

RiskBench: A Scenario-based Benchmark for Risk Identification

Chi-Hsi Kung¹ Chieh-Chi Yang¹ Pang-Yuan Pao¹ Shu-Wei Lu¹ Pin-Lun Chen¹
 Hsin-Cheng Lu² Yi-Ting Chen^{1‡}

Abstract—Intelligent driving systems aim to achieve a zero-collision mobility experience, requiring interdisciplinary efforts to enhance safety performance. This work focuses on risk identification, the process of identifying and analyzing risks stemming from dynamic traffic participants and unexpected events. While significant advances have been made in the community, the current evaluation of different risk identification algorithms uses independent datasets, leading to difficulty in direct comparison and hindering collective progress toward safety performance enhancement. To address this limitation, we introduce RiskBench, a large-scale scenario-based benchmark for risk identification. We design a scenario taxonomy and augmentation pipeline to enable a systematic collection of ground truth risks under diverse scenarios. We assess the ability of ten algorithms to (1) detect and locate risks, (2) anticipate risks, and (3) facilitate decision-making. We conduct extensive experiments and summarize future research on risk identification. Our aim is to encourage collaborative endeavors in achieving a society with zero collisions. We have made our dataset and benchmark toolkit publicly at this [project webpage](#).

I. INTRODUCTION

In pursuit of a collision-free mobility society, numerous industries and academic institutions have dedicated their efforts to enhancing the safety capabilities of Intelligent Driving Systems (IDS). This paper primarily centers on risk identification, the process of identifying events that could influence the decision-making of IDS. Specifically, we delve into the evaluation of algorithms' ability to identify and assess risks stemming from dynamic traffic participants and unexpected events, such as collisions and construction zones.

Predicting potential hazards is an inherently challenging aspect of risk identification. A variety of risk identification algorithms have been attempted, such as collision anticipation [1], [2], [3], [4], [5], trajectory prediction and collision checking [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], object importance [16], [17], [18], [19], [20], human gaze prediction [21], [22], [23], [24], [25], driver behavior change prediction [26], [27]. Although significant progress has been made, current evaluation strategies utilize independent datasets, which pose challenges in terms of making direct comparisons and impede collective efforts aimed at improving safety performance. For instance, collision anticipation methods are evaluated on incidentDB [2]

¹ C.-H. Kung, C.-C. Yang, S.-W. Lu, P.-Y. Pao, P.-L. Chen, and Y.-T. Chen is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan.

² H.-C. Lu is with the Department of Computer Science, National Taiwan University, Taipei, Taiwan.

[‡] Corresponding Author.

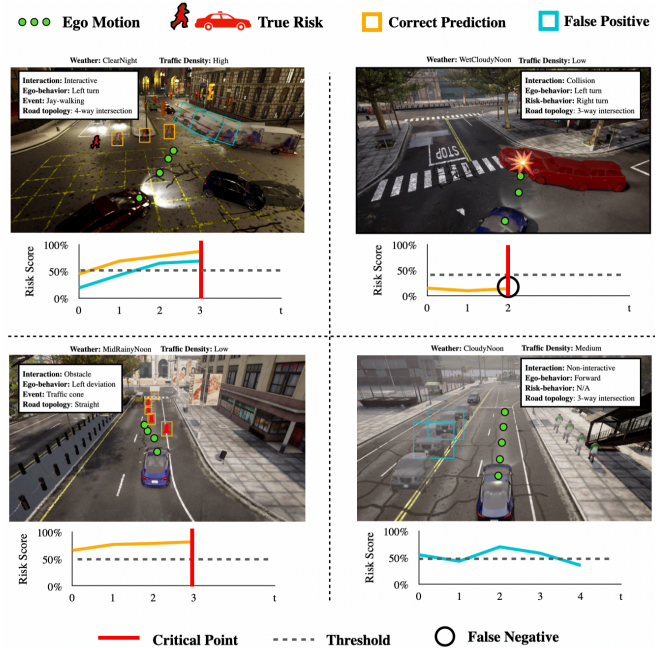


Fig. 1: Scenario-based Evaluation for Risk Identification. We establish four distinct interaction types to cover various risk definitions explored in the community. RiskBench evaluates an algorithm's ability to identify risks stemming from dynamic traffic participants and unexpected events based on localization, anticipation, and planning awareness.

or CTA datasets [5], whereas trajectory prediction-based methods [10], [13], [28], [29], [30], [31] are evaluated on pre-processed trajectory datasets [32], [33], [34]. Additionally, object importance or risk region prediction-based methods use separate datasets [16], [17], [19], [20]. There is currently an absence of a standardized benchmark for comprehensive evaluation of different risk identification algorithms.

We present RiskBench, a benchmark designed for risk identification, with a specific focus on assessing risks stemming from dynamic traffic participants (vehicles, pedestrians, and motorcyclists) and unexpected events (collisions and construction zones) through a scenario-based methodology. This approach is prevalent in the automotive field [35], [36], [37], enabling comprehensive scenario-based performance analysis [38], [39]. We design a scenario taxonomy that enables the methodical collection of ground truth risks. The taxonomy comprises four interaction types: *Interactive*: yielding to dynamic risks [40], [39], [41], [61], *Collision*: crashing scenario [1], [4], [42], *Obstacle*: interacting with

static elements [40], [32], and *Non-interactive*: normal driving [43], [44], [45], [32], aiming to cover diverse risks explored in the community. The four interaction types are depicted in Fig. 1.

