# Measuring What Matters: Scenario-Driven Evaluation for Trajectory Predictors in Autonomous Driving

**Longchao Da[1], David Isele[2], Hua Wei[1], Manish Saroya[2*]**

[1]Arizona State University
[2]Honda Research Institute, USA
longchao@asu.edu

## Abstract

Being able to anticipate the motion of surrounding agents is essential for the safe operation of autonomous driving systems in dynamic situations. While various methods have been proposed for trajectory prediction, the current evaluation practices still rely on error-based metrics (e.g., ADE, FDE), which reveal the accuracy from a post-hoc view but ignore the actual effect the predictor brings to the self-driving vehicles (SDVs), especially in complex interactive scenarios: a high-quality predictor not only chases accuracy, but should also captures all possible directions a neighbor agent might move, to support the SDVs' cautious decision-making. Given that the existing metrics hardly account for this standard, in our work, we propose a comprehensive pipeline that adaptively evaluates the predictor's performance by two dimensions: accuracy and diversity. Based on the criticality of the driving scenario, these two dimensions are dynamically combined and result in a final score for the predictor's performance. Extensive experiments on a closed-loop benchmark using a real-world dataset show that our pipeline yields a more reasonable evaluation than traditional metrics by better reflecting the correlation of the predictors' evaluation with the autonomous vehicles' driving performance. This evaluation pipeline shows a robust way to select a predictor that potentially contributes most to the SDV's driving performance.

## 1 Introduction

Trajectory prediction is a fundamental component of autonomous driving systems. It allows the planner to anticipate the future movements of surrounding agents, including vehicles, pedestrians, and cyclists, enabling proactive and safe decision-making (Cui et al. 2021). A typical self-driving vehicle (SDV) pipeline consists of perception, planning, and control modules (Rosique et al. 2019). The perception module detects nearby agents and uses a *trajectory predictor* to forecast their possible future behaviors. The planner then generates the ego vehicle's trajectory based on these predictions, traffic rules, and map information. High-quality predictions are crucial to ensure safety, efficiency, and passenger comfort.

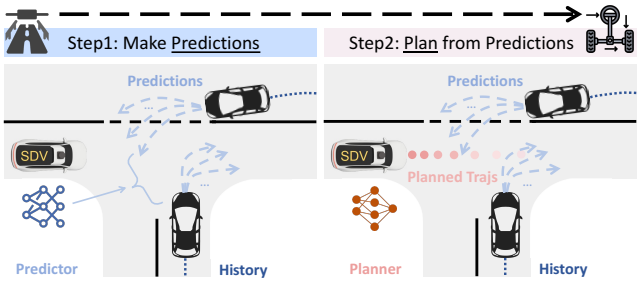


Figure 1: Illustration of how the predictor informs the planner in an SDV system. From the scene observation to the final control, it highlights two steps: the *predictor* estimates future trajectories of surrounding agents, then the *planner* leverages such information to plan for ego vehicle's future trajectory.

Trajectory prediction has been widely studied, with approaches spanning from physics-based predictors like Constant Velocity (CV) (Schöller et al. 2020; Isele et al. 2024) to the Learning-based methods, which have significantly improved prediction by modeling interactions and multimodality (Girgis et al. 2021; Nayakanti et al. 2022). The Transformer-based predictors, such as MTR (Shi et al. 2022) can naturally incorporate high-definition map context during trajectory forecasting, and recent graph-based models, e.g., LaneRCNN (Zeng et al. 2021), Path-Aware Graph Attention (Da and Zhang 2022), and GOHOME (Gilles et al. 2022), explicitly encode HD map structure to further enhance the plausibility of their multimodal outputs.

Given various methodologies, it becomes a realistic question of how to select predictors to provide real-world driving planners with a reliable reference (Da et al. 2024). Most prevalent strategies rely on displacement metrics, such as *Average Displacement Error (ADE)* and *Final Displacement Error (FDE)*, which measure the distance between predictions and ground truth. However, the predictor is not an isolated module: its output guides the planner, affecting the vehicle's decisions about braking, accelerating, or changing lanes. If evaluation criteria fail to capture meaningful differences among predictors that cause downstream changes, predictors that perform well on benchmark metrics may still lead to unsafe or inefficient driving behaviors in practice.

*Corresponding author: manish_saroya@honda-ri.com. Work done during the internship of Longchao Da at the HRI, San Jose.

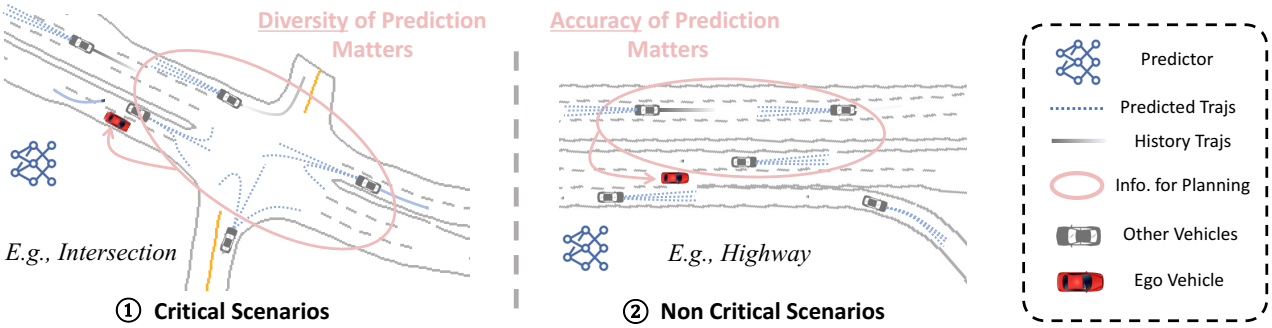


Figure 2: The illustration of predictor evaluation based on scenario criticality. As shown in the figure, in ① critical situations, such as multiple vehicles interacting in the crossing, prediction diversity is more favored to make informative decisions, while on the right side ②, in simple scenarios like highways, the vehicles are moving relatively static, the straightforward accuracy is favored. Only three modalities are shown for presentation; scenarios are from a real-world dataset (Caesar et al. 2020).

