SafeShift: Safety-Informed Distribution Shift for Robust Trajectory Prediction in Autonomous Driving

Benjamin Stoler[1*] Ingrid Navarro[1*] Meghdeep Jana[1] Soonmin Hwang[2] Jonathan Francis[3] Jean Oh[1]

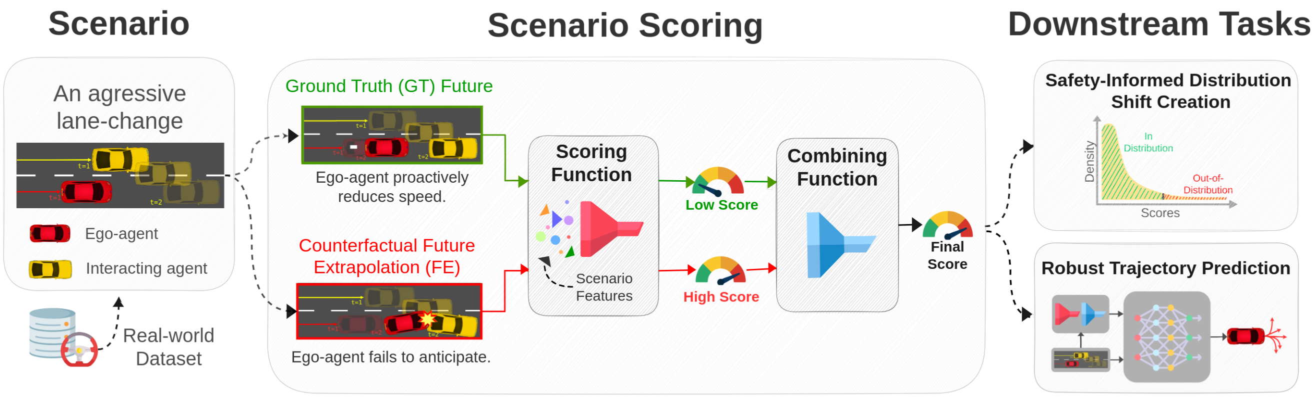


Fig. 1: An overview of SafeShift. Our framework consists of a scoring methodology that uses counterfactual probing to characterize and score scenarios, exploring *what-if* scenarios where proactive maneuvers were not performed, thus resulting in safety-criticality or near misses. We also apply and assesses this scoring approach on two downstream tasks: safety-informed *distribution shift creation*, where challenging scenarios are found and held out for evaluation; and *robust trajectory prediction*, where trajectory prediction algorithms are assessed under this distribution shift and remediated.

***Abstract*— As autonomous driving technology matures, the safety and robustness of its key components, including trajectory prediction is vital. Although real-world datasets such as Waymo Open Motion provide recorded real scenarios, the majority of the scenes appear benign, often lacking diverse safety-critical situations that are essential for developing robust models against nuanced risks. However, generating safety-critical data using simulation faces severe simulation to real gap. Using real-world environments is even less desirable due to safety risks. In this context, we propose an approach to utilize existing real-world datasets by identifying safety-relevant scenarios naively overlooked, e.g., near misses and proactive maneuvers. Our approach expands the spectrum of safety-relevance, allowing us to study trajectory prediction models under a safety-informed, distribution shift setting. We contribute a versatile scenario characterization method, a novel scoring scheme for reevaluating a scene using counterfactual scenarios to find hidden risky scenarios, and an evaluation of trajectory prediction models in this setting. We further contribute a remediation strategy, achieving a 10% average reduction in predicted trajectories' collision rates. To facilitate future research, we release our code for this overall SafeShift framework to the public: github.com/cmubig/SafeShift**

I. Introduction

As autonomous driving (AD) technologies are increasingly deployed in the wild, the safety and robustness of the autonomous systems remain chief concerns [1]–[3]. One key AD task is that of trajectory prediction, wherein the future trajectories of agents in a scene must be predicted, given a brief historical observation. These predictions may be used in the downstream portion of conventional vehicle control stacks, to inform an ego-agent's motion planner as it attempts to find possible conflict-free and traffic infraction-free paths. Thus, improving the agent's robustness and its ability to detect possibly safety-critical scenarios is of paramount importance in ensuring the overall acceptable performance of autonomous vehicles in real-world deployments [4].

It is appealing to train trajectory prediction models using large real-world motion prediction datasets, such as the Waymo Open Motion Dataset (WOMD) [5], as they consist of recorded scenarios capturing the behaviors of various agents—human drivers and vulnerable road users (VRUs), e.g., cyclists and pedestrians—under real-world traffic layouts and densities. One inherent challenge in using such datasets, however, is that the frequency of vehicle infractions and other safety-critical scenarios therein is quite low. The prior art regards this issue as the "curse of rarity" [6]–[8] and, as a result, industry and academia have resorted to validating autonomous driving agents via on-road tests [9], [10], where those valuable rare events are also potentially dangerous to other drivers and VRUs, or via simulated experiments [6]–[8], wherein the artificial behaviors of agents and inaccurate world physics in the simulators can leave models unprepared and inadequate for real-world deployment [11], [12].

*First authors BS and IN contributed equally. Work done by MJ and SH while at CMU. [1]School of Computer Science, Carnegie Mellon University, Email: {benstoler, jeanoh}@cmu.edu, ingridn@cs.cmu.edu. [2]Department of Automotive Engineering, Hanyang University, Email: soonminh@hanyang.ac.kr. [3]Bosch Center for Artificial Intelligence, Email: jon.francis@us.bosch.com.

Recently, several works have identified a potential solution to this challenge of robust training, by generating "new" traffic scenes that serve as training samples for otherwise rare events and/or as difficult test-cases to challenge already-trained models. Unfortunately, despite recent advances in safety-critical scenario generation methods [13]–[15], generating non-trivially challenging cases that match the realism, frequency, and difficulty of safety-critical scenarios that agents might encounter in the real world remains an open problem. An effective and under-explored alternative lies somewhere in the middle: we propose an approach to mine large-scale datasets of real-world vehicle deployments to find and leverage meaningful *safety-relevant* scenarios that may be hidden in the data. Our key insight is that, in autonomous driving, safety-relevance includes not just scenarios where observed agents act in a safety-critical manner, but also scenarios where agents are able to avoid infractions through proactive maneuvers. Therefore, we propose to leverage counterfactual probes to additionally characterize *what-if* scenarios where these proactive maneuvers were *not* performed. Such fine-grained scenario characterization enables trajectory forecasting models to more easily distill diverse defensive driving skills [16] from existing datasets, e.g., preemptive braking as illustrated in Figure 1 (left).