We construct the RiskBench dataset in the CARLA simulator [52], resulting in **6916** scenarios from 4 interaction types, 6 actor-behavior categories, 14 different maps, 237 unique road structures, 4 traffic violations, 3 traffic densities, and 21 weather and lighting conditions. Simulation-based benchmarking has been widely adopted in both the research community [52], [48], [39], [53], [54], [55], [56], [57], [47] and industry (e.g., Waymo [58] and Waabi [59]). For instance, SafeBench [39] generates safety-critical scenarios in the CARLA simulator for evaluating the performance of driving models. Similarly, DOS [47] also assesses driving models within the CARLA environment, placing particular emphasis on the analysis of various occlusion events.

Riskbench mitigates the following challenges when it comes to benchmarking risk identification algorithms in real-world settings: First, simulation serves as a crucial preliminary phase preceding costly real-world testing and deployment. It aids researchers and developers in pinpointing critical flaws and vulnerabilities. Second, current large-scale real-world datasets are collected naturalistically [32], [44], [60], [33]. They lack the capacity to regulate the occurrence frequency of various scenarios [61]. This limitation hampers the thorough scenario-based performance analysis. Third, static real-world datasets are inadequate for supporting interactive evaluation, which is essential for validating the impact of risk identification on subsequent planning processes.

We devise three metrics to assess the performance of a risk identification algorithm, drawing inspiration from the ISO26262 Hazard and Risk Assessment Methodology [62]. Specifically, we evaluate algorithm’s ability to (1) localize risks, (2) anticipate risks, and (3) facilitate decision-making. The first two metrics are well-established in collision detection and anticipation literature [1], [2], [42], [63]. Our contribution lies in the third metric, which is a planning-aware metric called Influenced Ratio (IR). The key insight is that if the identified risks are accurate, the planner should possess the ability to generate a collision-free path. On the contrary, if a model makes an inaccurate identification, the planner could potentially guide the ego vehicle into a situation with a higher risk of collision or near-collision.

We benchmark ten risk identification algorithms, including rule-based, trajectory prediction and collision checking-based, collision anticipation, and behavior prediction-based algorithms. Our scenario-based performance analysis shows that existing risk identification algorithms demand considerable endeavor. We identify two crucial perspectives for future research: (1) ensuring temporal consistency of risk identification and (2) improving object-centric representation learning (Please see Video Submission for details).

Our contributions are summarized as follows:

- We introduce **RiskBench**, a scenario-based benchmark of identifying risks induced by traffic participants and unexpected events, alongside a scenario taxonomy for

methodical collection of ground truth risks.

- We develop a scenario-based benchmark to systematically evaluate the ability of risk identification algorithms to (1) localize risks, (2) anticipate risks, and (3) facilitate decision-making.
- We conduct extensive scenario-based analysis on ten algorithms and discuss the limitations of existing risk identification algorithms.

II. RELATED WORK

Traffic Scene Datasets and Benchmarks. Various datasets and benchmarks have been released in recent years for object detection and tracking [45], [64], [32], and activity recognition and intention prediction of traffic participants [65], [60], [66], [14], [67]. We observe substantial progress because of the standardized benchmarks. There are independent efforts tailored for risk identification, such as collision anticipation [1], [2], [5], [4], risk region prediction [3], object importance [16], [17], [19], trajectory prediction and collision checking [33], [49], human gaze prediction [21], [22], [24], [25], risk localization and captioning [50], behavior change detection and causal inference [60]. Additionally, scenario-based safety assessment benchmarks are explored [39], [40]. Table I provides a summary of datasets and benchmarks that are relevant to risk identification and scenario-based performance analysis.

Our work complements existing datasets and benchmarks by constructing a common benchmark that entails different definitions of risks explored in the community. In addition, we propose a comprehensive evaluation protocol, including the planning-aware metric, which has been explored. SafeBench [39] is the most relevant work among existing sets. The authors apply safety-critical scenario generation algorithms to generate challenging testing cases to evaluate the safety and robustness of autonomous driving algorithms. In pursuit of collaborative efforts, we believe there is a great synergy to combine the initiatives described in [40], [39].

Planning-aware Metrics. The drawback of prevalent metrics (e.g., precision/recall) used in perception algorithms is that they penalize all incorrect detection equally, without considering the downstream tasks. Recently, planning-aware metrics have been explored for perception and trajectory forecasting in the community [68], [69]. Philion et al., [68] design the planning-aware metric for perception through training a robust planner, using the nuScenes dataset [32], to plan a driving trajectory based on its semantic observations (detection). Recently, Renz et al., [69] proposed the Relative Filtered Driving Score to quantify the ability of planners to identify relevant objects. The concept is highly relevant to the proposed planning-aware metric, IR. We design IR to evaluate the ability of risk identification algorithms instead of planners. It will be our future work to consider the two tasks jointly as they go hand in hand.

III. THE RISKBENCH DATASET

This section presents the scenario taxonomy, and the data collection and augmentation pipeline.

TABLE I: Comparison with traffic scene datasets and benchmarks relevant to risk identification and scenario-based analysis. In scenario types, I, C, O, and N represent *Interactive*, *Collision*, *Obstacle*, and *Non-interactive*, respectively. For road topology, F, T, S, R, and X represent *4-way intersection*, *T-intersection*, *Straight*, *Roundabout*, and *Not Categorized*, respectively. For Risk Identification metrics, L, A, C, and P represent Risk Localization, Risk Anticipation, Temporal Consistency, and Planning Awareness, respectively. The column, Sensors, involves V and T, which stand for Vision and Trajectory, respectively.

