estimation [42]. Incorporating Early-Fusion (EF) and Mid-Fusion (MF) brings substantial performance gains across all metrics. Notably, with Mid-Fusion enhancement, our method achieves competitive results, matching or approaching the performance of the current SOTA on several key metrics.

C. Ablation Studies

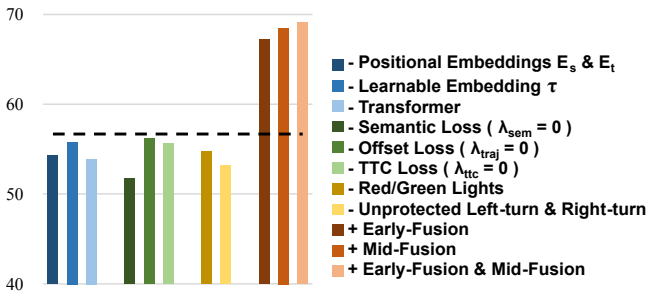

Fig. 4. Component Analysis on DAD Dataset Using AP. "- Red/Green Lights": removing the 'red light' and 'green light' classes from the semantic prediction task. "- Unprotected Left-turn & Right-turn": excluding unprotected left-turn and right-turn scenarios from the training set.

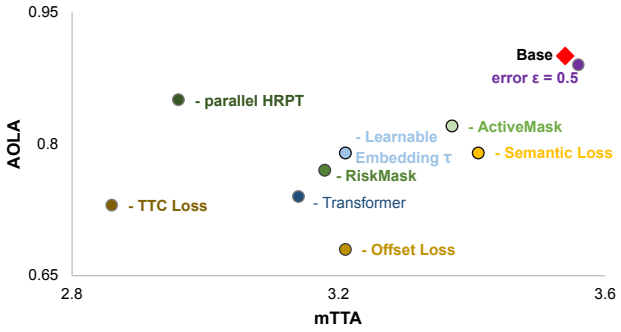

Fig. 5. Component Analysis on CARLA Using mTTA and AOLA.

We conduct systematic ablation studies to evaluate the contributions of individual components in our framework across two evaluation settings. Results are summarized in Figure 4 (DAD dataset, AP metric) and Figure 5 (CARLA simulation, mTTA and AOLA metrics).

On the DAD benchmark, we analyze core architectural components including the Positional Embeddings, Learnable

TABLE I

PERFORMANCE COMPARISON ON THE CARLA SIMULATION BENCHMARK.

Metric	Algo.	Base Traffic Scenarios								Avg.
		Straight Obstacle	Turning Obstacle	Lane Changing	Vehicle Passing	Red-light Running	Unprotected Left-turn	Right-turn	Crossing Negotiation	
AP(%)	DSA	65	75	80	75	70	70	80	75	73.8
	Liao et al.	80	**85**	**85**	**80**	80	75	80	70	79.3
	Ours	**90**	**85**	**85**	**80**	**90**	**80**	**85**	**80**	**84.3**
mTTA	DSA	2.58	2.26	2.52	2.27	2.65	2.13	2.57	2.37	2.42
	Liao et al.	3.22	3.11	3.29	2.93	3.24	3.16	3.49	3.27	3.21
	Ours	**3.68**	**3.81**	**3.49**	**3.3**	**3.77**	**3.21**	**3.54**	**3.52**	**3.54**
ALAO	Liao et al.	0.86	0.84	0.85	0.84	0.87	0.83	0.84	0.81	0.84
	Ours	**0.91**	**0.89**	**0.92**	**0.89**	**0.92**	**0.88**	**0.92**	**0.9**	**0.9**

TABLE II

PERFORMANCE COMPARISON ON THE DAD AND CCD REAL-WROLD DATASETS.

Model	DAD			CCD	
	AP(%)↑	mTTA(s)↑	AOLA↑	AP(%)↑	mTTA(s)↑
DSA [6]	48.1	1.34	-	98.7	3.08
adaLEA [40]	52.3	3.43	-	99.2	3.45
Ustring [7]	53.7	3.53	-	99.5	3.74
DSTA [11]	56.1	3.66	-	99.6	3.87
Liao et al. [33]	**69.2**	<u>4.26</u>	**0.89**	99.7	3.93
GG [26]	63.6	**4.45**	-	**99.9**	**4.96**
Ours	56.7	3.72	0.65	99.2	3.67
Ours + EF	67.2	4.13	<u>0.85</u>	<u>99.8</u>	3.87
Ours + MF	<u>68.5</u>	4.23	**0.89**	**99.9**	<u>3.95</u>

The best and second-best results per column are in **bold** and <u>underlined</u>, respectively; EF/MF denotes Early/Mid-Fusion, while "-" indicates unavailable values.