Under this paradigm of scenario characterization, we propose the `SafeShift` framework for identifying and studying the most safety-relevant scenarios in a widely-available autonomous driving dataset. The more extreme scenarios are held out as an out-of-distribution (OOD) test-set, thus acting as a stand-in for the valuable and rare, long-tailed events. In this way, we avoid both the challenges of attempting to generate new safety-critical scenarios as well as the challenges in performing simulation-to-real transfer; instead, we optimize the usefulness of existing data. To the best of our knowledge, prior work that focuses on creating artificial distribution shifts have not done so based on safety-relevance, instead focusing on, e.g., lane or global location characteristics [17], [18], speed of driving [19], or the city that the data was captured in [19]. Furthermore, prior efforts in scenario characterization under distribution shift settings rely on empirical, dataset-specific heuristics [20]–[22].

Our main contributions, illustrated in Figure 1, are thus as follows: 1) A versatile approach for scenario characterization in autonomous driving, focused on capturing safety-relevant scenarios; 2) A methodology for scoring safety-criticality for the purposes of crafting a distribution shift, including novel progress in incorporating the aforementioned fuller spectrum of safety-relevance, and improving safety performance therein; and 3) An evaluation of existing socially-aware trajectory prediction approaches in this safety-informed distribution shift setting, utilizing WOMD [5] as an exemplar. Our developed remediation strategy for this setting reduces the predicted trajectories' collision rates by an average of 10%, across the tested models.

II. Related Work

A. Socially-Aware Trajectory Prediction

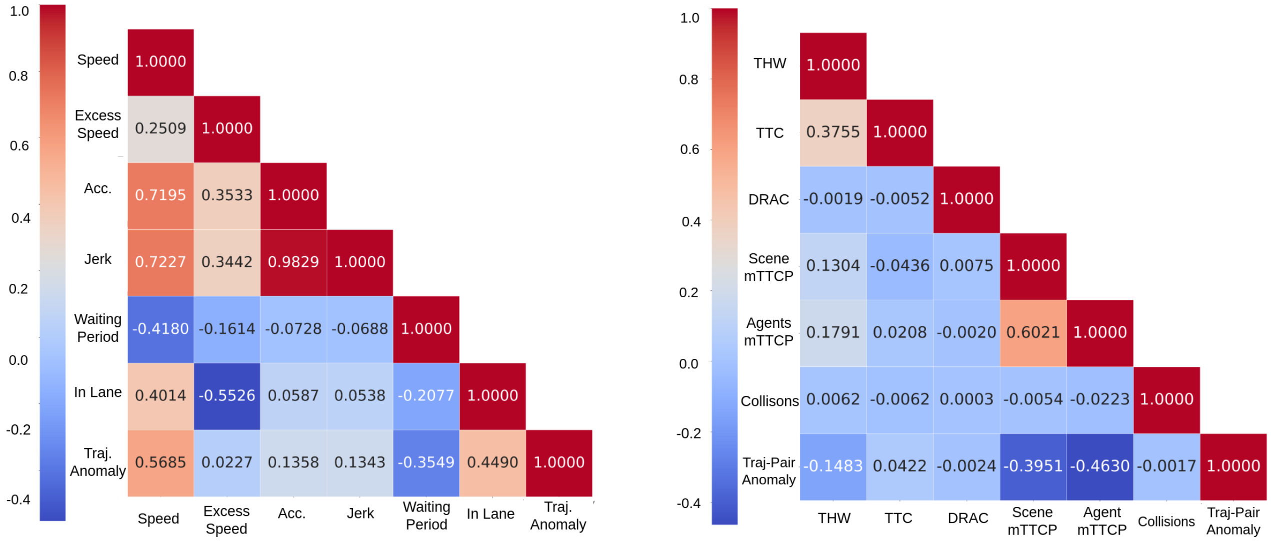
Motion prediction in crowded environments is a well-researched task in the domains of autonomous driving and motion in human crowds [23]. Most current approaches for motion prediction are data-driven, i.e., they focus on characterizing behavior and interactions observed in the data. To capture a multi-modal distribution of possible futures, generative frameworks are frequently used [24]–[29]. To model joint behavior and social cues, various techniques such as social pooling [30], rasterized representations [31], and attention-based methods [24], [27], [32], [33] have been employed. Several state-of-the-art techniques have also explored learning richer representations for motion prediction, such as modeling context information as road graphs or polylines [34], [35] and goal conditioning [25].

B. Robustness Assessment in Trajectory Prediction

One approach to examine robustness for trajectory prediction is robustness to adversarial attacks. Recent studies have shown that state-of-the-art prediction models often lack basic social awareness and collision avoidance when faced with these attacks [36]–[38]. A significant disadvantage with these techniques however is that they ultimately rely on simulating realistic agent behavior, which often incurs a simulation-to-real gap [11]–[13], [39]. Another approach to ensuring robustness involves studying models' performances under a data domain distribution shift setting, recognizing that AD models will ultimately encounter unseen scenarios in the wild. Some approaches involve identifying domains based on meta characteristics of the scene, such as road shape characteristics, side-of-driving, and average speeds [17], [19]. Another recent method explores clustering scenes into domains based on several features, including lane deflection angles, global bounds of the scenario and trajectories, and lane shape information [18]. Many of these works also include domain adaptation or remediation strategies to reduce the impact of the distribution shift, such as by leveraging Frenet coordinates [18], [40], few-shot adaptation [17], or motion-based style transfer [41]. However, to the best of our knowledge, no work has attempted to create distribution shifts based on safety-relevance or study remediation therein.

C. Critical Scenario Identification in Autonomous Driving

Many existing datasets rely on mining the immense amount of collected data from road-tests for interesting scenes [5], [42], [43], considering surface-level metrics such as traffic density and kinematic complexity. Therefore, prior frameworks for critical scenario identification (CSI) have been designed to expand upon these initial dataset characterizations [44], [45]. These frameworks typically focus on creating taxonomies for categorizing conflict scenarios, as well as for developing metrics and validation methods to describe and cope with them. In [21], scenarios are instead hierarchically scored along metrics related to anomaly, interestingness, and relevance, better handling more complex maneuvers. Another recent work expands beyond



(a) Correlation Analysis for Individual Features

(b) Correlation Analysis for Social Features

Fig. 2: Pearson correlation coefficients for each pair of metrics, showing how the features complement each other. Analysis performed on WOMD [5].

this by defining complexity aspects relating to the road graph layout, surrounding objects, and topology of agents' paths [22]. However, the use of these surrogate metrics for CSI alone, without applying counterfactual reasoning, can fail to identify more subtle safety-relevant scenarios, as illustrated in Figure 1. Furthermore, these metrics often rely on empirical weighting and thresholding schemes, as well as on privileged information not uniformly available in AD datasets (e.g., global reference frames, drivable area identification, etc.) [21], [22]; thus they cannot be applied to several key datasets, including WOMD.