Dataset	# of Categorized Scenarios	Scenario Types				Road Topology				Risk Id. Metrics				Sensors	
		I	C	O	N	F	T	S	R	L	A	C	P	V	T
RiskBench	6916	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CARLA Challenge [40]	23	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
SafeBench [39]	8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Longest6 [46]	7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DOS [47]	100	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GTACrash [48]	N/A	✓	✓	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓
nuPlans [49]	75	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
DAD [1]	4	✓	✓	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓
DADA [4]	2,000	✓	✓	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓
CTA [42]	18	✓	✓	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓
BDD-A [22]	N/A	✓	✓	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓
DRAMA [50]	17785	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TrafficNet [51]	30	✓	✓	✓	✓	X	X	X	X	✓	✓	✓	✓	✓	✓

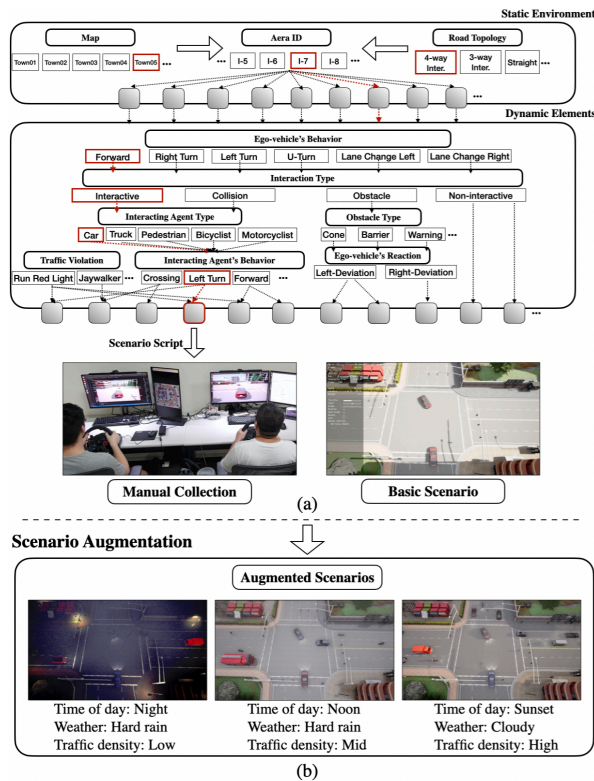


Fig. 2: **Scenario Collection Pipeline.** The scenario taxonomy is designed to enable a systematic collection of ground truth risks induced by dynamic traffic participants and unexpected events. The taxonomy includes various attributes such as road topology, scenario types, ego vehicle behavior, and traffic participants' behavior. From this taxonomy, if a scenario script is set, two human subjects can act accordingly. To form the final scenario dataset, we augment the collected scenario by changing attributes, including time of day, weather conditions, and traffic density.

A. The Taxonomy of Scenarios

The motivation for designing a scenario taxonomy is to enable a systematic collection of **ground truth risks** induced

by dynamic traffic participants and unexpected events.

Static Environment contains the following attributes, i.e., *Map*, *Area ID*, and *Road Topology*. The attribute *Map* has 8 towns constructed in the CARLA ecosystem. We assign all road segments and junctions with a *Area ID*. In total, we have unique 231 IDs. *Road topology* denotes the road structure of each *Area ID*, e.g., 4-way intersection. Note that, we reconstruct real-world maps from CAROM-Air [70] in the CARLA simulator because the simulator provides a limited number of roundabouts. Note that, we select 6 out of 13 semantic maps provided by CAROM-Air.

Dynamic Elements involve the following attributes, i.e., *Interaction Type*, *Interacting Agent Type*, *Interacting Agent's Behavior*, and *Ego-vehicle's Behavior*. To collect a scenario, we first set a value for *Ego-vehicle's Behavior* then determine the value of *Interaction Type*. If the value of *Interaction Type* is *Interactive* and *Collision*, we first set a value for *Interacting Agent Type* and *Interacting Agent's Behavior*. Note that the assignment to *Interacting Agent Type* and *Interacting Agent's Behavior* define the type and behavior of the **ground truth risk** in a scenario. Then we set a value for *Interacting Agent Behavior*. If the value of *Interaction Type* is *Obstacle*, we first set a value for *Obstacle Type*. Note that the assignment to *Obstacle Type* defines the type of the **ground truth risk**. Then we set a value for *Ego-vehicle's reaction*. Please refer to our website for the full description of the values of each attribute.

B. Data Collection

Fig. 2 depicts the overall pipeline. We script a scenario by specifying attribute values. For instance, we specify the values of *Map*, *Area ID*, *Road Topology*, *Ego-vehicle's Behavior*, *Interaction Type*, *Interacting Agent Type*, and *Interacting Agent's Behavior* as *Town05*, *I-7*, *4-way intersection*, *Forward*, *Interactive*, *Car*, and *Left Turn*, respectively. To authenticate interactions and address the rarity of scenarios, we enlist two human subjects for script execution. One participant assumes the role of the ego vehicle, while the other takes on the role of the interacting agent, i.e., acting

