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

Uncertainty-Aware Vision-based Risk Object Identification via Conformal Risk Tube Prediction

Kai-Yu Fu Yi-Ting Chen†
National Yang Ming Chiao Tung University

***Abstract*— We study object importance-based vision risk object identification (Vision-ROI), a key capability for hazard detection in intelligent driving systems. Existing approaches make deterministic decisions and ignore uncertainty, which could lead to safety-critical failures. Specifically, in ambiguous scenarios, fixed decision thresholds may cause premature or delayed risk detection and temporally unstable predictions, especially in complex scenes with multiple interacting risks. Despite these challenges, current methods lack a principled framework to model risk uncertainty jointly across space and time. We propose Conformal Risk Tube Prediction, a unified formulation that captures spatiotemporal risk uncertainty, provides coverage guarantees for true risks, and produces calibrated risk scores with uncertainty estimates. To conduct a systematic evaluation, we present a new dataset and metrics probing diverse scenario configurations with multi-risk coupling effects, which are not supported by existing datasets. We systematically analyze factors affecting uncertainty estimation, including scenario variations, per-risk category behavior, and perception error propagation. Our method delivers substantial improvements over prior approaches, enhancing vision-ROI robustness and downstream performance, such as reducing nuisance braking alerts. For more qualitative results, please visit our project webpage: https://hcis-lab.github.io/CRTP/**

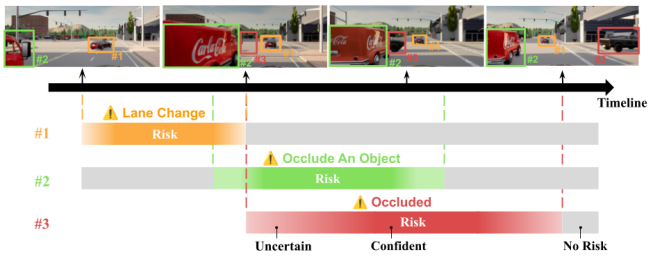

Fig. 1: **Risk Tube Prediction**. Our formulation models risk uncertainty jointly across space and time by representing potential hazards as spatiotemporal risk tubes. In this example, the green-boxed truck (#2) moves forward and may occlude part of the scene, creating the possibility of a hidden object (#3) emerging from the occluded region. Risk tubes illustrate how potential hazards evolve over time, while uncertainty is visualized through semi-transparent shading that gradually becomes opaque as observations reduce ambiguity and confidence increases.

I. INTRODUCTION

With over 1.19 million road traffic deaths annually [1], improving the safety of intelligent driving systems (IDS) has been a longstanding goal in the community. A key capability in this effort is visual risk object identification (Vision-ROI), which aims to localize potential hazards and estimate their associated risk levels or importance scores. The community has explored a variety of approaches, including collision prediction [2]–[4], trajectory prediction and collision checking [5]–[7], object importance estimation [8]–[12], human gaze prediction [13]–[16], and behavior-based prediction [17]–[22]. In this paper, we study object importance-based Vision-ROI, where risk objects are defined by human annotators' subjective assessment. This formulation directly reflects human perception of driving risk and is a common supervision signal in driving datasets.

Existing object-importance–based Vision-ROI approaches are largely deterministic, implicitly assuming reliable perception and stable scene dynamics. However, real-world traffic environments are inherently uncertain due to factors such as occlusions, sensor noise, and incomplete observations that may conceal potential hazards. In these scenarios, fixed decision thresholds can lead to temporal boundary misalignment (i.e., premature or delayed risk detection and release) and fragmented predictions that flicker between risky and non-risky states. Such behaviors are undesirable in safety-critical systems because they can produce unstable risk assessments near decision boundaries. This gap motivates the need to develop uncertainty-aware Vision-ROI systems that adapt their risk assessment to the spatiotemporal complexity of the scene and operate reliably across diverse traffic configurations. [23]–[25].

To bridge this gap, we propose **Risk Tube Prediction** (fig. 1), an uncertainty-aware formulation for Vision-ROI that jointly models uncertainty over spatial extent and temporal horizon. Instead of predicting risk for individual objects at a single time step, our formulation represents risk as a spatiotemporal tube that captures how potential hazards evolve over time. This representation is motivated by two key observations. First, risk in driving scenarios is inherently temporal: objects that are currently safe may become hazardous due to future interactions, motion patterns, or road topology. Predicting risk solely at the frame level therefore fails to capture the temporal development of hazards. Second, uncertainty often arises not only from whether an object is

† Corresponding Author

risky, but also from where and when the risk may occur. Occlusions, partial observations, and complex multi-agent interactions can lead to ambiguity in both spatial and temporal localization of risks.

To obtain reliable uncertainty estimates under our formulation, we examine existing approaches, including Bayesian methods [26], ensembles [25], Kalman filtering [27], and uncertainty embedding [28]. However, these methods often produce uncertainty estimates that do not consistently reflect predictive correctness, resulting in miscalibration [29]. Moreover, inaccurate uncertainty estimates exacerbate temporal boundary misalignment and lead to fragmented predictions, resulting in false alarms, missed risks, and unnecessary or delayed braking responses.

To this end, we present **Conformal Risk Tube Prediction**, a framework that integrates Conformal Prediction (CP) [30], [31] to construct calibrated risk tubes capturing both the spatial and temporal uncertainty of potential hazards. However, vanilla CP is insufficient in our setting because different risk categories (such as occlusion-induced or interaction-driven risks) exhibit distinct spatiotemporal characteristics that complicate calibration and reduce reliability. To address this challenge, we introduce a spatiotemporal feature-alignment loss that encourages category-consistent appearance–motion representations. We further employ category-aware conformal calibrators to ensure reliable risk score calibration and predictive coverage across heterogeneous risk types.

To evaluate our approach, we construct a **Multiple Coexisting Risks** dataset, in which multiple risk categories occur within a single scenario, a setting that is not addressed in the existing datasets [3], [4], [32]–[39]. Our dataset enables comprehensive evaluation under multi-risk conditions. We systematically analyze factors that influence uncertainty estimation, including scenario configurations, category-specific behaviors, and the propagation of perception errors, to assess the robustness of our method. Experimental results demonstrate clear improvements over prior uncertainty-modeling baselines, achieving higher calibrated risk coverage, better temporal alignment, and fewer fragmented predictions. Furthermore, we show that risk tubes enable timely yet minimal braking alerts [40], outperforming existing Vision-ROI methods. Our contributions are summarized as follows:

- We introduce an uncertainty-aware Vision-ROI formulation, **Conformal Risk Tube Prediction** that models spatiotemporal uncertainty of potential hazards more reliably than existing approaches.
- We construct a **Multiple Coexisting Risks dataset** that enables systematic evaluation of concurrent multi-risk scenarios, a setting that challenges existing Vision-ROI methods.
- Extensive experiments demonstrate that our framework improves the robustness of Vision-ROI and supports more reliable downstream responses, such as reducing nuisance braking alerts.