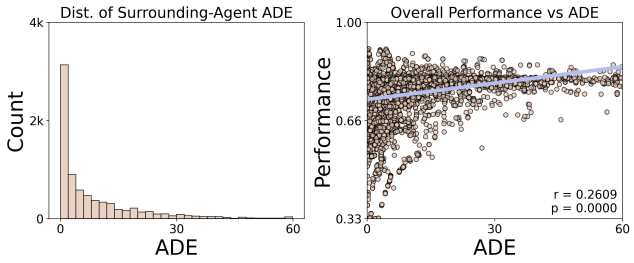


Figure 3: The correlation analysis between ADE and driving performance. We could observe that the prediction's ADE shows a weak and non-insightful relationship with the actual driving performance. i.e., given a predictor that is evaluated by an error-based metric, low error does not necessarily mean a better driving performance.

In our preliminary study [1], we verified the above error-based measures indeed fall short in representing such downstream impact of predictors to planners (Weng et al. 2023; Phong et al. 2023; Shridhar et al. 2020): as shown in Figure. 3, ADE hardly provides meaningful insight: low displacement error doesn't imply improved SDV's driving performance, vice versa. Given this, we conclude the first challenge: solely relying on error-based measures is not enough to quantify the predictor's performance, considering the real impact on SDV.

Under the analysis of these SDVs' driving scenarios and trajectories, it appears that an important factor neglected by the displacement error is the complexity of the scenario. As shown in Figure 2, in relatively simple scenarios, such as highway driving, where surrounding vehicles and the ego share a steady uniform motion, the predictors only need to extrapolate future paths accurately from past observations. By contrast, in complex scenarios, such as intersections, the requirement for the predictor is not only to cover the correct trajectory in its prediction, but also to anticipate every plausible maneuver a neighboring agent might take. Such comprehensive coverage is essential for planners to make early, conservative decisions under uncertainty (Grewal, Tonella, and Stocco 2024). Since real-world systems lack access to real-time ground truth feedback, and high error tolerance can be dangerous, generating a diverse ensemble of possible futures becomes critical. Thus the second challenge arises as: How to reasonably quantify the prediction diversity that helps the planner with rich information to make safer, robust decision-making in challenging scenarios.

Recent works introduced diversity and uncertainty-aware metrics, such as Average Minimum Volume (AMV), energy scores (Shahroudi, Lepson, and Kull 2024), and lane-aware distances (Greer, Deo, and Trivedi 2021), to encourage predictors to produce multiple plausible futures. However, these metrics are typically computed in an *open-loop* setting, that do not account for how prediction errors propagate through the planning pipeline. A few studies have begun to explore the closed-loop evaluation (Phong et al. 2023), but they apply the same criteria regardless of context. Consequently, a scenario-aware evaluation framework that dynamically weights diversity versus accuracy based on traffic context is necessary.

In this paper, we verify that conventional metrics for trajectory prediction fail to reflect the planner's performance. Then, to resolve this, we propose `ED-Eva` that balances the **E**rror metric and the **D**iversity adaptively by the scenarios, in a closed-loop fashion. Specifically, we design an adaptive classifier that takes vehicles' interaction graph as input and outputs a weighting factor to determine the importance of prediction *diversity* and *accuracy*. To quantify prediction diversity in a robust and interpretable way, we propose the *GMM-Area Diversity (GAD)*. It measures the spatial spread of prediction by fitting a Gaussian Mixture Model (GMM) and computing the area of the resulting uncertainty ellipse, and identifies how well the predictor represents multimodal futures. Experiments on multiple trajectory predictors and planners show `ED-Eva` achieves higher correlation with real driving outcomes compared to error-based evaluations, revealing its potential for a robust evaluation system.

In summary, the main contributions of this paper are as follows:

- We identify a critical gap in existing predictor evaluation

[1] In total of 9059 real-world scenarios were tested.

methods, demonstrating that error-based metrics fail to reflect downstream planner performance.

- We propose a *scenario-aware evaluation framework*: `ED-Eva` that adaptively balances prediction **E**rror and **D**iversity during **Eva**luation, aligning predictor performance with planning outcomes.
- We introduce a novel metric named GMM-Area Diversity (GAD), to robustly quantify prediction spread.
- Comprehensive experiments across multiple predictors and planners show that our framework better correlates with downstream planner performance.

## 2 Related Work

The prediction quality is evaluated by various metrics. Existing methods can be broadly categorized into two groups: *accuracy-based metrics* and *diversity or uncertainty-aware metrics*. Furthermore, we discuss the preliminary work identifying the criticality of scenarios and how our method differs.

**Error-based Prediction Evaluation:** Error-based metrics, such as *Average Displacement Error (ADE)* and *Final Displacement Error (FDE)*, measure geometric closeness between predicted and ground-truth trajectories (Zhang et al. 1988). These metrics are widely used in trajectory prediction benchmarks due to their simplicity. Several variants have been proposed to account for the multimodal nature of trajectory prediction. Notably, *aveADE*/*aveFDE* average errors across all predicted modes (Sun, Guo, and Wang 2023), while *minADE*/*minFDE* report the best-case error (Li et al. 2024). However, recent studies have shown error-based measures are insufficient for autonomous driving systems by the failure to capture multi-agent interactions (Weng et al. 2023), to account for the planner's downstream decision-making (Phong et al. 2023), and the lack of alignment with system-level objectives such as safety and comfort (Shridhar et al. 2020).

**Diversity and Uncertainty Evaluation:** Diversity and uncertainty-aware evaluation methods encourage predictors to produce multiple plausible future trajectories rather than collapsing to a single mode. Metrics such as *Average Minimum Volume (AMV)* (Mohamed et al. 2022), *energy scores* (Shahroudi, Lepson, and Kull 2024), and scalibration-based assessments (Carrasco Limeros et al. 2024) could measure how well predictions capture the multimodal nature of traffic behaviors. Lane-aware distance metrics (Schmidt et al. 2023) and auxiliary losses (Greer, Deo, and Trivedi 2021) have also been proposed to promote spatial plausibility and social compliance. However, these metrics are typically evaluated independently of the downstream planner and do not consider how prediction diversity or spread impacts planning outcomes like safety or comfort.