III. Preliminaries

In this section, we define the task of trajectory prediction under distribution shifts. First, we consider the set of scenarios that comprise a motion prediction dataset as $\mathcal{S}$. Thus, we denote $s \in \mathcal{S}$ as a single scenario taken from this corpus. The scenario s consists of all agent tracks $\mathbf{X}$, map information, and meta information. Agent tracks are the time-varying locations of every observed agent in the scene, in a Cartesian frame, where $\mathbf{X}_t^{(i)}$ denotes the location of agent i at timestep t. The map contains road information, e.g., lane locations and lane connectivity, and the meta information provides additional task specifications, such as the list of which agents are to be predicted.[1]

The information in s is further split into a history and future portion: $s_{hist} = \{s_1, s_2, ..., s_{T_{obs}}\}$ and $s_{fut} = \{s_{T_{obs+1}}, s_{T_{obs+2}}, ..., s_{T_{tot}}\}$, where T_{obs} denotes the timestep before the prediction horizon and T_{tot} denotes the total length of the scene. Thus, the task of trajectory prediction is to jointly estimate the values of $\mathbf{X}_{fut}^{(i)}$, using only s_{hist}, for all agents $i \in \{1, 2, 3, ...\}$.

Under distribution shift conditions, $\mathcal{S}$ may be split into two sets—$\mathcal{S}_{ID}$, representing the in-distribution set, and $\mathcal{S}_{OOD}$ representing the out-of-distribution set. The task of robust trajectory prediction then, is to minimize the drop in performance on safety-relevant metrics for trajectory prediction (i.e., collision rates) when prediction models are tested on $\mathcal{S}_{OOD}$, after being trained and validated only on $\mathcal{S}_{ID}$.

The remaining sections in this work are organized as follows. In Section IV and Section V, we propose a novel scenario-characterization framework and scoring methodology. Then in Section VI and Section VII, we show how to leverage our framework for introducing safety-informed distribution shifts in a given autonomous driving dataset and for developing remediation strategies to improve the robustness of trajectory prediction models. Finally in Section VIII we discuss the results and implications of these experiments.

IV. Scenario Features

We propose a hierarchical scheme as in [21], [22], where low-level, base features are computed within a scenario and then later aggregated to form a score representing a scenario's overall safety-relevance. We consider base features across two main categories: *individual* features related to single-agent behavior and *social* features relevant to the interactions between agents.

For both of these categories, accurate lane assignment is highly important but is nontrivial, e.g., VRUs often do not adhere to lanes. Whereas a simple method of snapping to the best-fitting local lane has been used in previous work [18], we instead leverage a probabilistic approach [46] to find valid lane *sequences* for agents. Additionally, we permit lane assignments based on physically plausible lane deflection angles rather than the lane connectivity graph alone. We excluded some features utilized in previous frameworks and

[1] The exact format of s, such as the origin of the Cartesian frame or specific set of map information provided, varies from dataset to dataset.

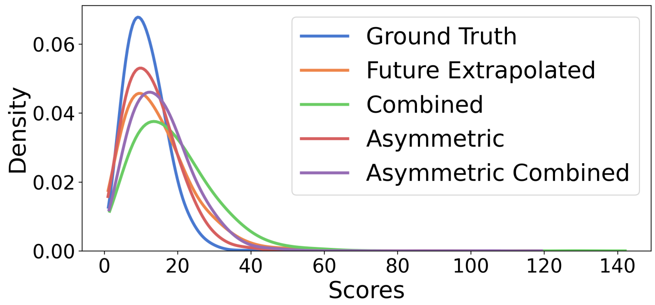


Fig. 3: PDF of our score variations, exhibiting long-tailed behavior. Analysis performed in WOMD [5].

TABLE I: Trajectory scoring variations.

Variation	IndScore	SocScore
Ground Truth (GT)	$\mathbf{X}_{GT}^{(i)}$	$(\mathbf{X}_{GT}^{(i)}, \mathbf{X}_{GT}^{(j)})$
Future Extrapolated (FE)	$\mathbf{X}_{FE}^{(i)}$	$(\mathbf{X}_{FE}^{(i)}, \mathbf{X}_{FE}^{(j)})$
Asymmetric (AS)	$\mathbf{X}_{FE}^{(i)}$	$(\mathbf{X}_{FE}^{(i)}, \mathbf{X}_{GT}^{(j)})$
Combined (CO)	$\max(\texttt{TrajScore}_{\texttt{GT}}, \texttt{TrajScore}_{\texttt{FE}})$	
Asymmetric Combined (AC)	$\max(\texttt{TrajScore}_{\texttt{GT}}, \texttt{TrajScore}_{\texttt{AS}})$	

datasets [20], [21], such as driving region-based anomaly detection, that require the knowledge of global, city coordinates which are not generally available across all AD datasets. Instead, to identify anomalies, we utilize a traffic primitive extraction and clustering approach pioneered in [47]. This process produces cluster centers for both single trajectories and trajectory pairs, allowing us to easily measure anomalies.

Individual Features: We primarily focus on metrics derived from relative positional data of a trajectory, such as speed, acceleration, and jerk. We additionally implement metrics to incorporate map context, including waiting period (WP) [48], speed difference with the lane's speed limit, and the percentage of time that the agent is following a lane. Finally, we include a trajectory anomaly value, derived from its distance to the nearest individual traffic primitive cluster.

Social Features: We use widely studied and accepted safety surrogate metrics, as in [21], [49], [50]. These include time headway (THW), time-to-collision (TTC), deceleration rate to avoid crash (DRAC), and the difference between minimum time to conflict points (ΔmTTCP) in both agent trajectories and road graph locations of interest (e.g., crosswalks, stop signs). We then incorporate a measure of collisions directly, counting situations where two agents' center points or segmented paths overlap at a given timestep. Finally, analogous to the individual trajectory anomaly, we add a trajectory-pair anomaly value using paired traffic primitive clusters.

Our full feature selection, along with a correlation analysis is shown in Figure 2. For the individual features, the kinematic-based ones correlate positively, as could be expected, while the other features are largely weakly correlated. Similarly, for the social features, TTC and THW have a weak correlation, as they both involve a leader-follower scenario. The two forms of ΔmTTCP are also relatively strongly correlated, as agent trajectories are required to be somewhat intertwined for both. This analysis implies that the selection and extraction of base features are largely complementary, without excessive overlap in coverage.

V. Scenario Scoring

Using the base features described in Section IV, we define a safety-relevance scoring function that can characterize a given scenario. We then propose a counterfactual re-scoring approach where we re-characterize the same scenario by taking *what-if* alternatives into account.

A. Scoring Functions

We start by hierarchically aggregating the base features to create overall trajectory and scene scores as follows. Let $\mathbf{V}_{ind}$ be the total set of extracted individual features, $\mathbf{V}_{soc}$ be the set of social features, and $v \in \mathbf{V}$ represent a single feature taken from one of these sets (e.g., acceleration, TTC, etc.). Then, let $v_t^{(i)}$ be such an extracted individual feature v for trajectory i at timestep t. Similarly, $v_t^{(i,j)}$ denotes a social feature over trajectories i and j together.