II. RELATED WORK

A. Visual Risk Object Identification

Visual risk object identification (Vision-ROI) is a core capability of intelligent driving systems that aim to reduce accident frequency and severity. Prior works can be categorized into four paradigms. First, objects predicted to be involved in collisions are treated as risk objects [2]–[7]. Second, risk objects are defined by human annotators' subjective assessments [8]–[12]. Third, objects fixated by human gaze are considered risk objects [13]–[16]. Fourth, objects influencing the driver's or the ego vehicle's behavior are labeled as risk objects [17]–[22].

In this work, we focus on object importance–based Vision-ROI methods, which are typically deterministic and ignore predictive uncertainty. Such overconfident outputs [41] may compromise safety [42], [43]. Existing uncertainty-aware vision methods introduce error intervals [44], predefined candidate sets [45], or variance heat maps [46], yet they lack a principled mechanism to model uncertainty that jointly evolves across space and time. We therefore propose Risk Tube Prediction, an uncertainty-aware formulation that jointly models uncertainty over spatial extent and temporal horizon. By marginalizing variability in both location and timing, it yields more robust risk estimates.

B. Uncertainty Quantification

Uncertainty quantification enables driving systems to identify when predictions are unreliable [25]. In driving applications, methods generally fall into two families: direct modeling and statistical approaches. Direct modeling approaches [27], [44], [46]–[50] include Bayesian formulations [50] treat network weights as random variables and estimates predictive uncertainty via posterior sampling or variational approximations. Kalman Filter-based method [27] that provides state uncertainty through the state covariance in a Gaussian dynamical model. Ensemble methods [46] train multiple networks with different initializations and interpret the dispersion of their predictions as uncertainty. Uncertainty Embedding [28] methods capture uncertainty by allowing each input embedding to occupy a distributional region in the latent space rather than a fixed point. However, these approaches often degrade under distribution shift [25], [26], suffer calibration errors [29], incur high computational cost [46], and yield unreliable test-time behavior.

Conformal Prediction (CP) [30], [31] is a widely used statistical inference that constructs prediction sets with coverage guarantees for the true target, while the set size provides an informative measure that dynamically reflects model uncertainty. CP has already been applied to object detection [51], multi-object tracking [52] and trajectory prediction [53], where coverage is especially valuable for safety-critical driving. We present, to our knowledge, the first application of CP to Vision-ROI. However, vanilla CP is insufficient: traffic scenes contain heterogeneous risk categories (e.g., occlusion, interaction) whose distinct characteristics confound calibration. We propose a category-aware CP framework with

TABLE I: Comparison of risk identification datasets by risk category and single- vs. multi-risk settings. Risk categories are Interaction (I), Collision (C), Obstacle (Obs), Occlusion (Occ), and Normal Driving (N).

Dataset	Risk Category					Single/Multiple
	I	C	Obs	Occ	N	
MCR (Ours)	✓	✓	✓	✓		M
RiskBench [32]	✓	✓	✓		✓	S
DADA [3]		✓				S
CTA [4]	✓	✓	✓			S
Drive Anywhere [33]				✓		S
INTERACTION [34]	✓					S
DOS [35]				✓		S
OccluRoads [36]				✓		S
SafeBench [37]		✓			✓	S
DeepAccident [38]		✓		✓		S
ROL [39]			✓		✓	S

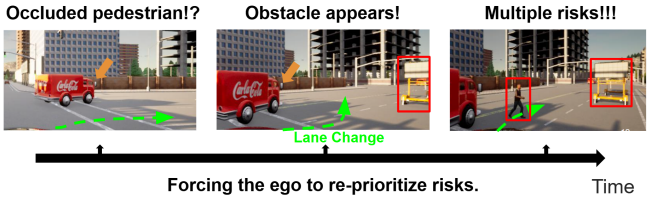

Fig. 2: The presence of multiple risks complicatedly reshapes object-ego interactions in both space and time.

a spatiotemporal feature-alignment loss, achieving improved calibration and more precise risk localization.

C. Dataset for Risk Identification

Existing risk identification datasets and benchmarks adopt different definitions of risk and are evaluated under specific risk categories. For example, prior studies have examined risk scenarios including occlusion (hidden hazards) [35], [36], [38], collision (forced crashes) [3], [4], [32], [39], interaction (dynamic social events) [32], [34], [37], and obstacle (static blockages) [32], [33], [37]. However, these scenarios rarely consider concurrent occurrence, limiting the evaluation of multi-risk coupling across categories, which ultimately complicates uncertainty estimation and risk assessment.

To address this gap, we construct the **Multiple Coexisting Risks** (MCR) dataset, integrating all four risk categories within shared scenarios. Within a single scenario, multiple risk categories can occur concurrently or in sequence. Built in CARLA [54], MCR supports scripted hazard behaviors and controllable traffic density, providing approximately 1000 scenarios for comprehensive multi-risk evaluation.

III. The Multiple Coexisting Risks Dataset

We present a scenario taxonomy and data collection pipeline. Fig. 2 shows an example of multi-risks scenarios.

Scenario Taxonomy. We design a taxonomy (Fig. 3) with static (red) and dynamic (blue) attributes to systematically collect ground-truth risk instances from multiple coexisting categories.

Static attributes define the scene environment. CARLA [54] provides towns with diverse layouts, e.g., Town02 (simple, many T-junctions) and Town05 (grid city with multi-lane intersections). Following [32], the *Map* selects the town, while *Road Topology* specifies local structures (straight roads, T-junctions, four-way intersections).

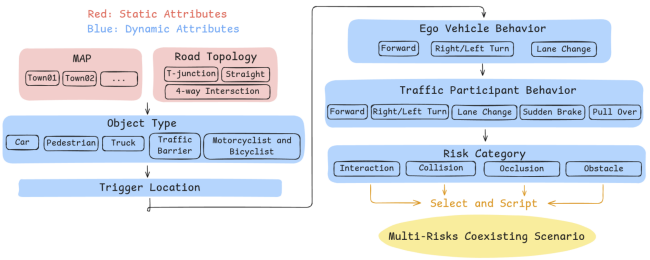

Fig. 3: The scenario taxonomy specifies attributes including road topology, risk trigger location, risk category, object type, and behavior. Given a scenario configuration, we script hazard behaviors, control traffic conditions, and further augment the scenario by varying traffic density in CARLA.