Type Embedding τ, and Transformer module. The multi-task learning objectives (Semantic, Offset, and TTC Losses) and training data configuration are examined. Our Sim2Real strategy, employing frozen pre-trained models, demonstrates effective domain adaptation and enhances robustness to real-world variations.

Complementary evaluation on CARLA simulation assesses the hierarchical reasoning mechanism of the HRPT, including its parallel structure, RiskMask, and ActiveMask components. The importance of the Learnable Embedding τ, Transformer module, and multi-task losses is validated. All components contribute significantly to the final performance in both experimental settings.

V. CONCLUSION AND FUTURE WORK

In this work, we introduce **NSF-HRPT**, a novel framework for quantitative risk assessment, built on a dual-stage paradigm unifying learning and reasoning. At its core, an **NSF** learns continuous spatiotemporal representations from simulation data, encoding scene semantics, trajectory predictions, and probabilistic Time-to-Collision (TTC) distributions. The proposed **HRPT** enables efficient parallel risk reasoning via spatial-hierarchical computation, while our **Sim2Real** strategy boosts real-world generalizability with no retraining required. Extensive experiments demonstrate state-of-the-art performance on synthetic benchmarks, and competitive accuracy for both TTC estimation and risk localization on real-world datasets.

Future research will pursue three key directions to extend this work. First, we will develop hybrid training paradigms that integrate limited real-world data with simulation data to enhance the NSF's domain adaptation capabilities, potentially through semi-supervised or adversarial training techniques. Second, we plan to incorporate large language models (LLMs) to provide higher-level risk interpretation and reasoning, enabling more intuitive and actionable warning generation for autonomous systems. Finally, we aim to construct comprehensive real-world datasets incorporating bird's-eye-view trajectory annotations and detailed risk assessments, which will serve as essential benchmarks for advancing safety-critical perception research. These efforts will collectively strengthen the framework's robustness, interpretability, and applicability, ultimately facilitating safer and more reliable deployment in real-world autonomous driving systems.

ACKNOWLEDGMENT

We acknowledge the use of large language models (LLMs) during the preparation of this manuscript. LLMs were used exclusively in an auxiliary capacity for Chinese-to-English translation and text polishing, LaTeX code generation, and assisting with proofreading to ensure logical consistency and formatting standardization.

This work was partly supported by Ningbo Youth Science and Technology Innovation Leading Talent Project (2024QL044) and Ningbo Key R&D Program (2025Z047).

REFERENCES

[1] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[2] J. Breitenstein, J.-A. Termöhlen, D. Lipinski, and T. Fingscheidt, "Corner cases for visual perception in automated driving: Some guidance on detection approaches," *arXiv preprint arXiv:2102.05897*, 2021.

[3] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[4] K. Chen, R. Ge, H. Qiu, R. Ai-Rfou, C. Qi, X. Zhou, Z. Yang, S. Ettinger, P. Sun, Z. Leng *et al.*, "Womd-lidar: Raw sensor dataset benchmark for motion forecasting," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "Nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[6] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Asian Conference on Computer Vision*, 2016.

[7] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[8] H. Liu, L. Zhang, S. K. S. Hari, and J. Zhao, "Safety-critical scenario generation via reinforcement learning based editing," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[9] C. Xu, A. Petiushko, D. Zhao, and B. Li, "Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[10] J. Zhang, C. Xu, and B. Li, "Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2024.

[11] M. M. Karim, Y. Li, R. Qin, and Z. Yin, "A dynamic spatial-temporal attention network for early anticipation of traffic accidents," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[12] Y. Guan, H. Liao, C. Wang, X. Liu, J. Zhang, and Z. Li, "World model-based end-to-end scene generation for accident anticipation in autonomous driving," *Communications Engineering*, 2025.

[13] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, 2017.

[14] J. Zhou, L. Wang, Q. Meng, and X. Wang, "Diffroad: Realistic and diverse road scenario generation for autonomous vehicle testing," *arXiv preprint arXiv:2411.09451*, 2024.