**Scenario Criticality Identification:** Understanding the criticality of scenarios is important to decide the risks of taking certain actions (Zhang et al. 2022). There are early works applying Long-short term memory (LSTM) for scene understanding and risk assessment (Wang et al. 2021), while later graph-based interaction modeling provides a better capability in dynamic scene change capturing (Malawade et al. 2022; Yu et al. 2021; Wang et al. 2024). Even though there are multiple works exploring the scene risk and criticality assessment, only few works consider the prediction evaluation based on scenario semantics (Sánchez et al. 2022) and (Chen, Pourkeshavarz, and Rasouli 2024). But they rely on predefined scenario categorizations prior to evaluation, which limits the adaptability of the evaluation method itself, due to the lack of automatic, context-aware criticality assessment. Besides, they do not consider the closed-loop impact from the prediction to the planner's action.

Unlike existing work, this paper focuses on the closed-loop of predictor and planner, proposes a scenario-adaptive evaluation framework that dynamically adjusts the preference between error-based measure and diversity of predictions.

## 3 Methodology

A good predictor should help to improve the SDV's driving performance, and a robust evaluation method should identify the correlation between the prediction and driving performance. In this section, we first define our problem for evaluating trajectory predictors based on their impact on closed-loop driving performance, then propose a concrete evaluation framework `ED-Eva`, that simultaneously considers geometric error and diversity based on the scenarios.

### Problem Formulation

We consider an autonomous driving system [2] that operates within a set of scenarios $\mathcal{S}$. Each scenario $s \in \mathcal{S}$ specifies the initial conditions, such as road geometry, agent states, and traffic configurations. Let $\mathcal{D}$ denote a distribution over scenarios, representing the conditions under which we perform the evaluation. We define the following components:

**Definition 1 (Predictor and Predicted Trajectory)** *A trajectory predictor $P_i$ takes as input the observation of scenario $s$, including the ego vehicle's state, surrounding agents, and environmental context, and produces* predicted trajectories*:*

$$P_i : s \mapsto \hat{\tau}^{(i)} = \left(\hat{\tau}_1^{(i)}, \ldots, \hat{\tau}_M^{(i)}\right) \tag{1}$$

*where $\hat{\tau}_j^{(i)}$ is the predicted trajectory of agent $j$, and $i$ denotes the prediction comes from predictor $P_i$.*

We denote a *fixed* motion-planning policy $\pi$ as a planner by:

**Definition 2 (Planner)** *A fixed planner $\pi$ generates an executed ego trajectory $\tau^{\text{ego}}$ conditioned on the current scenario and the predicted trajectories:*

$$\tau^{\text{ego}} = \pi\left(s, \hat{\tau}^{(i)}\right) \tag{2}$$

*where $\hat{\tau}^{(i)}$ includes the trajectories of surrounding agents, the planner will perform maneuvers on vehicles based on the provided information.*

To measure how the ego planner performs after integrating the predictions, we define a *driving performance* function:

**Definition 3 (Driving Performance)** *The quality of an executed ego-trajectory is measured by the following function:*

$$R\left(\tau^{\text{ego}}, s\right) \tag{3}$$

[2] *Autonomous driving* and *self driving* are interchangeably used.

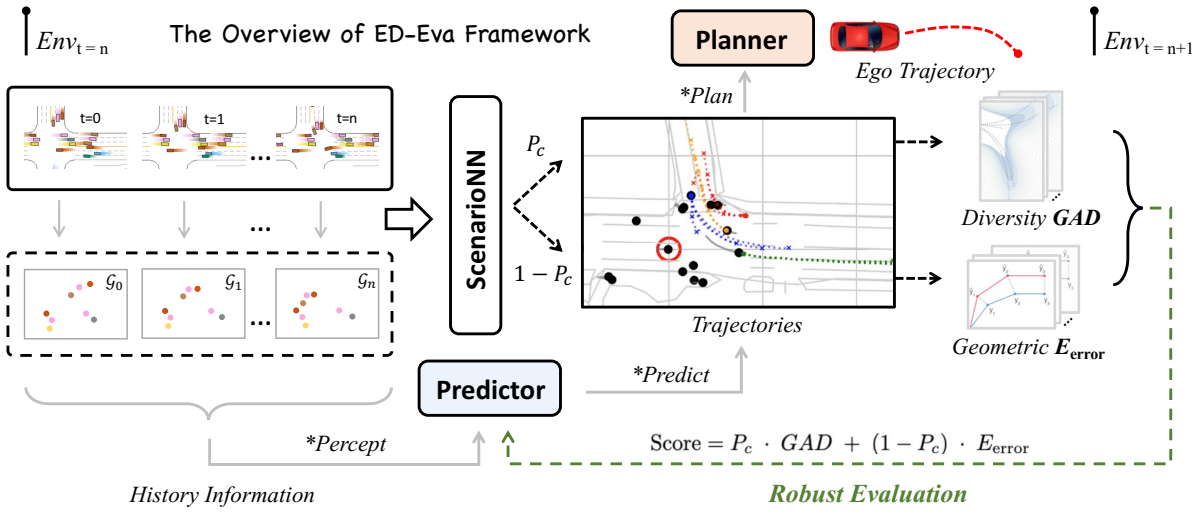


Figure 4: The overview of the proposed framework ED-Eva. At $\text{Env}_{t=n}$, given the $n$ steps of history information, the predictor makes predictions for future trajectories, and the planner decides the ego maneuvers based on predictions, then the system moves to the next step $\text{Env}_{t=n+1}$. Our evaluation exists in this loop, we first construct the spatial-temporal graph features and feed them to ScenarioNN, the network will provide critical probability $P_c$ based on complexity of the scenario, then we employ the diversity measure GAD to capture the prediction's reasonable spread and the geometric $E_{\text{error}}$ to quantify their displacement error, they are dynamically weighted by the context of the traffic and come to the final evaluation score that could reflect the predictor's influence to the overall driving performance.