To combine these extracted base features, we first convert them to a form in which a larger value corresponds to more safety-relevance (i.e., for features such as speed, we use v directly, but for features such as TTC, we use $1/v$). We then aggregate the individual features into an individual score. We take the maximum value for each metric incurred throughout the trajectory and then linearly combine them according to weights specified in [21]; let these weights be denoted as $\mathbf{W}_{ind}$ and $\mathbf{W}_{soc}$. Then, a trajectory's individual score is expressed in Equation (1), where " · " denotes the vector scalar product:

$$\texttt{IndScore}^{(i)} = \mathbf{W}_{ind} \cdot \left[\max_t(v_t^{(i)}) \mid v \in \mathbf{V}_{ind} \right] \tag{1}$$

Note that we do not perform any sort of value detection thresholding to avoid reliance on empirical decision making. Similarly, for each pair of trajectories, we compute a social score, as follows in Equation (2):

$$\texttt{SocScore}^{(i,j)} = \mathbf{W}_{soc} \cdot \left[\max_t(v_t^{(i,j)}) \mid v \in \mathbf{V}_{soc} \right] \tag{2}$$

An agent's trajectory score is then computed by adding together its *individual* score with the *social* score of all trajectory pairs it is involved in:

$$\texttt{TrajScore}^{(i)} = \texttt{IndScore}^{(i)} + \sum_{j \neq i} \texttt{SocScore}^{(i,j)} \tag{3}$$

We combine these `TrajScores` into a final `SceneScore` as follows. We begin by taking the weighted sum of all agents' scores in the scene, where each weight is inversely proportionate to its minimum distance to an agent marked as requiring prediction. Then, to regularize the effect of scene density, we normalize this total, proportionate to the total number of agents present.

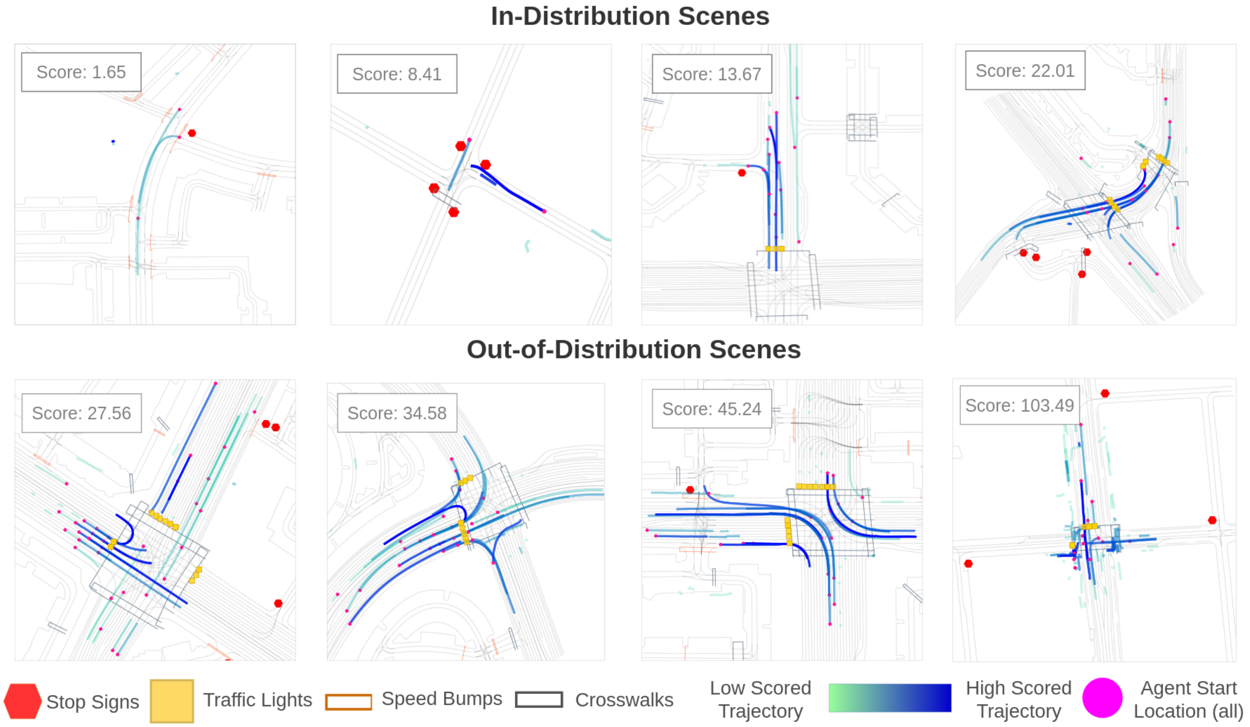


Fig. 4: Examples of WOMD [27] scenes by score. In-Distribution and Out-of-Distribution follow our `Scoring` split in Section VI-A.

B. Counterfactual Re-Scoring

The key insight of counterfactual re-scoring is to assess the safety-criticality of a scenario based on potential *what-if* cases in addition to the recorded ground truth event. We hypothesize that the characterization using counterfactual scenarios can capture the hidden risks better than using the ground truth record only, which will subsequently result in improved performance in downstream tasks such as robust trajectory prediction.

To find scenarios beyond just those with high aggregated criticality and/or surrogate criticality values, we wish to perform a counterfactual probe into what could happen if an agent were to simply maintain its current progress within a lane. This represents, e.g., the behavior of a distracted driver ignoring external factors. We craft this probe for an agent i by first extracting its assigned lane sequence in $\mathbf{X}_{hist}^{(i)}$. Next, we convert its coordinates to a Frenet frame [40], a coordinate system representing progress and displacements along the given lanes' centerlines. Finally, we perform a constant-velocity extrapolation in the Frenet frame, to compute a "future extrapolated" trajectory. For agents without a lane assignment, we perform the same steps in Cartesian space. We denote this future extrapolated trajectory as $\mathbf{X}_{FE}^{(i)}$, in contrast with the original ground truth trajectory, $\mathbf{X}_{GT}^{(i)}$.

To incorporate this method into the trajectory score in Equation (3), we extract the individual and social features of both $\mathbf{X}_{GT}^{(i)}$ and $\mathbf{X}_{FE}^{(i)}$. We first compute the individual score using $\mathbf{X}_{FE}^{(i)}$. To compute the social interaction scores, for a pair of interacting agents (i, j), we compute i's social score between $(\mathbf{X}_{FE}^{(i)}, \mathbf{X}_{GT}^{(j)})$ and j's score analogously. We denote this *asymmetric* score as $\texttt{TrajScore}_{AS}^{(i)}$. Similarly, we compute the reference ground truth score using exclusively the *GT* trajectories for both agents and denote this as $\texttt{TrajScore}_{GT}^{(i)}$. We then take the maximum value of these two scores into a final *asymmetric combined* trajectory score, $\texttt{TrajScore}_{AC}^{(i)}$. In Table I, we summarize these scoring variations and ablations.

We compute a `SceneScore` for these trajectory variations by utilizing the corresponding `TrajScore` (e.g., $\texttt{SceneScore}_{FE}$ uses $\texttt{TrajScore}_{FE}$ exclusively, etc.). As shown in Figure 3, this overall scene scoring method follows a long-tailed distribution as desired. The scores that incorporate future extrapolation have a much wider spread than just the ground truth, indicating a greater variety of scenarios captured.