[15] L. Feng, Q. Li, Z. Peng, S. Tan, and B. Zhou, "Trafficgen: Learning to generate diverse and realistic traffic scenarios," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[16] Y. Yang, X. Lu, Z. Zhang, Z. Wu, G. Li, L. Meng, Z. Ding, and Y. Xue, "Authsim: Towards authentic and effective safety-critical scenario generation for autonomous driving tests," *arXiv preprint arXiv:2502.21100*, 2025.

[17] J. Li, T. Wang, X. Peng, J. Chen, Z. Chen, B. Li, and X. Liu, "Safety2drive: Safety-critical scenario benchmark for the evaluation of autonomous driving," *arXiv preprint arXiv:2505.13872*, 2025.

[18] Y. Mei, T. Nie, J. Sun, and Y. Tian, "Seeking to collide: Online safety-critical scenario generation for autonomous driving with retrieval augmented large language models," *arXiv preprint arXiv:2505.00972*, 2025.

[19] J. Liu, Y. Guo, F. Zhong, T. Zhang, Z. Jing, S. Liang, J. Wang, M. Zhang, A. Liu, and X. Liu, "Adversarial generation and collaborative evolution of safety-critical scenarios for autonomous vehicles," *arXiv preprint arXiv:2508.14527*, 2025.

[20] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[21] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.

[22] D. Bogdoll, I. Hamdard, L. N. Rößler, F. Geisler, M. Bayram, F. Wang, J. Imhof, M. De Campos, A. Tabarov, Y. Yang *et al.*, "Anovox: A benchmark for multimodal anomaly detection in autonomous driving," in *European Conference on Computer Vision (ECCV)*, 2024.

[23] A. Nekrasov, M. Burdorf, S. Worrall, B. Leibe, and J. S. B. Perez, "Spotting the unexpected (stu): A 3d lidar dataset for anomaly segmentation in autonomous driving," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[24] D. Bogdoll, N. Ollick, T. Joseph, S. Pavlitska, and J. M. Zöllner, "Umad: Unsupervised mask-level anomaly detection for autonomous driving," *arXiv preprint arXiv:2406.06370*, 2024.

[25] R. Chen, W. Shao, B. Zhang, S. Shi, L. Jiang, and P. Luo, "Jisam: Alleviate labeling burden and corner case problems in autonomous driving via minimal real-world data," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[26] N. Thakur, P. Gouripeddi, and B. Li, "Graph(graph): A nested graph-based framework for early accident anticipation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[27] H. Liao, H. Sun, H. Shen, C. Wang, C. Tian, K. Tam, L. Li, C. Xu, and Z. Li, "Crash: Crash recognition and anticipation system harnessing with context-aware and temporal focus attentions," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[28] J. Zhang, Y. Guan, C. Wang, H. Liao, G. Zhang, and Z. Li, "Latte: A real-time lightweight attention-based traffic accident anticipation engine," *Information Fusion*, 2025.

[29] I. Bae, Y.-J. Park, and H.-G. Jeon, "Singulartrajectory: Universal trajectory predictor using diffusion model," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[30] W. Wang, C. K. Liu, and M. Kennedy III, "Egonav: Egocentric scene-aware human trajectory prediction," *arXiv preprint arXiv:2403.19026*, 2024.

[31] A. Rasouli, "A novel benchmarking paradigm and a scale-and motion-aware model for egocentric pedestrian trajectory prediction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[32] Z. Hou, E. Ma, F. Li, Z. Lai, K. Ho, Z. Wu, L. Zhou, L. Chen, C. Sun, H. Sun *et al.*, "Drivemrp: Enhancing vision-language models with synthetic motion data for motion risk prediction," *arXiv preprint arXiv:2507.02948*, 2025.

[33] H. Liao, Y. Li, C. Wang, Y. Guan, K. Tam, C. Tian, L. Li, C. Xu, and Z. Li, "When, where, and what? a benchmark for accident anticipation and localization with large language models," in *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024.

[34] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *International Conference on Computer Vision (ICCV)*, 2021.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2017.

[37] C. Xu, W. Ding, W. Lyu, Z. Liu, S. Wang, Y. He, H. Hu, D. Zhao, and B. Li, "Safebench: A benchmarking platform for safety evaluation of autonomous vehicles," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[38] W. G. Najm, J. D. Smith, and M. Yanagisawa, "Pre-crash scenario typology for crash avoidance research," John A. Volpe National Transportation Systems Center (U.S.) (Volpe), Technical report, 2007, united States. National Highway Traffic Safety Administration.

[39] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[40] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating traffic accidents with adaptive loss and large-scale incident db," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[41] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[42] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.