*which jointly captures safety (reflected by collisions), comfort (reflected by jerk), and efficiency (travel time) as in (Phong et al. 2023).*

Based on the Eq. 2, the driving performance induced by $P_i$ in scenario $s$ can be written as:

$$R(P_i, s) \;=\; R\Big(\pi\big(\hat{\tau}^{(i)},\, s\big),\, s\Big) \tag{4}$$

Under a distribution $\mathcal{D}$ over scenarios, the *expected driving performance* of predictor $P_i$ is:

$$\overline{R}(P_i) \;=\; \mathbb{E}_{s\sim\mathcal{D}}\big[R(P_i, s)\big] \tag{5}$$

which quantifies how $P_i$ contributes to the SDV's overall performance that consists of safety, comfort, and efficiency. We formalize the problem as below:

**Problem 1 (Robust Predictor Evaluation)** *Given a set of Predictors $\{P_i\}$, the goal is to find a measurement $\mathcal{M}$ that evaluates each predictor $P_i$ that reflects its driving performance $R(P_i, s)$, where $\mathcal{M}(P_i)$ is a scalar score for the $i^{th}$ predictor. If we select $P^*$ by:*

$$P^* \;=\; \arg\max_{P_i\in\mathcal{P}} \mathcal{M}(P_i) \tag{6}$$

*Then $P^*$ should* maximize $\overline{R}(P_i)$*, i.e., yield the best expected driving performance among the candidate predictors:*

$$\mathcal{M}(P_i) > \mathcal{M}(P_j) \;\Rightarrow\; \overline{R}(P_i) \;\geq\; \overline{R}(P_j) \tag{7}$$

Since traditional error-based metrics $\mathcal{M}$ (e.g., *ADE*), only measure pointwise errors between predicted and ground-truth trajectories, ignoring how these errors propagate through the planner. Thus, we seek to create a measurement $\mathcal{M}$ that captures *planner–predictor interactions*, resulting in a stronger correlation with real-world driving outcomes.

## Prediction Diversity Measure: GAD

It is important to understand the complexity of the scenario and consider its criticality based on Figure 2. Thus, in this section, we measure the ability to capture the complex future trajectories by quantifying the diversity of its predictions.

We adopt the concept of 'spread' to understand the diversity of a set of $N$ predicted trajectories over a horizon of $T_p$ time-steps. At each time $t$, we fit a two-dimensional Gaussian mixture model (GMM) for each prediction index $n$ to the ensemble of trajectory endpoints $\{(x_{t,i}^{(n)}, y_{t,i}^{(n)})\}_{i=1}^{K}$. Then we could collapse the GMM to a single 2×2 covariance matrix:

$$\Sigma_t^{(n)} = \sum_{k=1}^{K} w_k \left(\mu_k - \bar{\mu}\right)\left(\mu_k - \bar{\mu}\right)^{\top} + \sum_{k=1}^{K} w_k\, C_k, \tag{8}$$

where $w_k, \mu_k, C_k$ are the weight, mean, and covariance of component $k$, and $\bar{\mu} = \sum_k w_k \mu_k$, we then perform an eigen-decomposition as shown below ($Q$ is the orthogonal matrix of eigenvectors from the decomposition process):

$$\Sigma_t^{(n)} = Q \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} Q^{\top}, \qquad \lambda_1 \geq \lambda_2 \geq 0 \tag{9}$$

The lengths of the principal semi-axes of the one-standard-deviation uncertainty ellipse are: $a_1 = \sqrt{\lambda_1}, a_2 = \sqrt{\lambda_2}$. Based on this, we can derive its geometric area as:

$$\mathcal{A}\big(\Sigma_t^{(n)}\big) = \pi\, a_1\, a_2 = \pi\, \sqrt{\lambda_1\, \lambda_2} = \pi \sqrt{\det\big(\Sigma_t^{(n)}\big)} \tag{10}$$

Since each of the diversity measures of a predictor, it contains $\pi$, thus, we could simply drop the constant factor $\pi$ and define the per-prediction, per-time diversity score as below:

$$D\big(\Sigma_t^{(n)}\big) \;=\; \sqrt{\det\big(\Sigma_t^{(n)}\big)} \tag{11}$$

We average over all $N$ trajectories and $T_p$ steps yields the overall *GMM-Area Diversity* (GAD) below:

$$\text{GAD} \;=\; \frac{1}{N\,T_p}\sum_{n=1}^{N}\sum_{t=1}^{T_p}\sqrt{\det\!\big(\Sigma_t^{(n)}\big)} \tag{12}$$

The GAD captures spread along both principal axes rather than only the maximal direction, yielding truly bidirectional sensitivity. By fitting a full-covariance GMM and aggregating component covariances, it smooths over individual outliers for robust summarization. Besides, the required eigen-decomposition of a $2\times2$ matrix is closed-form and incurs negligible computational cost on top of GMM fitting.

### Prediction Displacement Measure $E_{\text{error}}$

In scenarios with low interaction complexity (e.g., highways), simple error-based metrics are often sufficient to gauge predictor performance. We therefore integrate *Error–Based Measure* $E_{\text{error}}$ into our framework. The general idea is to compare each predicted trajectory against the ground-truth future. We denote the set of $N$ predicted trajectories over a horizon of $T_p$ steps: $\{\hat{p}_t^{(n)} \mid t = 1,\dots,T_p;\ n = 1,\dots,N\}$ and the corresponding ground-truth trajectory by $\{p_t^* \mid t = 1,\dots,T_p\}$. The error measure could be represented by a generic function:

$$E_{\text{error}}\big(\{\hat{p}\},\{p^*\}\big) = \frac{1}{N\,T_p}\sum_{n=1}^{N}\sum_{t=1}^{T_p}\big\|\hat{p}_t^{(n)} - p_t^*\big\|. \tag{13}$$

This expression can be instantiated in multiple ways, e.g., by only the final timestep (*FDE*), by the best of $N$ modes (*minADE/minFDE*), by averaging top-$K$ modes (*aveADE*), or other common variants without changing the overall framework. Since our work tried to verify the feasibility of the framework, for simplicity, we adopt *ADE* as the $E_{\text{error}}$.