Dynamic attributes determine object behavior patterns. We first specify the *Object Type* (e.g., motorcycle, car, pedestrian) to reflect agent heterogeneity. The *Risk Trigger Location* defines where the ego vehicle interacts with traffic participants, governing when and where risk materializes. Varying this attribute generates diverse spatiotemporal configurations. We design maneuver patterns for both the *Ego Vehicle* and *Traffic Participants*, including forward motion, lane changes, turns, and sudden braking. We consider four *Risk Categories* (Interaction, Collision, Obstacle, and Occlusion), each with distinct spatiotemporal characteristics. A single scenario may include multiple categories by configuring each risk instance and composing them within the same scene.

Data Collection. We use the Scenario Runner API in CARLA [54] to script scenarios, specify trigger locations, and instantiate object types. Traffic participants follow interpolated trajectories between predefined start and end points. The ego vehicle is controlled by a rule-based planner [55] that performs obstacle avoidance, lane changes, and speed control according to the script. Additional random actors are spawned to increase environmental complexity and encourage interactions. Data are collected at 4 FPS, including RGB images and CARLA-provided metadata such as bounding boxes and velocities. Since risk evolves over time with rising and falling phases, temporal annotations are manually provided. In total, we generate approximately 1000 diverse scenarios with a balanced distribution of risk categories, enabling comprehensive evaluation under multi-risk settings.

IV. Preliminaries

A. Problem Formulation of Risk Tube Prediction

Given front-view image frames, the model outputs a **Risk Tube Prediction** $\mathcal{T}$ which encloses the set of uncertain risk objects along with the future time intervals during which each object may become hazardous.

$$\mathcal{T}_t = \left\{ \left(o, [\hat{t}_o^{\text{start}}, \hat{t}_o^{\text{end}}]\right) \ : \ o \in \hat{\mathcal{O}}_t \right\}$$

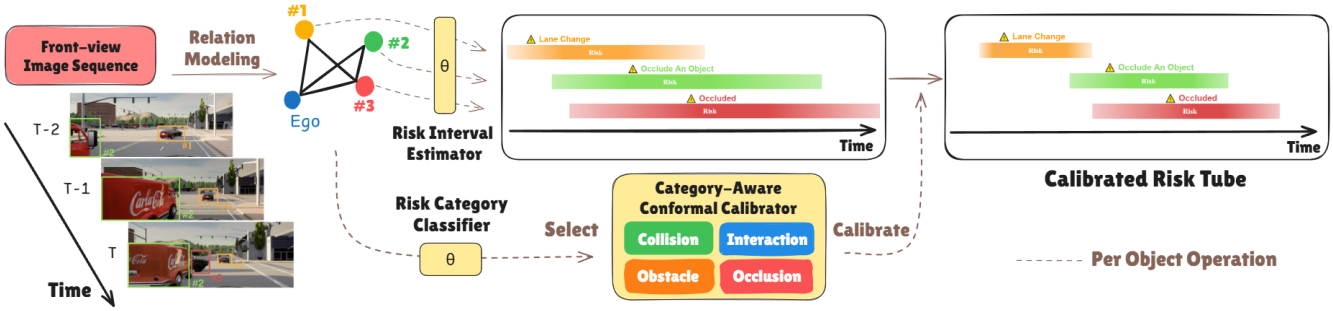

Fig. 4: Overview of the Conformal Risk Tube Prediction Framework. Given front-view images, the model performs spatiotemporal relation modeling and predicts each object's future risk interval. Then, based on a prediction of an object's risk category, the corresponding conformal calibrator is applied to calibrate its risk scores over the future interval. The calibrated risk tube provides a more precise temporal bound to fully cover the true risk interval of each hazardous object.

The predicted tube size should effectively reflect predictive uncertainty under diverse scenarios with multiple interacting risks. Ideally, a smaller tube indicates higher model confidence, precisely localizing the risk within a tighter interval, whereas higher uncertainty results in a larger tube. For evaluation, we consider an online setting, where the input consists of the past three front-view frames $I_{T-2:T}$. The tube's temporal support starts at $t = T$ and extends up to $t = T + 7$ (an 8-step horizon, $H = 8$).

B. Conformal Prediction (CP)

Consider a base predictor $f_\theta : \mathcal{X} \to \mathcal{Y}$ trained on a dataset D_{train}, for an input $x \in \mathcal{X}$, the model predicts $\hat{y} = f_\theta(x) \in \mathcal{Y}$ in the task-specific output. CP constructs a *prediction set* $C(X_{\text{test}}) \subseteq \mathcal{Y}$ for a new sample X_{test}. First, we define a nonconformity score $S(X, Y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ from the model's outputs that quantifies how inconsistent the prediction is with the ground truth. For classification, a common choice is $S(x, y) = 1 - f_\theta(x)_y$, where $f_\theta(x)_y$ denotes the predicted probability of the true label y. For regression, we usually design $S(x, y)$ as $|y - f_\theta(x)|$. Given a calibration set $D_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ of previously unseen, exchangeable pairs drawn from the same distribution as D_{train}. Compute nonconformity scores s_i for every $(X_i, Y_i) \in D_{\text{cal}}$, and sort them in ascending order. Then we take $\hat{q}_{1-\alpha}$ be the empirical $1 - \alpha$ quantile of $\{s_i\}_{i=1}^n$, which serves as the threshold indicating how much error is still acceptable. Note that $\alpha \in (0, 1)$ is the user-specified miscoverage level. For a new instance X_{test} (with unknown Y_{test} at inference time), we construct a CP set $C(X_{\text{test}}) = \{ y \in \mathcal{Y} : S(X_{\text{test}}, y) \le \hat{q}_{1-\alpha} \}$ for classification or a CP interval $\left[f_\theta(x) - \hat{q},\ f_\theta(x) + \hat{q} \right]$ for regression.

Since the core assumption of CP is that samples are *exchangeable*, the rank of the testing nonconformity score among the $n+1$ scores (n calibration plus itself) is uniformly distributed over $\{1, \ldots, n+1\}$. If we choose $k = \lceil (n+1)(1-\alpha) \rceil$ as the quantile $\hat{q}$, then the probability that the test score falls within this quantile is $\frac{k}{n+1}$, i.e., coverage is guaranteed to be at least $1 - \alpha$. Therefore, the CP sets or intervals are designed to satisfy *marginal coverage*, meaning they include the true label Y_{test}: $\mathbb{P}\{ Y_{\text{test}} \in C(X_{\text{test}}) \} \ge 1 - \alpha$. We randomly split the dataset at the scenario level into disjoint training, calibration, and test sets following an 8:1:1 ratio, ensuring that no scenario appears in more than one split.