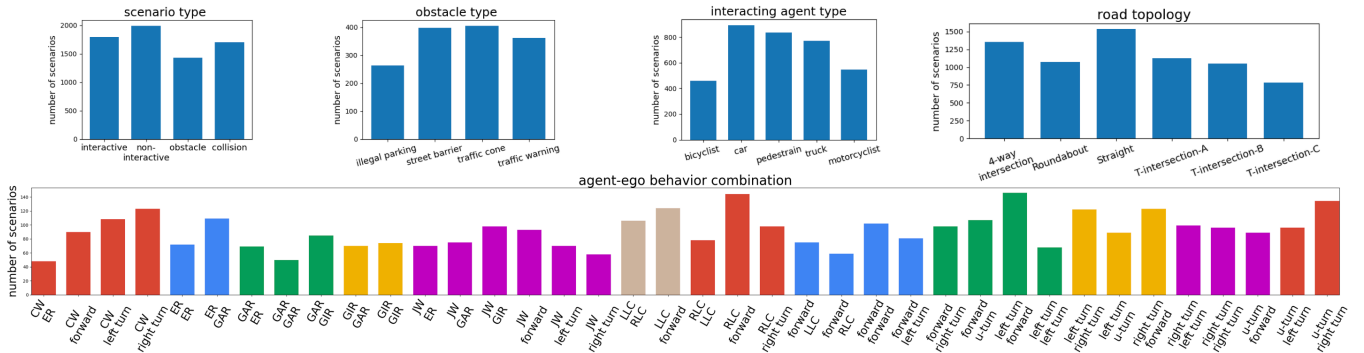


Fig. 3: **The RiskBench Dataset Statistics.** We denote *Crosswalking* as CW, *Jay-walking* as JW, *Left Lane Change* as LLC, *Right Lane Change* as RLC, *Go into Roundabout* as GIR, *Exit Roundabout* as ER, and *Go Around Roundabout* as GAR.

as the ground truth risk.

A Basic Scenario collection begins with both subjects’ movement and concludes upon interaction completion or the ego vehicle’s completion of its assigned behavior. We collect the trajectories of both agents and the actors’ blueprints, defined by the CARLA simulator [52].

C. Scenario Augmentation

To boost interaction complexity and visual challenges, we create a scenario augmentation pipeline. Specifically, we replay a **Basic Scenario** in CARLA using the recorded trajectories of the ego vehicle and interacting agent. We apply three augmentation strategies, i.e., random actor injection, time-of-day, and weather condition adjustment, as illustrated in Fig. 2 (b). For random actor injection, we spawn random actors around the ego vehicle as the distraction and ensure their trajectories will not interact with the ego vehicle. We set the level of traffic density to be *low*, *medium*, or *high*. For time-of-day and weather condition adjustment, we uniformly sample a value for time-of-day (i.e., noon, sunset, or night). We randomly select the cloudiness and rain intensity. Our data collection pipeline enables the creation of a well-organized dataset and the statistics are reported in Fig. 3.

IV. RISK IDENTIFICATION BENCHMARK

We present risk identification algorithms and implementation details, metrics, and benchmarking results.

A. Risk Identification Baselines

The baselines take a sequence of historical data as input and output a risk score for each road user (e.g., vehicle or pedestrian) or an unexpected event (e.g., collision or construction zone). We consider a road user or an unexpected event risk if the score exceeds a predefined threshold. We implement 10 risk identification algorithms and categorize them into the following four types.

- 1. Rule-based. Random:** We randomly pick an object. **Range:** We pick any objects within a predefined distance.
- 2. Trajectory Prediction and Collision Checking.** We follow the protocol of [9] to identify risk. Specifically, if there is an overlap between the ego vehicle’s and a road user’s future trajectories, the road user is the predicted risk.

We select **Kalman filter** [71], adversarial network **Social-GAN** [72], memory augmented network **MANTRA** [28], and recent query-based **QCNet** [31], which is 2nd place winner of the Argoverse challenges [73], [74].

3. Collision Anticipation. DSA [1] and **RRL** [75] identify risks via collision prediction with exponential loss [76], encouraging early anticipation of risk. Note that **RRL** uses ground truth risk supervision in its training.

4. Behavior Prediction-based. The key insight of Behavior-based is that a risky situation is identified if the ego vehicle behavior is influenced [26]. We manually label the time index, where the ego vehicle’s behavior is influenced in a scenario. We implement two baselines. **Behavior Prediction (BP)** [77] utilizes graph attention networks [78] to model interactions between traffic participants and the ego-vehicle to predict ego behavior. If the predicted behavior is “Influenced,” we select the object with the highest attention scores as the risk. **Behavior Change Prediction (BCP)** [26] identifies risk by applying a causal inference-based approach [26].

It is worth noting that the collision anticipation and behavior-based approaches are vision-based.

B. Evaluation Metrics

We devise three metrics that evaluate the ability of a risk identification algorithm to (1) identify locations of risks, (2) anticipate risks, and (3) facilitate decision-making.

- 1. Risk Localization:** We utilize **precision** and **recall** for risk localization evaluation, following the existing collision and anticipation literature [1], [2], [5]. We utilize the **False Alarm Rate** for *Non-interactive* scenarios (no risky dynamic elements or events), i.e., a positive prediction is a false alarm.
- 2. Risk Anticipation:** Risk Localization treats the prediction of each frame independently, without evaluating if a model can anticipate risks or unexpected events. Time-to-collision (TTC) [1], [2], [63] is one of the most common metrics to evaluate risk anticipation. However, they do not consider false identification of risk. To this end, we propose **Progressive Increasing Cost (PIC)**, an anticipation and localization-aware metric. Our insight is that we penalize a false identification more when we are closer to the risk than far from the risk. We propose to combine the notion of TTC, the exponential loss [76], and the F1-score to establish an anticipation and localization-aware metric. Given a scenario

TABLE II: **Risk Localization and Anticipation Evaluation.** The notations P, R, PIC, and FA denote precision, recall, progressive increasing cost, and false alarm rate, respectively.