### Scenario Classifier: `ScenarioNN`

Inspired by work (Malawade et al. 2022), we develop the scenario neural network to predict the critical probability. It builds upon a spatial-temporal graph network that leverages a graph $\mathcal{G}$ to describe the spatial information. It takes the states of the ego vehicle and its nearest $N$ [3] neighbors as input for one graph $\mathcal{G}_t$, and collects a fixed horizon $T = 15$ to capture the temporal features. At each timestep $t \in \{1,\dots,T\}$, the $\mathcal{G}_t$ is constructed by node feature tensor $\mathbf{X} \in \mathbb{R}^{T\times N\times F}$, where $N$ is also the number of nodes and $F$ contains 14 features per node: from 3D position $(x, y, z)$, velocity, acceleration, and a 5-dimensional relative-motion aggregation over its neighbors.

Then, the spatial relations are encoded by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ computed from the 2D positions at $t = 1$:

$$A_{ij} = \frac{\mathbb{I}\{\|p_i^1 - p_j^1\| < d_{\text{th}}\} + \delta_{ij}}{\sum_{k=1}^{N}\big(\mathbb{I}\{\|p_i^1 - p_k^1\| < d_{\text{th}}\} + \delta_{ik}\big)} \tag{14}$$

where $d_{\text{th}} = 5\,\text{m}$ and $\delta_{ij}$ adds self-loops before row-normalization. At each $t$, two graph-convolution layers as:

$$\mathbf{H}_t^{(\ell+1)} = \text{ReLU}\big(\mathbf{A}\,\mathbf{H}_t^{(\ell)}\,\mathbf{W}^{(\ell)}\big), \quad \mathbf{H}_t^{(0)} = \mathbf{X}_t \tag{15}$$

[3] $N$=7 in our setting.

We then mean-pool over the $N$ nodes to obtain a sequence $\{\mathbf{g}_t\}_{t=1}^{T} \subset \mathbb{R}^{d_g}$. The temporal patterns are captured by feeding this sequence into an LSTM of hidden size $d_h = 32$. Denoting the LSTM's final hidden state by $\mathbf{h}_T$, a linear head produces a logit:

$$\ell = \mathbf{w}^\top \mathbf{h}_T + b \tag{16}$$

Based on this, a sigmoid function gives the probability $P_c = \hat{y} = \sigma(\ell)$ that the scenario is "critical." During training, we balance positive (collision-check) and negative samples using weighted sampling and optimize binary cross-entropy. At test time, the probability will be used to weight the GAD and $E_{\text{error}}$ as shown in Eq 17.

### Overview of the `ED-Eva` Framework

As illustrated in Figure 4, our framework unifies: *Scenario Criticality Estimation*, *Diversity and Error Balance* in a closed-loop process. First, the spatio-temporal history of the ego vehicle and its neighbors is extracted into a sequence of graphs $\mathcal{G}_1 \dots \mathcal{G}_t$ and passed to the scenario classifier `ScenarioNN`, which outputs the probability $P_c$ that the current scenario is 'critical' (requiring diversity) and $1 - P_c$ that it is 'simple' scenario (accuracy). Second, the trajectory *predictor* generates $N$ future paths, which are scored independently by the diversity measure GAD and the Error Metric $E_{\text{error}}$. Please note that, in implementation, we take a negative error-based metric value to align the direction with diversity. Finally, these two signals are combined via:

$$\text{Score} = P_c \;\cdot\; GAD \;+\; (1 - P_c) \;\cdot\; E_{\text{error}} \tag{17}$$

This adaptive fusion ensures that in interactive safety-critical scenes, emphasis falls on diversity, and in low-risk scenarios, the focus remains on geometric accuracy.

## 4 Experiment

We conduct experiments to answer the following research questions (RQs):

**RQ1**: How does the `ED-Eva` perform compared to other displacement error-based measures?

**RQ2**: How does GAD metric measure diversity in various trajectory prediction?

**RQ3**: Can `ScenarioNN` distinguish criticality correctly?

**RQ4**: How do each of the components in the framework contribute to the `ED-Eva`?

### Experimental Setup

- **Predictors:** We evaluate four widely-used prediction methods: CV (Schöller et al. 2020) as a baseline predictor, AutoBot (Girgis et al. 2021) for attention-based interaction encoding, MTR (Shi et al. 2022) for explicit multimodal "mode" queries, and Wayformer (Nayakanti et al. 2022) for map-conditioned waypoint forecasting.
- **Planners:** The Frenet-based planner (Werling et al. 2010) is selected for vehicles controls. Specifically, the planner consumes a fixed set of $K = 6$ predicted trajectories (15 time steps) per agent and generates planned trajectories.