VI. Downstream Tasks

We showcase the utility of our scenario scores from Section V by applying them to two downstream tasks: 1) creating a safety-informed distribution shift to better evaluate trajectory prediction models; and 2) leveraging the scores to conduct remediation on such models, reducing the incurred drop in performance.

A. Distribution Shift Creation

We wish to evaluate and improve the robustness of trajectory prediction models when facing scenes more challenging/safety-critical than those on which they were trained. That is, we must split $\mathcal{S}$ in such a way that $\mathcal{S}_{ID}$ contains relatively low safety-criticality while $\mathcal{S}_{OOD}$ contains the most criticality. Thus, we propose the following approach of splitting $\mathcal{S}$ into the desired safety-informed subsets.

First, we implement a simple uniform, random training/validation/test split to analyze behavior absent of a domain shift context: `Uniform`. Next, as a baseline, we implement the cluster-based domain identification schema from [18], representing a recent approach for domain shift creation that focuses on other aspects of the scenarios instead of safety-relevance: `Clusters`. Finally, we incorporate a safety-informed approach leveraging our schema described in Section V: `Scoring`. We hold out the top 20% scoring scenes as the test set, then randomly partition the remaining scenes into training and validation.

B. Robust Trajectory Prediction

We propose a remediation strategy leveraging the proposed scores in Section V to increase downstream prediction model performance on challenging, more safety-relevant scenarios. Inspired by the difficulty-weighting of samples, as discussed in [51], we utilize $\texttt{TrajScore}_{AC}^{(i)}$ for each agent i to linearly weigh its contribution to a prediction model's loss function, out of the N total agents in a mini-batch:

$$\texttt{WeightedLoss}^{(i)} = \frac{1}{N}\sum_{i}^{N}\texttt{Loss}^{(i)} * \texttt{Score}_{AC}^{(i)} \qquad (4)$$

Equation (4) is then applied after computing the loss function for a given model, but before invoking the optimization pass. This encourages the model to not treat all scenarios and agents' trajectories as equal and to care about more safety-relevant situations. Next, because the future-extrapolated score depends only on information available in s_{hist}, we can incorporate $\texttt{TrajScore}_{FE}^{(i)}$ into a model directly, to add a sense of counterfactual understanding to its inductive biases. We encode this score for each agent i with a simple multilayer perceptron. Then, we concatenate this feature directly with the context encoding representation used in each model (i.e., a function of trajectory histories, lane embedding, etc.) before passing it to the model's trajectory decoding stage.

We also propose to incorporate a collision-aware loss objective within each model. Many models in AD trajectory forecasting produce multi-modal futures, where they output K possible future modes for each agent, along with a scalar, confidence value for each [27], [29], [31]. We add in a cross-entropy (CE) loss objective upon these confidence values, where the "correct" mode is the mode that minimizes collisions with other agents' ground truth futures. In the case where a model already has a CE loss objective (e.g., to minimize the distance to the agent's ground truth future), we linearly weigh the two target values according to a regularization parameter.

VII. Experimental Setup

Dataset: We utilize WOMD [5] as an exemplar dataset to validate our approach, as it contains a particularly wide variety of scenarios. This variety is highlighted in terms of both geographic and roadway diversity, as well as scene complexity and traffic density [52], [53]. We utilize a subset from the publicly available training and validation sets from WOMD, consisting of roughly $170k$ scenarios. We consider our three different data splits (Section VI-A)—`Uniform`, `Clusters`, `Scoring`—to create $\mathcal{S}_{ID}$ for training and validation (roughly $135k$ scenes), and $\mathcal{S}_{OOD}$ for testing (roughly $35k$ scenes).

Baselines: We implement two representative baseline models to validate the efficacy of our distribution shift and remediation strategies. First, we include MTR [27], which, as of this writing, is the current top-performing model on WOMD leaderboards. Second, we implement a version of A-VRNN [24], where we utilize social pooling [30] instead of a graph attention layer for the hidden state refinement. While both models are designed to be "socially-aware," neither is explicitly structured to predict safe futures. We follow the same training procedure performed by MTR, where the models are trained for 30 epochs, and learning rate reduction begins after epoch 20.

As a baseline remediation strategy, we implement the Frenet-based domain normalization approach in [18]. This approach converts all coordinates into a trajectory's Frenet frame, before passing the coordinates to a trajectory prediction model. In order to obtain reasonable performance, we use both the Cartesian and Frenet coordinates *together* via concatenation, rather than replacing the former with the latter. We then implement our proposed remediation approach, described in Section VI-B for both models.

Metrics: To measure safety-criticality, we use collision rate (CR), as the average number of collisions of each predicted trajectory to the ground truth of the other agents, as in [54], where collisions with the same external agent over multiple timesteps only count once. We also utilize standard trajectory prediction metrics, as used in the WOMD challenge, including Average Displacement Error (ADE) and Final Displacement Error (FDE). These two metrics are used in a best-of-K manner to report the mode with the smallest distance to the ground-truth future, over all predicted timesteps, and just the final predicted timestep, respectively. Another important metric used is mean Average Precision (mAP). This metric categorizes predicted modes into buckets (e.g., straight, stationary, u-turn, etc.), and punishes mode collapse for overlapping predictions.

VIII. Results

A. Distribution Shift Results

In Figure 4, we highlight some examples of scenarios identified in $\mathcal{S}_{ID}$ and $\mathcal{S}_{OOD}$ for our `Scoring` method described in Section V. The ID scenes contain both simple scenes with few interactions, as well as moderately safety-relevant scenes with lane changes and intersections. The OOD scenes appear significantly more safety-relevant, with more diverse maneuvers, such as u-turns, larger and more