V. METHODOLOGY

The overall framework is shown in Fig. 4. Given an input image sequence, the base model extracts per-object features and produces future-interval risk scores forming a risk tube. We train a risk-category classifier and apply a category-aware conformal predictor to calibrate the tube. A spatiotemporal feature-alignment loss aligns features of objects within the same risk category across time and space.

A. Base Model

Global ego features are extracted from RGB clips using an I3D backbone [56]. RoIAlign [57] produces per-object features from detected boxes, with phantom boxes added for occluded regions to capture hidden objects. A GCN [58] treats each object as a node and performs message passing with the ego node. Temporal relations are modeled via an LSTM-like module [59]. A linear layer predicts each object's risk score over eight timesteps, forming the deterministic risk tube, trained with binary cross-entropy on active-interval labels. Conformal prediction is applied afterward for uncertainty estimation and calibration.

B. Category-Aware Conformal Calibrator

Vanilla CP is insufficient as risk categories differ in spatiotemporal signatures, confusing calibration. We train an MLP classifier to assign each object to a category (occlusion, obstacle, interaction, collision) and maintain a dedicated conformal calibrator per category [60].

Let $f_\theta(x) \in [0, 1]^{H=8}$ denote the model's predicted 8 timesteps risk score interval. For each timestep t, the model produces $\text{pred}_t = f_\theta(x)_t \in [0, 1]$, and the ground truth label is $g_t \in \{0, 1\}$ indicating risk (1) or no risk (0). We define the nonconformity score $S_t = \left| g_t - \text{pred}_t \right|$. For each risk category c and prediction horizon step $t \in \{1, \ldots, 8\}$, we compute nonconformity scores $\{S_t^{(i)}\}_{i=1}^{n_c}$, where n_c denotes the number of calibration samples belonging to category c in the calibration set D_{cal}. Then we take the empirical

$(1-\alpha)$-quantile: $\hat{q}^{(c)}_{t,1-\alpha} = \mathrm{Quantile}_{1-\alpha}\big(\{S^{(i)}_t\}_{i=1}^{n_c}\big)$. At inference time, we define the buffer zone at time t $\mathrm{BZ}_t = \big[\hat{q}^{(c)}_{t,1-\alpha},\ 1-\hat{q}^{(c)}_{t,1-\alpha}\big]$. We calibrate the tube as follows: risk if $\mathrm{pred}_t \geq 1-\hat{q}^{(c)}_{t,1-\alpha}$, no risk if $\mathrm{pred}_t \leq \hat{q}^{(c)}_{t,1-\alpha}$, and treat $\mathrm{pred}_t \in \big(\hat{q}^{(c)}_{t,1-\alpha}, 1-\hat{q}^{(c)}_{t,1-\alpha}\big)$ as ambiguous buffer zone to mitigates oscillation near the decision boundary. We continuously update the quantile online [60], which can be interpreted as dynamically adapting the buffer zone used for determining whether an object is risky.

C. Spatiotemporal Feature Alignment

Objects in the same risk category often share appearance and motion patterns. To encourage features of such objects to align in space in a manner that is consistent with how their states evolve over time, we define a spatiotemporal alignment loss. Let $F_i(t) \in \mathbb{R}^H$ denote the latent node feature with length H of object i at time t. For pairs (i,k) that belong to the *same* risk category (with $i \neq k$), we measure their *spatial* similarity at time t via cosine similarity, and we measure each object's *temporal* similarity between t and $t{+}1$. We penalize the mismatch between the spatial similarity and the temporal similarity, averaged over a set $\mathcal{P}$ of valid triplets (t,i,k).

- Spatial Similarity:

$$\cos^{(t,i,k)}_{\mathrm{spat}} = \frac{\langle F_i(t),\, F_k(t)\rangle}{\|F_i(t)\|\,\|F_k(t)\|}, \quad \text{for } k \neq i \quad \text{and } \mathrm{risk\_type}_k = \mathrm{risk\_type}_i \tag{1}$$

- Temporal Similarity:

$$\Delta^{(t,i,k)}_{\mathrm{temp}} = \frac{\langle F_i(t),\, F_i(t{+}1)\rangle}{\|F_i(t)\|\,\|F_i(t{+}1)\|} - \frac{\langle F_k(t),\, F_k(t{+}1)\rangle}{\|F_k(t)\|\,\|F_k(t{+}1)\|} \tag{2}$$

- Alignment Loss:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{P}|}\sum_{(t,i,k)\in\mathcal{P}} \Big(\cos^{(t,i,k)}_{\mathrm{spat}} - \Delta^{(t,i,k)}_{\mathrm{temp}}\Big)^2 \tag{3}$$

VI. EXPERIMENTS

Our experiments aim to answer the following research questions (RQ). (**RQ1**) Is Conformal Risk Tube Prediction robust to spatiotemporal variations across risk categories? (**RQ2**) Does it remain robust under propagated perception errors? (**RQ3**) How does the Risk Tube benefit downstream tasks compared with other Vision-ROI methods?

A. Baselines

All baselines share the same base model (Sec. V-A) and use their estimated normalized uncertainty to construct buffer zones (Sec. V-B) for fair comparison on RQ1 and RQ2. **Rule-based:** Every object is marked *risky* at all time steps. **HD (Hard Decision):** Deterministic classifier using a fixed threshold to label each risk interval element. **BNN [44]:** Bayesian neural network estimates predictive uncertainty via posterior sampling; variance across MC samples is used as uncertainty. **KF [61]:** Kalman Filter provides state uncertainty through the state covariance; magnitude is used as uncertainty. **OCP [60]:** Online conformal prediction updates quantiles adaptively to provide coverage guarantees over time. **UE [28]:** Uncertainty Embedding maps features to a Gaussian latent (μ, σ^2) and trains with task loss plus KL penalty, capturing feature uncertainty.

For RQ3, we compare the following Vision-ROI methods using the same backbone. **Distance:** Object is risky if distance to ego is below 10m. **Collision Anticipation (CA) [10]:** This method predicts which object will be involved in a collision; we use the predicted collision score as each object's risk score. **Behavior Prediction (BP):** [17] The approach outputs the ego vehicle's current action (go/stop). When the predicted action is *stop*, the object with the highest attention score is designated as the risk object. **Trajectory Prediction (TP):** [5] We predict future 2D trajectories on image for objects and mark an object as risky if its predicted trajectory intersects the ego vehicle's path.