| Evaluation Method | Frenet (Planner) + Different Predictors | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MTR (Shi et al. 2022) | | | | Autobot (Girgis et al. 2021) | | | | Wayformer (Nayakanti et al. 2022) | | | |
| | **Efficient** | **Discomfort** | **Unsafety** | **Overall*** | **Efficient** | **Discomfort** | **Unsafety** | **Overall*** | **Efficient** | **Discomfort** | **Unsafety** | **Overall*** |
| *-ADE* | -0.2936 | 0.1821 | 0.1150 | -0.3399 | -0.3052 | 0.3365 | 0.0162 | -0.3737 | -0.3106 | 0.3189 | 0.0545 | -0.3991 |
| *-FDE* | -0.2919 | 0.2544 | 0.0035 | -0.3462 | -0.2988 | 0.3710 | 0.0109 | -0.3966 | -0.3125 | 0.3554 | 0.0116 | -0.4202 |
| *-minADE* | -0.2721 | 0.1931 | 0.1115 | -0.3375 | -0.3005 | 0.3338 | 0.0186 | -0.3708 | -0.3050 | 0.3154 | 0.0581 | -0.3851 |
| *-minFDE* | -0.2407 | 0.2689 | 0.0155 | -0.3404 | -0.2798 | 0.3646 | 0.0103 | -0.3851 | -0.2922 | 0.3496 | 0.0190 | -0.4098 |
| *-aveADE* | -0.2899 | 0.1862 | 0.1175 | -0.3427 | -0.3056 | 0.3363 | 0.0156 | -0.3734 | -0.3107 | 0.3183 | 0.0562 | -0.3991 |
| *-aveFDE* | -0.2867 | 0.2620 | 0.0176 | -0.3567 | -0.2986 | 0.3700 | 0.0081 | -0.3948 | -0.3119 | 0.3558 | 0.0148 | -0.4212 |
| Diversity(AMV) | -0.1075 | -0.0341 | 0.2104 | -0.0823 | -0.2359 | -0.1210 | 0.0029 | -0.1613 | -0.3914 | 0.3105 | 0.0126 | -0.3418 |
| Diversity (GAD) | -0.6227 | 0.3618 | 0.0977 | -0.5905 | 0.1184 | -0.5739 | -0.1900 | 0.6058 | -0.7779 | 0.2772 | -0.1364 | -0.5871 |
| ED-Eva (GAD, ADE) | 0.0357 | -0.1451 | -0.1597 | +0.2200 | -0.0474 | -0.2900 | -0.1235 | +0.2652 | -0.0689 | -0.2186 | -0.1082 | +0.1843 |

Table 1: The correlation analysis on different evaluation methods' results and the Planner's driving performance. The table shows two sets of information: First, the driving performance of the Frenet Planner equipped with three different predictors that run in testing scenarios. Second, during the testing, the predictor's performance was measured by baseline evaluations. The error-based method reported a negative value to align the direction: intuitively, the lower ADE, the better; the higher the overall performance, the better. By negative error results, we simply expect the correlation in the 'Overall*' column, the higher the better. For Discomfort and Unsafety, since Discomfort represents the jerk and Unsafety is from collision, so the negative correlation is expected for ED-Eva, which means the higher the performance our method quantifies, the lower the collision or jerk value is.

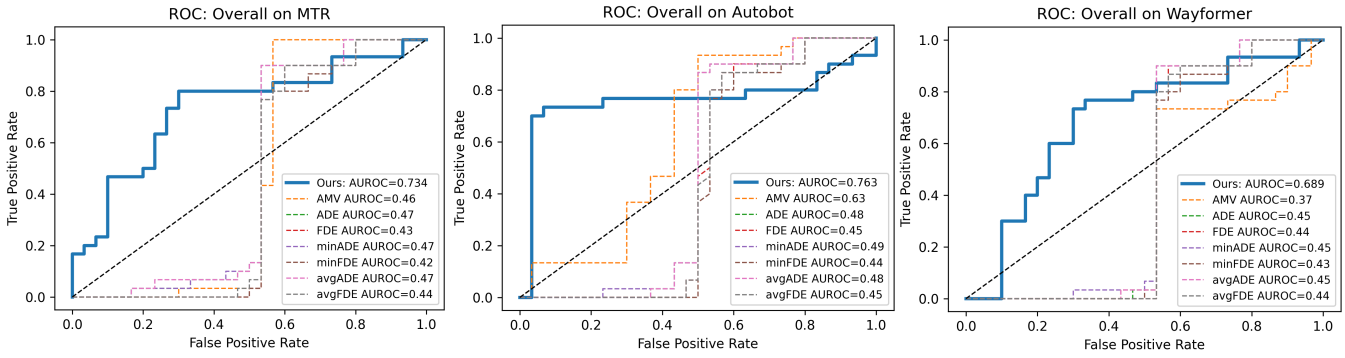


Figure 5: The *AUROC* of our method ED-Eva compared with baseline evaluation metrics. It reflects that our method (blue lines) holds a stronger capability to rank the predictions in different scenarios that best align with their true driving performance score.

- **Basline Evaluation Methods:** *ADE* and *FDE* (best mode), *minADE* and *minFDE*, as well as *aveADE* and *aveFDE* (over 6 modalities), the error-based metrics all apply the same rule that the lower value, the better (↓).
- **Metrics:** Since we investigate the performance of the 'evaluators' based on how much they reflect the driving performance, thus, we apply the Pearson correlation coefficient to reflect their correlation. The performance consists of three dimensions and is aligned to the final performance based on existing work (Phong et al. 2023), the higher the value of the performance, the better (↑). We also use *AUROC* (Stocco et al. 2020) to show the ranking based on the whole test scenario.
- **Dataset and Model Preparation:** The test set we use is the nuscene dataset (Caesar et al. 2020) consists of 9059 driving scenarios. The learning-based predictors are all trained on the Waymo dataset (Sun et al. 2020) for 100 epochs using the *Unitraj* framework (Feng et al. 2024).

## The Overall Performance of ED-Eva (RQ1)

To demonstrate how each predictor evaluation method reveals the overall correlations to the driving performance, we present the results as shown in Table 1. The ED-Eva provides a measure where the higher the evaluation score, the better the performance of a predictor leads to a planner. To make a consistent comparison, we show the negative values for the error-based measure (since intuitively, the higher the error, the worse the performance it leads to). We can observe that only our method consistently and positively correlates with the overall driving performance (as shown in red color compared to those in blue). Besides, we validate that our method can reveal the discomfort and unsafety as well, since they are calculated by jerk and collision, thus the negative correlation is favored and indeed reflected by the result.

Although Pearson's linear correlation is only moderate, since our evaluation really reflects rank consistency rather than strict linearity, we can better illustrate alignment via *AUROC* curves (Figure 5). A higher *AUROC* indicates closer agreement with overall driving performance. Across three different predictor configurations, ED-Eva (blue curve) consistently encloses more area than competing metrics. Note that we do not claim any predictor is 'best' in all scenarios; rather, ED-Eva provides a scenario-aware, robust score for comparing methods on the particular cases that matter.