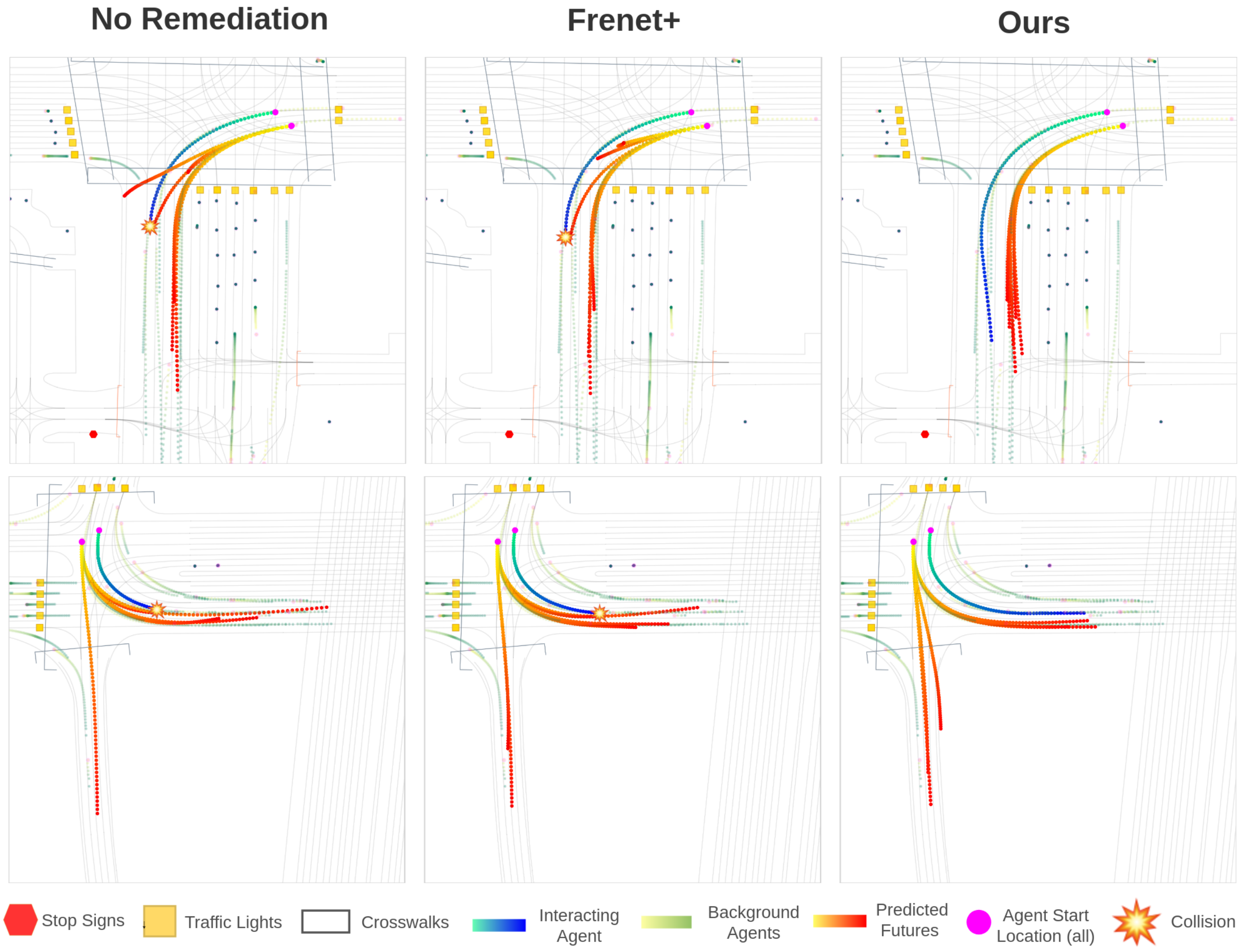


Fig. 5: Qualitative examples of remediation approaches applied to MTR across two distinct scenarios. Trajectories progress from the pink starting points.

dangerous intersections, and many more VRUs navigating alongside vehicles.

Our quantitative results for the trajectory prediction experiments are summarized in Table II. The metric values reported are averaged over the three classes of vehicles, pedestrians, and cyclists. The Δ_{val} value in the final column indicates the increase in collision rate in the OOD test value compared to the ID validation value. In the `Uniform` split, as expected, results between $\mathcal{S}_{ID}$ and $\mathcal{S}_{OOD}$ are quite similar. For the `Clusters` [18] split, we note that while a slight drop in metric performance for ADE / FDE and mAP occurred, the collision rate actually *decreased* from validation to test. We suspect this is because the domains identified by this strategy have no sense of safety-criticality, affirming the importance of using such metrics when selecting scenes. Finally, our `Scoring` strategy resulted in the largest increase in collision rates between $\mathcal{S}_{ID}$ and $\mathcal{S}_{OOD}$, both in terms of absolute value and percentage change. This increase occurs in both the ground truth tracks, as well as in our tested methods, more than doubling the in-distribution rates.

B. Robust Trajectory Prediction Results

We show our remediation experiment results in Table III. Our proposed method was the most effective in reducing collisions for the tested models, as shown by the Δ_{test} values. For MTR in particular, we observe the test collision rates are lowered by 14%, while for A-VRNN, the rates decrease by 6%. This resulted in an average decrease of 10%, reducing the gap to the ground truth collision rate. However, our method does result in a slight decrease in performance on other metrics for MTR. This is likely because MTR has an existing CE loss to select the best mode based on these other metrics, meaning the collision loss objective is in contention with its original objective.

Furthermore, the Frenet+ strategy [18] appeared ineffective in remediating the drop in performance on the `Scoring` data split. We suspect this is due to the presence of more object types than just vehicles; cyclists and pedestrians are often not in lanes, so incorporating such lane information may have been more harmful than beneficial. Additionally, even for vehicles following well-defined lanes, the Frenet+

TABLE II: Distribution shift experiments in WOMD [5]. ADE / FDE is in meters. Δ_{val} is the change in test collision rate (CR) from the corresponding val CR. A more drastic **increase** is better.

Data Split	**Method**	**Validation Set (In-Distribution)**			**Testing Set (Out-of-Distribution)**		
		ADE / FDE	mAP	CR	ADE / FDE	mAP	CR (Δ_{val})
Uniform	GT	- / -	-	0.008	- / -	-	0.009 (+12.5%)
	MTR [27]	0.73 / 1.58	0.30	0.062	0.73 / 1.59	0.31	0.061 (−1.60%)
	A-VRNN [24]	1.80 / 4.63	0.06	0.057	1.82 / 4.67	0.06	0.058 (+1.80%)
Clusters [18]	GT	- / -	-	0.009	- / -	-	0.007 (−22.2%)
	MTR	0.69 / 1.50	0.35	0.060	0.71 / 1.55	0.33	0.051 (−15.0%)
	A-VRNN	1.79 / 4.59	0.08	0.062	1.82 / 4.70	0.07	0.049 (−21.0%)
Scoring (Ours)	GT	- / -	-	0.005	- / -	-	**0.017** (+**240**%)
	MTR	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	**0.100** (+**127**%)
	A-VRNN	1.99 / 5.26	0.05	0.042	2.13 / 5.55	0.05	**0.099** (+**136**%)

GT: Ground truth tracks

TABLE III: Robust trajectory prediction experiments in WOMD [5]. ADE / FDE is in meters. Δ_{test} is the change in test CR from the *un-remediated* test CR for each method. A more drastic **decrease** is better.

Data Split	**Method**	**Validation Set (In-Distribution)**			**Testing Set (Out-of-Distribution)**		
		ADE / FDE	mAP	CR	ADE / FDE	mAP	CR (Δ_{test})
Scoring (Ours)	GT	- / -	-	0.005	- / -	-	0.017 (-)
	MTR	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	0.100 (-)
	MTR + F+ [18]	0.73 / 1.59	0.32	0.043	0.75 / 1.59	0.30	0.099 (−1.00%)
	MTR + Ours	0.83 / 1.80	0.25	0.037	0.89 / 1.91	0.22	**0.086** (−**14.0**%)
	A-VRNN	1.99 / 5.26	0.05	0.042	2.13 / 5.55	0.05	0.099 (-)
	A-VRNN + F+	2.05 / 5.24	0.06	0.041	2.23 / 5.73	0.06	0.103 (+4.04%)
	A-VRNN + Ours	1.76 / 4.61	0.06	0.039	1.91 / 4.94	0.06	**0.093** (−**6.06**%)

GT: Ground truth tracks, **F+**: Frenet+ Strategy [18]

strategy can still incur collisions, particularly at intersections and unprotected turns.