	Interactive			Collision			Obstacle			Non-inter.	All
	P (%)	R (%)	PIC	P (%)	R (%)	PIC	P (%)	R (%)	PIC	FA(%)	F-1
Random	9.5	50.0	16.7	19.2	49.8	15.7	7.2	50.3	11.4	80.8	15.1
Range (5m)	38.7	10.7	20.3	85.6	27.7	4.1	45.0	18.3	10.2	3.3	30.0
Range (10m)	43.6	69.1	7.1	69.3	62.4	3.8	33.3	88.5	0.8	15.2	53.6
Kalman filter [71]	37.0	54.0	15.4	59.9	58.1	4.2	35.3	61.2	10.2	18.8	46.7
Social-GAN [72]	39.4	57.7	14.1	60.8	45.4	7.8	36.9	69.3	4.0	16.3	46.0
MANTRA [28]	36.3	62.1	13.3	58.3	45.4	7.1	35.7	71.8	2.5	16.5	45.4
QCNet [31]	40.0	56.6	14.1	60.3	45.4	7.3	38.2	68.9	4.4	14.5	46.4
DSA [1]	54.7	19.7	20.9	79.4	45.9	3.7	54.2	47.2	17.5	3.3	46.8
RRL [75]	49.4	15.4	22.4	87.3	64.6	4.6	47.1	19.0	16.0	5.1	48.6
BP [77]	36.4	16.3	25.6	74.4	9.8	25.1	22.4	3.0	27.9	4.7	16.6
BCP [26]	48.7	29.2	22.8	76.8	20.7	20.8	32.1	35.2	20.5	4.7	33.4

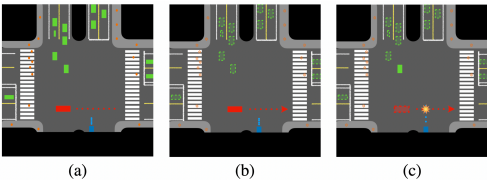


Fig. 4: **Planning-aware Metric.** An ideal planner should be able to yield to the crossing vehicle as shown in (a). If a risk identification model successfully identifies the vehicle as the risk, we hide all the other objects. The planner should plan a slow-down path, as shown in (b). In contrast, if the model identifies the wrong object as risk (e.g., the green one), the planner will plan a path to (nearly) collide with the **True Risk**, as shown in (c).

with T frames, we calculate the F-1 score for each frame t and define the metric as $PIC = -\sum_{t=1}^n e^{-(T-t)} \log(F1_t)$.

3. Planning Awareness: We propose **Influenced Ratio (IR)** to evaluate if a risk identification algorithm can facilitate decision-making. Assume we access an ideal planner that can interact with traffic participants and unexpected events. Our insight is that, if the identified risks are accurate, the planner should plan a path without collisions, as illustrated in Figure 4 (b). Specifically, IR is defined as $IR = |D_{orig} - D_{post}| / D_{orig}$, where the parameter D_{orig} denotes the closest distance while interacting between the ego vehicle and the true risk in a scenario using the full observation. The parameter D_{post} denotes the closest distance while interacting between the ego vehicle and the true risk with the masked observation. The value of IR is between 0 and 1. If the value is closer to 1, there is a higher possibility that there is a false identification influencing the planner’s decision-making, resulting in safety-critical situations. We also adopt **Collision Rate**, a commonly used metric for evaluating planning and decision-making [49], [40].

C. Implementation Details

Dataset. The testing set comprises a total of 1689 scenarios, including 521 interactive, 375 collision, 322 obstacle, and

471 non-interactive scenarios. We split train/val/test according to maps provided by CARLA [52].

Inputs. To make a fair comparison among all methods, vision-based algorithms (i.e., **DSA** [1], **RRL** [75], **BP** [77], and **BCP** [26]) access the ground-truth tracklets as input. We restrict the field of view of all methods to the front view to ease the multi-view fusion problem.

Planning-aware Evaluation. Our planner is the privileged model proposed in [79], which takes a bird’s-eye-view of semantic maps as input and outputs waypoints. The model is trained to imitate demonstrations collected in the RiskBench. The evaluation is conducted on *Interactive* and *Obstacle* types. We exclude *Collision* scenarios because we found the trained planner cannot handle collision scenarios and leave it for future work.

D. Risk Localization and Anticipation Evaluation

We evaluate algorithms’ ability to localize risks via Precision, Recall, and False Alarm Rate (FA rate) and to anticipate risk via Progressive Increasing Cost (PIC). We also report an overall F-1 score, considering all scenarios. The complete scenario-based performance analysis is shown in Table II.

Rule-based. We find that **Range (5m)** identifies risks only when the risk is very close, resulting in unsatisfactory recall in all safety-critical scenarios. Although increasing the distance threshold to 10m can significantly improve performance in most scenarios, **Range (10m)** increases FA in *Non-interactive* scenarios compared to **5m**. Hyperparameter tuning for rule-based methods to strike a balance between precision and recall across diverse scenarios is challenging. In collision scenarios, **Trajectory Prediction and Collision Checking** outperforms the other interaction types in terms of precision and PIC. This is due to the short-term prediction horizon of these methods when the ego-vehicle decelerates, preventing trajectory overlap. This is because the predicted trajectories become short when the ego-vehicle slows down, which normally results in no overlap of trajectories. The same trend is seen in “Non-interactive” scenarios, where all trajectories remain fast, resulting in inferior FA.