B. Evaluation Metrics

We first describe metrics for RQ1 and RQ2.

Coverage: The ratio of GT risk objects whose *active risk interval* is fully covered by the prediction (equivalently, the GT interval is a subset of the predicted interval).

Tube Volume (TV): Serving as an indicator of predictive uncertainty. Formally, for object o with predicted risk interval $\widehat{\mathcal{I}}_o$ in the tube, $\mathrm{TV} = \frac{1}{|\widehat{\mathcal{O}}|}\sum_{o\in\widehat{\mathcal{O}}}|\widehat{\mathcal{I}}_o|$. At a fixed coverage level, larger TV implies greater uncertainty.

Temporal Consistency (TC): Quantifies fragmented prediction. We define the number of temporal switches $\mathrm{T}(y) = \sum_{t=0}^{H-1}\mathbf{1}[y_t \neq y_{t+1}]\,, (H=8)$. Temporal consistency compares the switch counts of prediction and ground truth: $\mathrm{TC} = 1 - \frac{\left|\mathrm{T}(\widehat{\mathcal{I}}^{\mathrm{pred}}_o) - \mathrm{T}(\widehat{\mathcal{I}}^{\mathrm{gt}}_o)\right|}{H-1}$. Higher values indicating closer temporal behavior to ground truth.

Boundary Alignment (BA): Quantifies temporal boundary misalignment. We evaluate alignment near the *risk start* boundary T_s and the *risk end* boundary T_e. Let $m(t) = \mathbf{1}\{\mathrm{pred}(t) = \mathrm{gt}(t)\} \in \{0,1\}$ be the per–time-step match indicator. Let $w_\theta(t) = \exp(-|t-\theta|/\tau)$ be the penalty weights around a boundary $\theta \in \{T_s, T_e\}$, where τ controls how fast the penalty decays as t moves away from the boundary. The boundary score is the locally weighted accuracy around θ: $\mathrm{PIC}^*_\theta = 1 - \frac{\sum_t w_\theta(t)\,[1-m(t)]}{\sum_t w_\theta(t)}$. The final metric averages both sides: $\mathrm{BA} = \frac{1}{2}\big(\mathrm{PIC}^*_{T_s} + \mathrm{PIC}^*_{T_e}\big)$, where larger is better.

Risk-IOU: Let $\widehat{\mathcal{I}}$ and $\mathcal{I}^*$ be the predicted and ground-truth active risk intervals, respectively. Define the interval IoU as $\mathrm{IoU}(\widehat{\mathcal{I}}, \mathcal{I}^*) = \frac{|\widehat{\mathcal{I}} \cap \mathcal{I}^*|}{|\widehat{\mathcal{I}} \cup \mathcal{I}^*|}$. We combine it with Temporal Consistency (TC) and Boundary Alignment (BA) to obtain: $\mathrm{RiskIOU} = \mathrm{IoU}(\widehat{\mathcal{I}}, \mathcal{I}^*) \times \frac{\mathrm{TC}+\mathrm{BA}}{2}$.

We then describe metrics for RQ3.

Average Brake Counts: Measures how often the model triggers braking alerts on average.

Misaligned Brake Counts (MBC): Quantify whether the braking timing is correct. For a video clip (length=L) with predicted brake sequence $y_t \in \{0,1\}$ and ground-truth sequence $\hat{y}_t \in \{0,1\}$ over $t = 1, \ldots, L$, the misaligned brake count is the sum of *false negative brakes* and *false positive*

TABLE II: Multi-Risks Coupling Effects: results on **One Risk** and **Multi-Risks**. Higher is better for Coverage, Temporal Consistency (TC), Boundary Alignment (BA), and Risk IoU; lower is better for Tube Volume (TV). The best results are highlighted in bold, and the second are underlined.

Scenario	Method	Coverage ↑	Tube Volume ↓	TC ↑	BA ↑	Risk IoU ↑
One Risk	Rule Based	**1.000**	23.261	**0.857**	0.600	0.475
	HD	0.667	**7.432**	0.644	0.694	0.515
	BNN [44]	0.810	18.049	0.727	0.650	0.518
	KF [61]	0.778	14.560	0.716	0.720	0.529
	OCP [60]	0.801	15.253	0.703	0.751	0.549
	UE [28]	0.821	18.166	0.718	0.709	0.542
	Ours	0.851	12.988	0.734	**0.800**	**0.637**
Multi-Risks	Rule Based	**1.000**	28.844	**0.857**	0.582	0.494
	HD	0.625	**6.507**	0.642	0.661	0.505
	BNN [44]	0.787	23.269	0.693	0.602	0.508
	KF [61]	0.715	18.005	0.704	0.654	0.519
	OCP [60]	0.742	17.265	0.688	0.665	0.532
	UE [28]	0.798	18.300	0.707	0.664	0.527
	Ours	0.827	15.641	0.708	**0.752**	**0.569**

brakes: $\mathrm{MBC} = \sum_{t=1}^{L} \Big[\mathbf{1}(y_t = 0,\ \hat{y}_t = 1) + \mathbf{1}(y_t = 1,\ \hat{y}_t = 0)\Big]$.

C. Results and Discussions

RQ1: Is Conformal Risk Tube Prediction robust to spatiotemporal variations across risk categories? We compare methods under both One-Risk and Multi-Risk settings in Table II. The rule-based baseline trivially achieves Coverage = 1.0 by marking all timesteps as risky, inflating Tube Volume and yielding poor boundary alignment. HD produces the smallest tubes but fails to cover full risk intervals, resulting in low coverage and overconfident predictions. Other direct uncertainty modeling methods (BNN, KF, OCP, UE) also struggle. In contrast, our method maintains high coverage with moderate Tube Volume, achieving strong boundary alignment and the best Risk-IoU in both settings. Performance drops for all methods in the Multi-Risk setting, reflecting the increased difficulty under interacting risks. Nevertheless, our approach remains robust due to improved uncertainty modeling and category-aware calibration, enabling more precise spatiotemporal localization (Fig. 5).

We further analyze performance by risk category in Table III. While baselines improve only in specific categories, our method achieves consistently strong results across most metrics and categories, demonstrating robustness. This result indicates that category-aware conformal calibration is beneficial. All methods, including ours, perform worse in occlusion scenarios, likely because 2D phantom boxes overlap with foreground objects degrade feature quality. This suggests the need for more dedicated occlusion modeling.