### Verification on the Diversity Measure (RQ2)

In this section, we conduct a case analysis on the diversity measure method GAD and interpret how it works within the evaluation pipeline. As shown in the Figure 6, we took Autobot and Wayformer as two predictors for the case study; the two predictors were fed with the exact same history observation and are supposed to make a prediction from the red dot. The predictions are in multimodality, as shown in gray dotted lines. Then we apply the GAD to quantify the diversity following Eq. 12, the built GMM distributions are shown in the zoomed-in window, with their GAD value attached, we can observe that the GAD successfully models the spread (right side) and produces higher diversity on the right side.

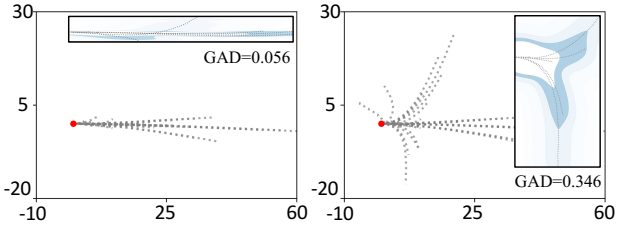


Figure 6: The proposed diversity measure GAD based on the GMM distribution and the ellipse-area quantification. This case shows two prediction outputs of the same observation; the red dot is the last step observation. It shows that the GAD correctly quantifies the right figure with a larger value 0.346, while the left side is more concentrated with GAD = 0.056.

### The Performance of ScenarioNN (RQ3)

In this section, we validate the ScenarioNN 's performance. The scenario criticality predictor is pretrained on our MetaDrive dataset (Li et al. 2022). As Figure 7 shows, it achieves 89% precision on held-out scenarios. Although not perfect, this accuracy already benefits the overall evaluation pipeline, and future work is encouraged to further improve the performance of scenario-understanding. We also demonstrate the case analysis on the right side, at t=15, the overall scenario looks stable, thus the ScenarioNN provides a low critical score of 0.12. At t=17, we observe the top vehicle shows vibration and crosses into another direction of the lane, but it does not affect the overall agent, thus its $P_c$ is still low as 0.34. When it comes to t=20, the red agent speeds up and comes back to the vehicle team, the $P_c$ is predicted as 0.91 with a high likelihood of crashing, which means in such a situation, it is important to predict the diverse directions the vehicles might move to take conservative actions and avoid collision. At t=22, the agent team is approaching the intersection, its predicted $P_c$ is also high, and the predictor should be predicting diverse trajectories one might move to.

### Ablation Study on ED−Eva (RQ4)

We conduct an ablation study to understand how each component contributes to the effectiveness of our method. As in the Figure 8, from left to right, we progressively remove the components. It reflects that if we only remove ScenarioNN,

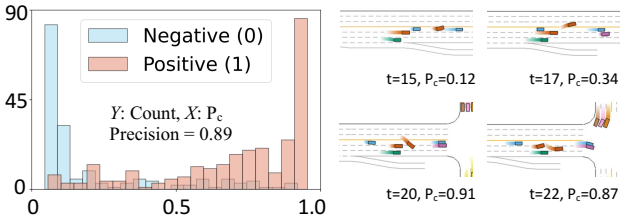


Figure 7: The performance of ScenarioNN, the left figure is the distribution of the predicted probability for $P_c$, the color is the binary groundtruth (critical / non-critical), the scenario network correctly identifies the criticality level with a high precision of 89%. Right side shows a snapshot of the testing scenario along with the frame time and predicted $P_c$.

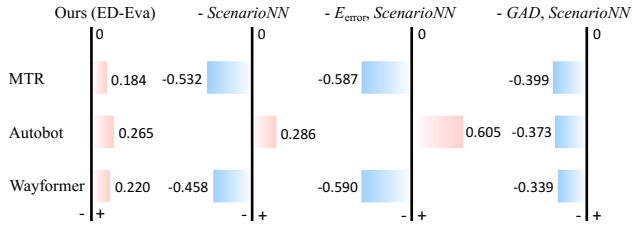


Figure 8: The ablation study of ED−Eva method. The leftmost is the correlation of our method to the overall driving performance across three predictors, the higher the better.

the consistent correlation is altered. Then, if we remove both diversity measure GAD and ScenarioNN, it fails to provide an insights indicator for overall performance. This provides a better understanding that the ScenarioNN is crucial to make reasonable evaluations, while the diversity and accuracy matter in different situations.

## 5 Conclusion

We identified a fundamental mismatch between traditional error-based metrics and the impact of trajectory predictors on downstream planning performance in autonomous driving tasks. To address this, we introduced a Scenario-Driven Evaluation Pipeline that adaptively balances error measure and diversity according to the criticality of the driving scenario. We propose the GMM-Area Diversity (GAD) metric, which robustly quantifies the multimodal spread of the predictions, and we apply a graph-based scenario classifier that determines when the diversity or accuracy should dominate the evaluation. The experiments show that ED−Eva correlates more strongly with actual driving performance than the conventional metrics. It suggests that moving beyond displacement errors toward a more context-adaptive, planner-centric evaluation is essential for selecting predictors that truly enhance self-driving performance. To apply in real-world scenarios, the GAD can be easily bounded by traffic physical factors, such as using the farthest reachable distance under maximum speed regulated by the local policy, and we admit that the design of the classifier can be further enhanced.

## Acknowledgments

The work was partially supported by NSF award #2442477. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

## References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Carrasco Limeros, S.; Majchrowska, S.; Johnander, J.; Petersson, C.; Sotelo, M. Á.; and Fernández Llorca, D. 2024. Towards trustworthy multi-modal motion prediction: Holistic evaluation and interpretability of outputs. *CAAI Transactions on Intelligence Technology*, 9(3): 557–572.

Chen, C.; Pourkeshavarz, M.; and Rasouli, A. 2024. Criteria: a new benchmarking paradigm for evaluating trajectory prediction models for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 8265–8271. IEEE.

Cui, A.; Casas, S.; Sadat, A.; Liao, R.; and Urtasun, R. 2021. Lookout: Diverse multi-future prediction and planning for self-driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16107–16116.

Da, F.; and Zhang, Y. 2022. Path-aware graph attention for hd maps in motion prediction. In *2022 International conference on robotics and automation (ICRA)*, 6430–6436. IEEE.