TABLE III: **Planning-aware Evaluation.** IR and CR denote **Influenced Ratio** and **Collision Rate**, respectively.

Risk ident. Algorithm	Interactive		Obstacle	
	IR	CR (%)	IR	CR (%)
Auto-pilot [52]	0.45	33.6	0.52	58.7
Full observation	0.00	0.0	0.00	0.0
Ground truth risk	0.02	0.4	0.04	5.3
Random	0.11	7.7	0.44	48.1
Range (10m)	0.01	6.2	0.24	26.4
Kalman Filter [71]	0.06	6.9	0.33	41.8
Social-GAN [72]	0.16	10.0	0.49	55.3
MANTRA [28]	0.15	9.3	0.51	56.7
QCNet [31]	0.05	10.0	0.52	57.2
DSA [1]	0.01	0.8	0.37	38.5
RRL [75]	0.01	1.2	0.49	54.3
BP [77]	0.14	9.3	0.41	47.1
BCP [26]	0.14	9.7	0.34	39.4

TABLE IV: **Temporal Consistency** is calculated within 1, 2, and 3 seconds before critical/collision points.

	1s	2s	3s
QCNet [31]	50.2%	26.9%	18.5%
DSA [1]	14.0%	4.6%	3.8%
RRL [75]	19.0%	8.5%	4.7%
BP [77]	4.2%	2.4%	1.9%
BCP [26]	6.9%	3.9%	3.4%

Collision Anticipation. RRL outperforms all learning-based models in terms of the F-1 score on all scenarios since the method directly learns to localize risks. We notice **RRL** and **DSA** do not perform well in *Interactive* and *Obstacle* scenarios because they are intrinsically different from *Collision*. **Behavior Prediction-based.** **BP** [77] faces difficulty identifying the risk affecting ego-vehicle’s action. **BCP** [26] performs better due to the causal inference. However, the two methods generally have low recall due to weak supervision.

E. Planning-aware Evaluation

We study the influence of risk identification on the downstream driving task in Table III. Initially, we test the built-in *Auto-pilot* in CARLA to run the collected scenarios. However, *Auto-pilot* can not perform well in *Interactive* and crashes in 58% of *Obstacle*. We thus adopt the privileged model proposed in [79] as our planner. We compare the planner [79] using the *Full observation* and restricted observation where the planner only observes the ground truth risk. The planner with restricted observation performs slightly inferior in both *Interactive* and *Obstacle* scenarios because we do not want the planner to overfit the recorded trajectories, which will make the evaluation invalid.

For *Rule-based* algorithms, *Range* achieves favorable results because it owns the nearly full observation by detecting every near object as a risk. While *trajectory prediction and collision checking* algorithms perform favorably in terms of **Recall** in Table II, we observe the trend is not aligned in the planning-aware evaluation. This is because the models cannot successfully predict the risk until the risk is nearby. Thus, the planner cannot react in time, especially for *Obstacle*,

requiring more buffer to avoid. In contrast, while vision-based algorithms are not robust temporally (see PIC results in Table II), they can predict the risk earlier by incorporating visual information. Therefore, *Behavior prediction-based* models perform on par with *Trajectory Prediction and Collision Checking* in *interactive*, and better in the *obstacle*. Among them, **DSA** outperforms all baselines in *interactive* and **BCP** outperforms all learning-based methods in *obstacle*. Combined with the risk localization and anticipation benchmark and the planning-aware evaluation, the results suggest that a risk identification that incorporates visual information is inevitable but requires more effort to improve the robustness, which we will discuss in Section V.

V. DISCUSSION

Trajectory Prediction and Collision Checking. Despite the recent huge progress in trajectory prediction, we find out that the top-ranked method QCNet [31] performs on par with other trajectory prediction methods. In other words, a high-performing trajectory predictor does not always guarantee improved risk identification performance. This is due to shortcomings in the “Collision Checking” mechanism. Our evaluation, which takes planning into account, further strengthens this observation. We hope our findings will encourage the community to collaboratively address the safety challenge from a holistic system perspective.

Temporal Consistency. Inspired by the observation in planning-aware evaluation (Section IV-E), we study the temporal consistency of models’ predictions, as shown in Table IV. To evaluate temporal consistency, we determine if a risk is predicted accurately and consistently within a specified time frame, leading up to the critical/collision point. We discover the benchmarked vision-based algorithms lack temporal consistency. The finding raises significant reliability concerns for the deployment in the real world. Our hypothesis for subpar temporal consistency is that they are trained with per-frame supervision. We further conduct experiments to demonstrate that using a heuristic smoothing method can significantly improve downstream decision-making in our **Video Submission**. This evidence underscores the imperative for more robust spatial-temporal modeling, such as the Transformer [80], [81], [82], to enhance the performance of risk identification and subsequent tasks.

VI. CONCLUSION

We introduce RiskBench, a scenario-based benchmark focusing on assessing risks induced by dynamic traffic participants and unexpected events. We explicitly design a scenario taxonomy for systematic collection of risks and scenario-based performance analysis. We benchmark four types of existing risk identification algorithms by assessing them using three risk metrics, i.e., risk localization, risk anticipation, and planning awareness. In the future, we will explore more advanced risk-aware planners [83], [84], [85], [86] to enhance robustness of the planning-aware evaluation.