RQ2: Does the Conformal Risk Tube Prediction remain robust under propagated perception errors? Table IV shows that replacing ground-truth boxes with perception-based detections degrades all methods: Tube Volume increases, while Coverage, TC, BA, and Risk-IoU decrease. Despite this, our method remains the most robust, exhibiting the smallest Tube Volume inflation and the lowest performance drops across other metrics, whereas baselines suffer substantially larger penalties. Our approach enlarges tubes only as needed to absorb detection noise while better preserving temporal fidelity. Notably, our current formulation models spatial uncertainty at the object level (identity) rather than explicitly at the bounding-box level. The performance gap under detection inputs reveals the impact of spatial uncertainty in real perception pipelines, suggesting that finer region-level uncertainty modeling could further enhance robustness.

TABLE III: Per–Risk-Category Analysis.

Category	Method	Coverage ↑	TV ↓	TC ↑	BA ↑	Risk IoU ↑
Interaction	HD	0.615	8.101	0.652	0.756	0.578
	BNN [44]	0.792	21.652	0.757	0.817	0.625
	KF [61]	0.816	18.014	0.695	0.837	0.623
	OCP [60]	0.842	17.796	0.723	**0.839**	0.655
	UE [28]	0.823	18.618	0.766	0.811	0.643
	Ours	**0.876**	14.458	**0.796**	0.826	**0.681**
Collision	HD	0.609	6.646	0.656	0.698	0.516
	BNN [44]	0.833	22.039	0.657	0.734	0.534
	KF [61]	0.825	20.014	0.695	0.792	0.585
	OCP [60]	0.847	18.699	0.732	0.816	0.610
	UE [28]	0.833	19.489	0.712	0.798	0.602
	Ours	**0.865**	14.839	**0.788**	**0.839**	**0.674**
Occlusion	HD	0.661	7.996	0.654	0.648	0.508
	BNN [44]	0.811	24.681	0.757	0.721	0.532
	KF [61]	0.813	24.016	0.695	0.747	0.513
	OCP [60]	0.791	18.025	**0.786**	0.788	0.589
	UE [28]	0.806	22.486	0.753	0.756	0.571
	Ours	**0.828**	16.807	0.766	**0.807**	**0.604**
Obstacle	HD	0.728	10.359	0.751	0.680	0.570
	BNN [44]	0.802	19.879	0.757	0.714	0.581
	KF [61]	0.785	18.450	0.719	0.796	0.628
	OCP [60]	0.821	18.014	0.695	0.764	0.595
	UE [28]	0.811	17.540	0.724	0.783	0.638
	Ours	**0.847**	15.133	**0.779**	**0.828**	**0.682**

TABLE IV: Each entry shows the change in metrics when replacing **GT** bounding boxes with **perception** bounding boxes (Δ = Detected bbox result − GT bbox result).

Scenario	Method	ΔCoverage ↑	ΔTV ↓	ΔTC ↑	ΔBA ↑	ΔRisk IoU ↑
One Risk	HD	-0.102	+4.921	-0.143	-0.137	-0.150
	BNN [44]	-0.078	+3.361	-0.170	-0.118	-0.142
	KF [61]	-0.159	+4.412	-0.180	-0.141	-0.149
	OCP [60]	-0.088	+3.759	-0.136	-0.162	-0.148
	UE [28]	-0.083	+3.518	-0.114	-0.128	-0.139
	Ours	**-0.071**	**+3.021**	**-0.094**	**-0.100**	**-0.130**
Multi-Risks	HD	-0.171	+5.940	-0.224	-0.182	-0.208
	BNN [44]	-0.134	+6.207	-0.197	-0.175	-0.189
	KF [61]	-0.237	+5.956	-0.201	-0.172	-0.183
	OCP [60]	-0.138	+5.655	-0.185	-0.161	-0.178
	UE [28]	-0.144	+5.567	-0.181	-0.164	-0.172
	Ours	**-0.104**	**+5.180**	**-0.158**	**-0.136**	**-0.148**

RQ3: How does the Risk Tube benefit downstream tasks compared with other Vision-ROI methods? An intelligent driving system must react promptly to hazards to prevent accidents. We use braking alerts as the system's response mechanism. Triggering brakes solely based on object–ego distance produces frequent nuisance alerts. Incorporating Vision-ROI allows the system to focus on truly hazardous objects and suppress spurious triggers. We evaluate multiple Vision-ROI methods (Sec. VI-A) and our Risk Tube Prediction on downstream braking using Average Brake Count and Misaligned Brake Count. As shown in Table V, combining distance proximity (e.g., < 10,m) with Vision-ROI

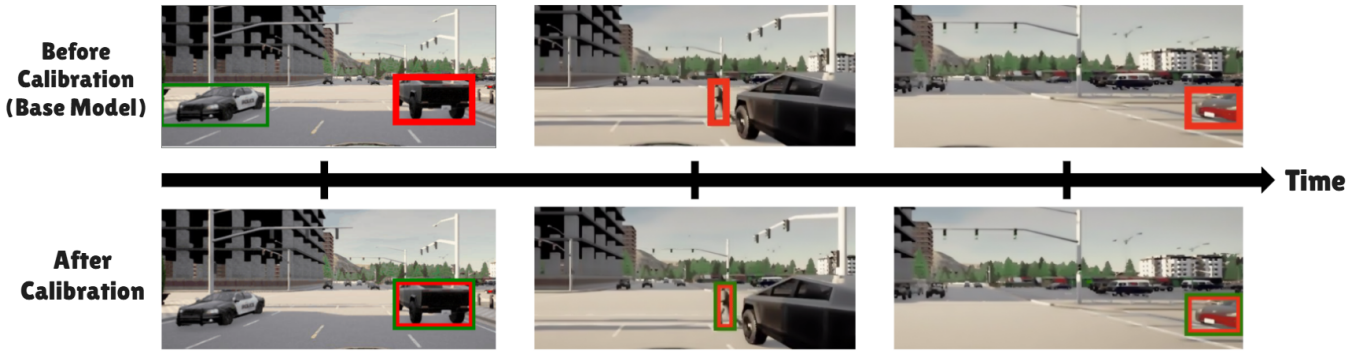

Fig. 5: Visualization of ROI results before and after calibration on a sampled scenario. All detected risk objects are shown with green bounding boxes, while ground truth risks are in red.