To gain further insight into the benefits of both the Frenet+ strategy and our remediation approach, we provide qualitative examples in Figure 5 using MTR as the prediction model. In these scenarios, the prediction with no remediation results in future modes that collide with an external agent. Meanwhile, the Frenet+ strategy is able to better stay in lanes than the un-remediated approach but still results in collisions. Finally, our remediation approach is able to avoid collisions, while still providing reasonable mode diversity and lane conformance.

C. Ablation Studies

As shown in Table IV, we performed a distribution shift ablation study focusing on the five variations of our scoring strategy discussed in Section V. We utilized MTR as it is the best model according to traditional metrics. Our full method, with asymmetric combined scoring, resulted in the largest increase in collision rate, while still incurring a moderate increase in the other metrics. This result confirms our hypothesis from Section V-B that our counterfactual probing technique indeed captures a fuller spectrum of safety-relevant scenes.

We also performed an ablation study focusing on aspects of our remediation strategy, as shown in Table V. While the collision loss objective alone was quite effective, the best performance was achieved utilizing our full approach, incorporating the scores as part of the models' inductive biases and loss weights as well.

IX. Conclusion

Developing autonomous driving trajectory prediction models through real-world datasets, such as WOMD, is often considered insufficient for ensuring robustness and safety. While such datasets provide realistic recorded scenarios, they rarely contain truly safety-relevant scenarios, falling victim to the "curse-of-rarity." Still, we proposed to further characterize these datasets and find hidden safety-relevant scenarios therein. We thus provided a versatile scenario characterization approach to score scenarios by a hierarchical combination of complementary individual and social features. By performing a counterfactual probe, emulating how a distracted agent may operate, we extended the spectrum of safety-relevance to additionally find hidden risky scenarios, without requiring unrealistic simulation or dangerous real-world testing.

TABLE IV: Scoring strategy ablation study. Results are from using MTR [27] on WOMD [5]. ADE / FDE is in meters. Δ_{val} is the change in test CR from val for the given method. The best distribution shift result is **bolded**.

Ablation Name	Scoring Strategy			Validation Set (In-Distribution)			Testing Set (Out-of-Distribution)		
	GT	FE	AS	ADE / FDE	mAP	CR	ADE / FDE	mAP	CR (Δ_{val})
Ground Truth	✓	-	-	0.72 / 1.57	0.32	0.041	0.75 / 1.64	0.29	0.088 (+115%)
Future Extrapolated	-	✓	-	0.73 / 1.61	0.33	0.046	0.72 / 1.55	0.31	0.097 (+111%)
Combined	✓	✓	-	0.73 / 1.60	0.32	0.048	0.74 / 1.59	0.29	0.098 (+104%)
Asymmetric	-	✓	✓	0.73 / 1.61	0.33	0.044	0.73 / 1.58	0.30	0.099 (+125%)
Asymmetric Combined	✓	✓	✓	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	**0.100 (+127%)**

GT: Ground truth, **FE**: Future extrapolated, **AS**: Asymmetric scoring.

TABLE V: Remediation strategy ablation study based on our proposed approach in Section VI-B utilizing MTR [27] on WOMD [5]. ADE / FDE is in meters. Δ_{test} is the change in test CR from the *un-remediated* MTR test CR.

Ablation Name	Remediation		Validation Set (In-Distribution)			Testing Set (Out-of-Distribution)		
	SC	CL	ADE / FDE	mAP	CR	ADE / FDE	mAP	CR (Δ_{test})
MTR [27]	-	-	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	0.100 (-)
MTR + Ours (SC only)	✓	-	0.74 / 1.63	0.31	0.046	0.74 / 1.61	0.29	0.103 (+3.00%)
MTR + Ours (CL only)	-	✓	0.81 / 1.77	0.27	0.038	0.88 / 1.92	0.23	0.093 (−7.00%)
MTR + Ours (Full)	✓	✓	0.83 / 1.80	0.25	0.037	0.89 / 1.91	0.22	**0.086 (−14.0%)**

SC: Score incorporation, **CL**: Collision loss objective.

Under a distribution shift setting where the most safety-relevant scenes were held out as out-of-distribution, we demonstrated that both ground truth, as well as our evaluated trajectory prediction models, incurred a significant increase in collision rates. We further contributed a remediation strategy, achieving a 10% average reduction in prediction collision rates.

Although our remediation strategy proved successful in reducing the test collision rate, the drop in performance was not remediated completely. Incorporating test-time refinement and collaborative sampling techniques, as highlighted in contemporaneous work, could prove a fruitful direction in improving this strategy further [4]. Another interesting future direction of this work would be to utilize our scoring strategy to assess safety-critical scenarios generated in simulation along the axes of realism, frequency, and type of safety-relevance created. Overall, we argue that trajectory prediction datasets can still be utilized in assessing safety in autonomous driving, and encourage future work to further this direction.

REFERENCES

[1] X. Guo and Y. Zhang, "Maturity in automated driving on public roads: a review of the six-year autonomous vehicle tester program," *Transportation research record*, vol. 2676, no. 11, pp. 352–362, 2022.

[2] J. Francis, B. Chen, S. Ganju, S. Kathpal, J. Poonganam, A. Shivani, V. Vyas, S. Genc, I. Zhukov, M. Kumskoy, *et al.*, "Learn-to-race challenge 2022: Benchmarking safe learning and cross-domain generalisation in autonomous racing," *arXiv preprint arXiv:2205.02953*, 2022.

[3] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, *et al.*, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, 2023.

[4] P. Kothari and A. Alahi, "Safety-compliant generative adversarial networks for human trajectory forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4251–4261, 2023.

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

[6] W. Ding, B. Chen, M. Xu, and D. Zhao, "Learning to collide: An adaptive safety-critical scenarios generating method," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2243–2250.

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

[8] S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, and H. X. Liu, "Dense reinforcement learning for safety validation of autonomous vehicles," *Nature*, vol. 615, no. 7953, pp. 620–627, 2023.

[9] N. Webb, D. Smith, C. Ludwick, T. Victor, Q. Hommes, F. Favaro, G. Ivanov, and T. Daniel, "Waymo's safety methodologies and safety readiness determinations," *arXiv preprint arXiv:2011.00054*, 2020.

[10] W. Huang, K. Wang, Y. Lv, and F. Zhu, "Autonomous vehicles testing methods review," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 163–168.