ACKNOWLEDGEMENT

The work is sponsored in part by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), the Yushan Fellow Program Administrative Support Grant, and the National Science and Technology Council (NSTC) under grants 110-2222-E-A49-001-MY3, 110-2634-F-002-051, 111-2634-F-002-022-, Mobile Drive Technology Co., Ltd (MobileDrive), and Industrial Technology Research Institute Mechanical and Mechatronics Systems Lab.

REFERENCES

- [1] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating Accidents in Dashcam Videos," in *ACCV*, 2016.
- [2] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh, "Anticipating Traffic Accidents with Adaptive Loss and Large-scale Incident DB," in *CVPR*, 2018.
- [3] 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.
- [4] J. Fang, D. Yan, J. Qiao, and J. Xue, "DADA: A Large-scale Benchmark and Model for Driver Attention Prediction in Accidental Scenarios," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [5] T. You and B. Han, "Traffic Accident Benchmark for Causality Recognition," in *ECCV*, 2020.
- [6] M. M. Minderhoud and P. H. Bovy, "Extended Time-to-collision Measures for Road Traffic Safety Assessment," *Accident Analysis & Prevention*, vol. 33, no. 1, pp. 89–97, 2001.
- [7] M. Althoff, O. Stursberg, and M. Buss, "Stochastic Reachable Sets of Interacting Traffic Participants," in *IV*, 2008.
- [8] A. Lawitzky, D. Althoff, C. F. Passenberg, G. Tanzmeister, D. Wollherr, and M. Buss, "Interactive Scene Prediction for Automotive Applications," in *IV*, 2013.
- [9] S. Lefèvre, D. Vasquez, and C. Laugier, "A Survey on Motion Prediction and Risk Assessment for Intelligent Vehicles," *ROBOMECH Journal*, vol. 1, p. 1, 2014.
- [10] B. Ivanovic and M. Pavone, "The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs," in *ICCV*, 2019.
- [11] M. Koschi and M. Althoff, "Set-Based Prediction of Traffic Participants Considering Occlusions and Traffic Rules," *IV*, vol. 6, 2020.
- [12] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions," in *CVPR*, 2019.
- [13] T. Salzmann, B. Ivanovic, P. Chakravarthy, and M. Pavone, "Trajectron++: Dynamically-Feasible Trajectory Forecasting With Heterogeneous Data," in *ECCV*, 2020.
- [14] S. Malla, B. Dariush, and C. Choi, "TITAN: Future Forecast using Action Priors," in *CVPR*, 2020.
- [15] L. Neumann and A. Vedaldi, "Pedestrian and Ego-vehicle Trajectory Prediction from Monocular Camera," in *CVPR*, 2021.
- [16] M. Spain and P. Perona, "Some objects are more equal than others: measuring and predicting importance," in *ECCV*, 2008.
- [17] E. Ohn-Bar and M. Trivedi, "Are all objects equal? Deep Spatio-temporal Importance Prediction in Driving Videos," *Pattern Recognition*, vol. 64, pp. 425–436, 2017.
- [18] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Nibbles, and M. Sun, "Agent-Centric Risk Assessment: Accident Anticipation and Risky Region Localization," in *CVPR*, 2017.
- [19] M. Gao, A. Tawari, and S. Martin, "Goal-oriented Object Importance Estimation in On-road Driving Videos," in *ICRA*, 2019.
- [20] Z. Zhang, A. Tawari, S. Martin, and D. Crandall, "Interaction Graphs for Object Importance Estimation in On-road Driving Videos," in *ICRA*, 2020.
- [21] 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.
- [22] Y. Xia, D. Zhang, J. Kim, and D. W. Ken Nakayama, Karl Zipser, "Predicting Driver Attention in Critical Situations," in *ACCV*, 2018.
- [23] Y. Xia, J. Kim, J. F. Canny, K. Zipser, T. Canas-Bajo, and D. Whitney, "Periphery-Fovea Multi-Resolution Driving Model Guided by Human Attention," in *WACV*, 2020.
- [24] A. Pal, S. Mondal, and H. I. Christensen, "'Looking at the right stuff' - Guided semantic-gaze for autonomous driving," in *CVPR*, 2021.
- [25] S. Bae, 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.
- [26] C. Li, S. H. Chan, and Y.-T. Chen, "Who Make Drivers Stop? Towards Driver-centric Risk Assessment: Risk Object Identification via Causal Inference," in *IROS*, 2020.
- [27] —, "DROID: Driver-centric Risk Object Identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [28] F. Marchetti, F. Becattini, L. Scidenari, and A. D. Bimbo, "MANTRA: Memory Augmented Networks for Multiple Trajectory Prediction," in *CVPR*, 2020.
- [29] Z. Zhou, L. Ye, J. Wang, K. Wu, and K. Lu, "Hivt: Hierarchical vector transformer for multi-agent motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [30] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2980–2987.
- [31] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [32] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
- [33] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov, "Large Scale Interactive Motion Forecasting for Autonomous Driving: The Waymo Open Motion Dataset," *arXiv*, 2021.
- [34] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kümmerle, H. Königshof, 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:1910.03088*, 2019.
- [35] Hala Elrofai and Jan-Pieter Paardekooper and Erwin de Gelder and Sytze Kalisvaart and Olaf Op den Camp, "StreetWise: Scenario-based Safety Validation of Connected Automated Driving," 2018.
- [36] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for Development, Test and Validation of Automated Vehicles," in *ECCV*, 2018.
- [37] U.S. Department of Transportation, National Highway Traffic Safety Administration, "A Framework for Automated Driving System Testable Cases and Scenarios," 2018.
- [38] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on Scenario-Based Safety Assessment of Automated Vehicles," *IEEE Access*, vol. PP, pp. 1–1, 05 2020.
- [39] 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 Track on Datasets and Benchmarks*, 2022.
- [40] "CARLA Autonomous Driving Challenge," <https://carlachallenge.org/>, 2022.
- [41] N. Agarwal and Y.-T. Chen, "Ordered atomic activity for fine-grained interactive traffic scenario understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8624–8636.
- [42] C.-H. Kung, S.-W. Lu, Y.-H. Tsai, and Y.-T. Chen, "Action-slot: Visual action-centric representations for multi-label atomic activity recognition in traffic scenes," *arXiv preprint arXiv:2311.17948*, 2023.
- [43] T. You and B. Han, "Traffic Accident Benchmark for Causality Recognition," in *ECCV*, 2020.
- [44] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *CVPR*, 2016.
- [45] 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 IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