TABLE V: Downstream braking alerts under the criteria "Vision-ROI flags risky *and* distance $< 10\,\text{m}$." Lower is better for both metrics. GT gives the empirical lower bound on brake counts; MBC is undefined for GT (shown as "—").

Scenario	Method	Average Brake Counts ↓	Misaligned Brake Counts ↓
One Risk	Ground Truth	16.22	—
	Distance (< 10 m)	43.26	29.47
	CA [10]	32.17	21.43
	BP [17]	31.09	20.09
	TP [5]	35.48	23.43
	Risk Tube (Ours)	**23.65**	**16.34**
Multi-Risks	Ground Truth	21.61	—
	Distance (< 10 m)	54.28	37.40
	CA [10]	40.78	28.98
	BP [17]	36.87	27.08
	TP [5]	36.44	26.05
	Risk Tube (Ours)	**28.41**	**20.68**

risk substantially reduces nuisance alerts. Notably, the Risk Tube achieves the lowest brake counts and misalignment. By providing calibrated estimates of when risk begins and ends, it minimizes unnecessary interventions while closely matching ground-truth braking behavior. Overall, calibrated risk tubes serve as a principled gating mechanism, reducing nuisance braking without compromising safety.

D. Ablation Study

We justify the design choices built upon the base model, with results presented in Table VI. Adding STFA improves Coverage, TC, and BA, while reducing TV compared with the Base model, indicating that aligning object features across space and time yields more stable risk intervals and fewer unnecessary expansions. Introducing CACC on top of STFA yields further improvements, achieving higher coverage, lower TV, and better Risk IoU compared to STFA alone. These results demonstrate that category-aware calibration adjusts risk scores and uncertainty according to the characteristics of each risk category and helps maintain nominal coverage while producing more temporally aligned risk tubes.

VII. CONCLUSION

We present Conformal Risk Tube Prediction, an uncertainty-aware formulation for Visual–ROI. We demonstrate that integrating conformal prediction can address temporal boundary misalignment, fragmented predictions,

TABLE VI: Ablation study of model components. STFA denotes spatiotemporal feature alignment. CACC denotes category-aware conformal calibrators.

Method	Coverage ↑	TV ↓	TC ↑	BA ↑	Risk IoU ↑
Base	0.771	23.269	0.633	0.672	0.527
Base + STFA	0.805 (+0.034)	20.053 (-3.216)	0.665 (+0.032)	0.708 (+0.036)	0.559 (+0.032)
Base + STFA + CACC (Ours)	**0.857 (+0.052)**	**15.641 (-4.412)**	**0.708 (+0.043)**	**0.732 (+0.024)**	**0.609 (+0.050)**

and miscalibrated uncertainty present in the existing object importance-based Vision-ROI algorithms. Through our extensive experiments on the proposed Multiple Coexisting Risks dataset, we show that the proposed method is effective and robust across diverse scenario configurations. Moreover, our method provides immediate yet minimal false alarms for downstream tasks such as braking warning.

Limitations and Future Work. Performance under *Occlusion* category remains weaker than the other risk types (Table III). We plan to conduct experiments on real-world settings. Currently, risk tube predictions are made independently at each timestep; conditioning future predictions on past tubes could further improve temporal consistency. Finally, we aim to extend our approach to other safety-critical tasks, such as lane-change avoidance and junction yielding, to enhance the generalizability of the proposed framework.

Acknowledgment: The work is sponsored in part by the National Science and Technology Council under grants 113-2628-E-A49-022-, 114-2628-E-A49-007-, 114-2634-F-A49-004-, and the Ministry of Education, the Yushan Fellow Program Administrative Support Grant.

REFERENCES

[1] W. H. Organization, *Global status report on road safety 2023*. WHO, 2023.
[2] R. Herzig, E. Levi, H. Xu, H. Gao, E. Brosh, X. Wang, A. Globerson, and T. Darrell, "Spatio-Temporal Action Graph Networks," in *ICCVW*, 2019.
[3] J. Fang, D. Yan, J. Qiao, J. Xue, and H. Yu, "DADA: Driver Attention Prediction in Driving Accident Scenarios," *TITS*, 2022.
[4] T. You and B. Han, "Traffic Accident Benchmark for Causality Recognition," in *ECCV*, 2020.
[5] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions," in *CVPR*, 2019.
[6] S. Malla, B. Dariush, and C. Choi, "TITAN: Future Forecast using Action Priors," in *CVPR*, 2020.
[7] L. Neumann and A. Vedaldi, "Pedestrian and Ego-vehicle Trajectory Prediction from Monocular Camera," in *CVPR*, 2021.

[8] M. Spain and P. Perona, "Some Objects Are More Equal Than Others: Measuring and Predicting Importance," in *ECCV*, 2008.

[9] E. Ohn-Bar and M. M. Trivedi, "Are all objects equal? Deep spatio-temporal importance prediction in driving videos," *Pattern Recognition*, 2017.

[10] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun, "Agent-Centric Risk Assessment: Accident Anticipation and Risky Region Localization," in *CVPR*, 2017.

[11] M. Gao, A. Tawari, and S. Martin, "Goal-oriented Object Importance Estimation in On-road Driving Videos," in *ICRA*, 2019.

[12] J. Li, H. Gang, H. Ma, M. Tomizuka, and C. Choi, "Important Object Identification with Semi-Supervised Learning for Autonomous Driving," in *ICRA*, 2022.

[13] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving," in *CVPRW*, 2016.

[14] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting Driver Attention in Critical Situations," in *ACCV*, 2018.

[15] A. Pal, S. Mondal, and H. I. Christensen, "Looking at the right stuff: Guided semantic-gaze for autonomous driving," in *CVPR*, 2020.

[16] S. Baee, E. Pakdamanian, I. Kim, L. Feng, V. Ordonez, and L. Barnes, "MEDIRL: Predicting the Visual Attention of Drivers via Maximum Entropy Deep Inverse Reinforcement Learning," in *ICCV*, 2021.

[17] C. Li, Y. Meng, S. H. Chan, and Y.-T. Chen, "Learning 3D-aware Egocentric Spatial-Temporal Interaction via Graph Convolutional Networks," in *ICRA*, 2020.

[18] C. Li, S. H. Chan, and Y.-T. Chen, "Who Make Drivers Stop? Towards Driver-centric Risk Assessment: Risk Object Identification via Causal Inference," 2020.

[19] P. Gupta, A. Biswas, H. Admoni, and D. Held, "Object Importance Estimation Using Counterfactual Reasoning for Intelligent Driving," *RA-L*, 2024.