Da, L.; Jenkins, P.; Schwantes, T.; Dotson, J.; and Wei, H. 2024. Probabilistic offline policy ranking with approximate Bayesian computation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 20370–20378.

Feng, L.; Bahari, M.; Amor, K. M. B.; Zablocki, É.; Cord, M.; and Alahi, A. 2024. Unitraj: A unified framework for scalable vehicle trajectory prediction. In *European Conference on Computer Vision*, 106–123. Springer.

Gilles, T.; Sabatini, S.; Tsishkou, D.; Stanciulescu, B.; and Moutarde, F. 2022. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 international conference on robotics and automation (ICRA)*, 9107–9114. IEEE.

Girgis, R.; Golemo, F.; Codevilla, F.; Weiss, M.; D'Souza, J. A.; Kahou, S. E.; Heide, F.; and Pal, C. 2021. Latent variable sequential set transformers for joint multi-agent motion prediction. *arXiv preprint arXiv:2104.00563*.

Greer, R.; Deo, N.; and Trivedi, M. 2021. Trajectory prediction in autonomous driving with a lane heading auxiliary loss. *IEEE Robotics and Automation Letters*, 6(3): 4907–4914.

Grewal, R.; Tonella, P.; and Stocco, A. 2024. Predicting safety misbehaviours in autonomous driving systems using uncertainty quantification. In *2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*, 70–81. IEEE.

Isele, D.; Gupta, P.; Liu, X.; and Bae, S. 2024. Gaussian lane keeping: A robust prediction baseline. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 3680–3687. IEEE.

Li, L.; Lin, X.; Huang, Y.; Zhang, Z.; and Hu, J.-F. 2024. Beyond minimum-of-N: Rethinking the evaluation and methods of pedestrian trajectory prediction. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, Q.; Peng, Z.; Feng, L.; Zhang, Q.; Xue, Z.; and Zhou, B. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3461–3475.

Malawade, A. V.; Yu, S.-Y.; Hsu, B.; Muthirayan, D.; Khargonekar, P. P.; and Al Faruque, M. A. 2022. Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction. *IEEE Internet of Things Journal*, 9(12): 9379–9388.

Mohamed, A.; Zhu, D.; Vu, W.; Elhoseiny, M.; and Claudel, C. 2022. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *European Conference on Computer Vision*, 463–479. Springer.

Nayakanti, N.; Al-Rfou, R.; Zhou, A.; Goel, K.; Refaat, K. S.; and Sapp, B. 2022. Wayformer: Motion forecasting via simple & efficient attention networks. *arXiv preprint arXiv:2207.05844*.

Phong, T.; Wu, H.; Yu, C.; Cai, P.; Zheng, S.; and Hsu, D. 2023. What truly matters in trajectory prediction for autonomous driving? *Advances in Neural Information Processing Systems*, 36: 71327–71339.

Rosique, F.; Navarro, P. J.; Fernández, C.; and Padilla, A. 2019. A systematic review of perception system and simulators for autonomous vehicles research. *Sensors*, 19(3): 648.

Sánchez, M. M.; Elfring, J.; Silvas, E.; and van de Molengraft, R. 2022. Scenario-based evaluation of prediction models for automated vehicles. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2227–2233. IEEE.

Schmidt, J.; Monninger, T.; Jordan, J.; and Dietmayer, K. 2023. LMR: Lane distance-based metric for trajectory prediction. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, 1–6. IEEE.

Schöller, C.; Aravantinos, V.; Lay, F.; and Knoll, A. 2020. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2): 1696–1703.

Shahroudi, N.; Lepson, M.; and Kull, M. 2024. Evaluation of trajectory distribution predictions with energy score. In *Forty-first International Conference on Machine Learning*.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2022. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35: 6531–6543.

Shridhar, S.; Ma, Y.; Stentz, T.; Shen, Z.; Haynes, G. C.; and Traft, N. 2020. Beelines: Evaluating Motion Prediction Impact on Self-Driving Safety and Comfort. *arXiv preprint arXiv:2011.00393*.

Stocco, A.; Weiss, M.; Calzana, M.; and Tonella, P. 2020. Misbehaviour prediction for autonomous driving systems. In *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 359–371.

Sun, D.; Guo, H.; and Wang, W. 2023. Vehicle trajectory prediction based on multivariate interaction modeling. *IEEE Access*, 11: 131639–131650.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Wang, H.; Lu, B.; Li, J.; Liu, T.; Xing, Y.; Lv, C.; Cao, D.; Li, J.; Zhang, J.; and Hashemi, E. 2021. Risk assessment and mitigation in local path planning for autonomous vehicles with LSTM based predictive model. *IEEE Transactions on Automation Science and Engineering*, 19(4): 2738–2749.

Wang, J.; Malawade, A. V.; Zhou, J.; Yu, S.-Y.; and Al Faruque, M. A. 2024. Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7493–7502.

Weng, E.; Hoshino, H.; Ramanan, D.; and Kitani, K. 2023. Joint metrics matter: A better standard for trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20315–20326.

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

Yu, S.-Y.; Malawade, A. V.; Muthirayan, D.; Khargonekar, P. P.; and Al Faruque, M. A. 2021. Scene-graph augmented data-driven risk assessment of autonomous vehicle decisions. *IEEE Transactions on Intelligent Transportation Systems*, 23(7): 7941–7951.

Zeng, W.; Liang, M.; Liao, R.; and Urtasun, R. 2021. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 532–539. IEEE.

Zhang, G.; Ouyang, R.; Lu, B.; Hocken, R.; Veale, R.; and Donmez, A. 1988. A displacement method for machine geometry calibration. *CIRP Annals*, 37(1): 515–518.

Zhang, X.; Tao, J.; Tan, K.; Törngren, M.; Sánchez, J. M. G.; Ramli, M. R.; Tao, X.; Gyllenhammar, M.; Wotawa, F.; Mohan, N.; et al. 2022. Finding critical scenarios for automated driving systems: A systematic mapping study. *IEEE Transactions on Software Engineering*, 49(3): 991–1026.