- [46] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *CVPR*, 2012.
- [47] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving," *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.
- [48] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 723–13 733.
- [49] H. Kim, K. Lee, G. Hwang, and C. Suh, "Crash To Not Crash: Learn to Identify Dangerous Vehicles using a Simulator," in *AAAI*, 2019.
- [50] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.
- [51] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1043–1052.
- [52] D. Zhao, Y. Guo, and Y. J. Jia, "TrafficNet: An Open Naturalistic Driving Scenario Library," in *IITSC*, 2017.
- [53] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *CoRL*, 2017.
- [54] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [55] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10914–10921, 2022.
- [56] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu, "SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 21 371–21 382.
- [57] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 107–124.
- [58] S. Li, S. Zhang, G. Chen, D. Wang, P. Feng, J. Wang, A. Liu, X. Yi, and X. Liu, "Towards benchmarking and assessing visual naturalness of physical world adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 324–12 333.
- [59] Waymo, "Simulation City: Introducing Waymo's most advanced simulation system yet for autonomous driving," https://waymo.com/blog/2021/06/SimulationCity.html?hl=zh_TW, 2021.
- [60] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, "Unisim: A neural closed-loop sensor simulator," in *CVPR*, 2023.
- [61] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Causal Reasoning," in *CVPR*, 2018.
- [62] R. Debouk and J. Joyce, "ISO 26262 Hazard and Risk Assessment Methodology," 2010.
- [63] A. Badki, O. Gallo, J. Kautz, and P. Sen, "Binary TTC: A Temporal Geofence for Autonomous Navigation," in *CVPR*, 2021.
- [64] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes," in *ICRA*, 2019.
- [65] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior," in *ICCVW*, 2017.
- [66] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A Large-scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction," in *ICCV*, 2019.
- [67] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi, "LOKI: Long Term and Key Intentions for Trajectory Prediction," in *ICCV*, 2021.
- [68] J. Phillion, A. Kar, and S. Fidler, "Learning to Evaluate Perception Models using Planner-Centric Metrics," in *CVPR*, 2020.
- [69] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "PlanT: Explainable Planning Transformers via Object-Level Representations," in *CoRL*, 2022.
- [70] D. Lu, E. Eaton, M. Weg, W. Wang, S. Como, J. Wishart, H. Yu, and Y. Yang, "CAROM air - vehicle localization and traffic scene reconstruction from aerial videos," in *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*. IEEE, 2023, pp. 10 666–10 673. [Online]. Available: <https://doi.org/10.1109/ICRA48891.2023.10160502>
- [71] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.
- [72] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks," in *CVPR*, 2018.
- [73] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.
- [74] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. Kaesemodel Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1. Curran, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4734ba6f3de83d861c3176a6273cac6d-Paper-round2.pdf
- [75] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. Carlos Nieves, and M. Sun, "Agent-centric Risk Assessment: Accident Anticipation and Risky Region Localization," in *CVPR*, 2017.
- [76] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent Neural Networks for Driver Activity Anticipation via Sensory-fusion Architecture," in *ICRA*, 2016.
- [77] C. Li, Y. Meng, S. H. Chan, and Y.-T. Chen, "Learning 3d-aware ego-centric spatial-temporal interaction via graph convolutional networks," in *ICRA*, 2020.
- [78] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *ICLR*, 2018.
- [79] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *CoRL*, 2020.
- [80] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *Advances in Neural Information Processing Systems*, 2022.
- [81] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Vip3d: End-to-end visual trajectory prediction via 3d agent queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5496–5506.
- [82] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [83] B. Chen, Z. Liu, J. Zhu, M. Xu, W. Ding, L. Li, and D. Zhao, "Context-aware safe reinforcement learning for non-stationary environments," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 689–10 695.
- [84] C. Diehl, T. Sievernich, M. Krüger, F. Hoffmann, and T. Bertram, "Umbrella: Uncertainty-aware model-based offline reinforcement learning leveraging planning," *arXiv preprint arXiv:2111.11097*, 2021.
- [85] H. Xu, X. Zhan, and X. Zhu, "Constraints penalized q-learning for safe offline reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8753–8760.
- [86] C. Diehl, T. S. Sievernich, M. Krüger, F. Hoffmann, and T. Bertram, "Uncertainty-aware model-based offline reinforcement learning for automated driving," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1167–1174, 2023.