[20] Z. Xiao, A. Yuille, and Y.-T. Chen, "Learning Road Scene-level Representations via Semantic Region Prediction," *arXiv preprint arXiv:2301.00714*, 2023.

[21] C. Li, S. H. Chan, and Y.-T. Chen, "DROID: Driver-Centric Risk Object Identification," *TPAMI*, 2023.

[22] P.-Y. Pao, S.-W. Lu, Z.-Y. Lu, and Y.-T. Chen, "Potential Field as Scene Affordance for Behavior Change-Based Visual Risk Object Identification," in *ICRA*, 2025.

[23] ISO, *ISO 21448:2022 road vehicles — safety of the intended functionality*. ISO, 2022.

[24] UNECE, *New Assessment/Test Method for Automated Driving (NATM) Guidelines for Validating Automated Driving System (ADS)*. UNECE, 2023.

[25] K. Wang, C. Shen, X. Li, and J. Lu, "Uncertainty Quantification for Safe and Reliable Autonomous Vehicles: A Review of Methods and Applications," *ITS*, 2025.

[26] R. D. de León Torres, M. Molina, and P. Campoy, "Survey of Bayesian Networks Applications to Intelligent Autonomous Vehicles," *arXiv preprint arXiv:1901.05517*, 2019.

[27] I. Klein, G. Revach, N. Shlezinger, J. E. Mehr, R. J. G. van Sloun, and Y. C. Eldar, "Uncertainty in Data-Driven Kalman Filtering for Partially Known State-Space Models," in *ICASSP*, 2022.

[28] S. J. Oh, K. Murphy, J. Pan, J. Roth, F. Schroff, and A. Gallagher, "Modeling Uncertainty with Hedged Instance Embedding," in *ICLR*, 2019.

[29] B. Ghoshal and A. Tucker, "On Calibrated Model Uncertainty in Deep Learning," in *ECML PKDD 2020*, 2022.

[30] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of machine learning research*, 2007.

[31] A. N. Angelopoulos, E. J. Candes, and R. J. Tibshirani, "Conformal PID Control for Time Series Prediction," *NeurIPS*, 2023.

[32] C.-H. Kung, C.-C. Yang, P.-Y. Pao, S.-W. Lu, P.-L. Chen, H.-C. Lu, and Y.-T. Chen, "RiskBench: A Scenario-based Benchmark for Risk Identification," in *ICRA*, 2024.

[33] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive Anywhere: Generalizable End-to-end Autonomous Driving with Multi-modal Foundation Models," in *ICRA*, 2023.

[34] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv preprint arXiv:1910.03088*, 2019.

[35] H. Shao, L. Wang, R. Chen, S. L. Waslander, H. Li, and Y. Liu, "ReasonNet: End-to-End Driving with Temporal and Global Reasoning," in *CVPR*, 2023.

[36] M. C. A. Nataly, M. S. Sergio, S. Carlota, and S. M. Angel, "Prediction of Occluded Pedestrians in Road Scenes using Human-like Reasoning: Insights from the OccluRoads Dataset," in *IV*, 2024.

[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 *NeurIPS*, 2022.

[38] T. Wang, S. Kim, W. Ji, E. Xie, C. Ge, J. Chen, Z. Li, and P. Luo, "DeepAccident: A Motion and Accident Prediction Benchmark for V2X Autonomous Driving," in *AAAI*, 2023.

[39] M. M. Karim, Z. Yin, and R. Qin, "An Attention-guided Multistream Feature Fusion Network for Early Localization of Risky Traffic Agents in Driving Videos," *IV*, 2023.

[40] J. Campbell, J. Brown, J. Graving, C. Richard, M. Lichty, T. Sanquist, L. P. Bacon, R. Woods, H. Li, D. N. Williams, Justin, and Morgan, *Human Factors Design Guidance for Driver-Vehicle Interfaces*. NHTSA, 2016.

[41] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, "A Survey of Uncertainty in Deep Neural Networks," *Artificial intelligence review*, 2022.

[42] S. Khaitan, Q. Lin, and J. M. Dolan, "Safe Planning and Control Under Uncertainty for Self-Driving," *TVT*, 2020.

[43] X. Tang, K. Yang, H. Wang, J. Wu, Y. Qin, W. Yu, and D. Cao, "Prediction-Uncertainty-Aware Decision-Making for Autonomous Vehicles," *IV*, 2022.

[44] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning," in *ACMMM*, 2020.

[45] X. Huang, S. McGill, B. C. Williams, L. Fletcher, and G. Rosman, "Uncertainty-Aware Driver Trajectory Prediction at Urban Intersections," in *ICRA*, 2019.

[46] C. J. Holder and M. Shafique, "Efficient Uncertainty Estimation in Semantic Segmentation via Distillation," in *ICCVW*, 2021.

[47] M. Pitropov, C. Huang, V. Abdelzad, K. Czarnecki, and S. Waslander, "LiDAR-MIMO: Efficient Uncertainty Estimation for LiDAR-based 3D Object Detection," in *IV*, 2022.

[48] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding," in *BMVC*, 2016.

[49] C. Tang, J. Chen, and M. Tomizuka, "Adaptive Probabilistic Vehicle Trajectory Prediction Through Physically Feasible Bayesian Recurrent Neural Network," in *ICRA*, 2019.

[50] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users," *CIM*, 2022.

[51] A. Timans, C.-N. Straehle, K. Sakmann, and E. Nalisnick, "Adaptive Bounding Box Uncertainties via Two-Step Conformal Prediction," in *ECCV*, 2024.

[52] S. Su, S. Han, Y. Li, Z. Zhang, C. Feng, C. Ding, and F. Miao, "Collaborative Multi-Object Tracking with Conformal Uncertainty Propagation," *RA-L*, 2024.

[53] X. Chen, R. Bhadani, and L. Head, "Conformal Trajectory Prediction with Multi-View Data Integration in Cooperative Driving," *arXiv preprint arXiv:2408.00374*, 2025.

[54] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *CoRL*, 2017.

[55] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-Enhanced Autonomous Driving Using Interpretable Sensor Fusion Transformer," in *CoRL*, 2022.

[56] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *CVPR*, 2018.

[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2018.

[58] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *ICLR*, 2017.

[59] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.

[60] A. Bhatnagar, H. Wang, C. Xiong, and Y. Bai, "Improved Online Conformal Prediction via Strongly Adaptive Online Learning," in *ICML*, 2023.

[61] Q. Li, R. Li, K. Ji, and W. Dai, "Kalman Filter and Its Application," in *ICINIS*, 2015.