[11] J. Francis, N. Kitamura, F. Labelle, X. Lu, I. Navarro, and J. Oh, "Core challenges in embodied vision-language planning," *Journal of Artificial Intelligence Research*, vol. 74, pp. 459–515, 2022.

[12] P. Huang, X. Zhang, Z. Cao, S. Liu, M. Xu, W. Ding, J. Francis, B. Chen, and D. Zhao, "What went wrong? closing the sim-to-real gap via differentiable causal discovery," in *Conference on Robot Learning*. PMLR, 2023, pp. 734–760.

[13] D. Xu, Y. Chen, B. Ivanovic, and M. Pavone, "Bits: Bi-level imitation for traffic simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2929–2936.

[14] Y. Cao, D. Xu, X. Weng, Z. Mao, A. Anandkumar, C. Xiao, and M. Pavone, "Robust trajectory prediction against adversarial attacks," in *Conference on Robot Learning*. PMLR, 2023, pp. 128–137.

[15] S. Suo, K. Wong, J. Xu, J. Tu, A. Cui, S. Casas, and R. Urtasun, "Mixsim: A hierarchical framework for mixed reality traffic simulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9622–9631.

[16] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[17] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 3145–3153.

[18] L. Ye, Z. Zhou, and J. Wang, "Improving the generalizability of trajectory prediction models with frenet-based domain normalization," *arXiv preprint arXiv:2305.17965*, 2023.

[19] M. Itkina and M. Kochenderfer, "Interpretable self-aware neural networks for robust trajectory prediction," in *Conference on Robot Learning*. PMLR, 2023, pp. 606–617.

[20] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, and L. Eckstein, "The exid dataset: A real-world trajectory dataset of highly interactive highway scenarios in germany," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 958–964.

[21] C. Glasmacher, R. Krajewski, and L. Eckstein, "An automated analysis framework for trajectory datasets," *arXiv preprint arXiv:2202.07438*, 2022.

[22] A. Sadat, S. Segal, S. Casas, J. Tu, B. Yang, R. Urtasun, and E. Yumer, "Diverse complexity measures for dataset curation in self-driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8609–8616.

[23] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *The International Journal of Robotics Research*, vol. 39, no. 8, pp. 895–935, 2020.

[24] A. Monti, A. Bertugli, S. Calderara, and R. Cucchiara, "Dag-net: Double attentive graph neural network for trajectory forecasting," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 2551–2558.

[25] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive GAN for predicting paths compliant to social and physical constraints," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 1349–1358.

[26] S. H. Park, G. Lee, J. Seo, M. Bhat, M. Kang, J. Francis, A. Jadhav, P. P. Liang, and L.-P. Morency, "Diverse and admissible trajectory forecasting through multimodal context understanding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 282–298.

[27] 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*, vol. 35, pp. 6531–6543, 2022.

[28] M. Bhat, J. Francis, and J. Oh, "Trajformer: Trajectory prediction with local self-attentive contexts for autonomous driving," *arXiv preprint arXiv:2011.14910*, 2020.

[29] X. Tang, S. S. Eshkevari, H. Chen, W. Wu, W. Qian, and X. Wang, "Golfer: Trajectory prediction with masked goal conditioning mnm network," *arXiv preprint arXiv:2207.00738*, 2022.

[30] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: human trajectory prediction in crowded spaces," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 961–971.

[31] S. Konev, K. Brodt, and A. Sanakoyeu, "Motioncnn: A strong baseline for motion prediction in autonomous driving," 2022.

[32] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," *CoRR*, vol. abs/2103.14023, 2021.

[33] J. Ngiam, V. Vasudevan, B. Caine, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, *et al.*, "Scene transformer: A unified architecture for predicting future trajectories of multiple agents," in *International Conference on Learning Representations*, 2021.

[34] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 541–556.

[35] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 525–11 533.

[36] E. Weng, H. Hoshino, D. Ramanan, and K. Kitani, "Joint metrics matter: A better standard for trajectory forecasting," *CoRR*, vol. abs/2305.06292, 2023.

[37] S. Saadatnejad, M. Bahari, P. Khorsandi, M. Saneian, S.-M. Moosavi-Dezfooli, and A. Alahi, "Are socially-aware trajectory prediction models really socially-aware?" *Transportation research part C: emerging technologies*, vol. 141, p. 103705, 2022.

[38] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 159–15 168.

[39] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," in *European Conference on Computer Vision*. Springer, 2022, pp. 335–352.

[40] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a frenet frame," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 987–993.

[41] P. Kothari, D. Li, Y. Liu, and A. Alahi, "Motion style transfer: Modular low-rank adaptation for deep motion forecasting," in *Conference on Robot Learning*. PMLR, 2023, pp. 774–784.

[42] 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.

[43] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[44] X. Zhang, J. Tao, K. Tan, M. Törngren, J. M. G. Sánchez, M. R. Ramli, X. Tao, M. Gyllenhammar, F. Wotawa, N. Mohan, *et al.*, "Finding critical scenarios for automated driving systems: A systematic literature review," *arXiv preprint arXiv:2110.08664*, 2021.

[45] H. Weber, J. Bock, J. Klimke, C. Rösener, J. Hiller, R. Krajewski, A. Zlocki, and L. Eckstein, "A framework for definition of logical scenarios for safety assurance of automated driving," *Traffic Injury Prevention*, vol. 20, pp. S65–S70, 06 2019.

[46] J. Schmidt, J. Jordan, D. Raba, T. Welz, and K. Dietmayer, "Meat: Maneuver extraction from agent trajectories," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 1810–1816.

[47] A. Guha, R. Lei, J. Zhu, X. Nguyen, and D. Zhao, "Robust unsupervised learning of temporal dynamic vehicle-to-vehicle interactions," *Transportation research part C: emerging technologies*, vol. 142, p. 103768, 2022.

[48] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.

[49] K. Vogel, "A comparison of headway and time to collision as safety indicators," *Accident Analysis & Prevention*, vol. 35, no. 3, pp. 427–433, 2003.

[50] J. Shen and G. Yang, "Crash risk assessment for heterogeneity traffic and different vehicle-following patterns using microscopic traffic flow data," *Sustainability*, vol. 12, no. 23, p. 9888, 2020.

[51] X. Zhou, O. Wu, W. Zhu, and Z. Liang, "Understanding difficulty-based sample weighting with a universal difficulty measure," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 68–84.

[52] Y. Wang, Z. Han, Y. Xing, S. Xu, and J. Wang, "A survey on datasets for the decision making of autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, 2024.

[53] M. Liu, E. Yurtsever, X. Zhou, J. Fossaert, Y. Cui, B. L. Zagar, and A. C. Knoll, "A survey on autonomous driving datasets: Data statistic, annotation, and outlook," *arXiv preprint arXiv:2401.01454*, 2024.

[54] P. Kothari, S. Kreiss, and A. Alahi, "Human trajectory forecasting in crowds: A deep learning perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7386–7400, 2021.