2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

CRITERIA: a New Benchmarking Paradigm for Evaluating Trajectory Prediction Models for Autonomous Driving

Changhe Chen[1], Mozhgan Pourkeshavarz[2*], Amir Rasouli[2]

***Abstract*— Benchmarking is a common method for evaluating trajectory prediction models for autonomous driving. Existing benchmarks rely on datasets, which are biased towards more common scenarios, such as cruising, and distance-based metrics that are computed by averaging over all scenarios. Following such a regiment provides a little insight into the properties of the models both in terms of how well they can handle different scenarios and how admissible and diverse their outputs are. There exist a number of complementary metrics designed to measure the admissibility and diversity of trajectories, however, they suffer from biases, such as length of trajectories.**

In this paper, we propose a new benChmarking paRadIgm for evaluaTing trajEctoRy predIction Approaches (CRITERIA). Particularly, we propose 1) a method for extracting driving scenarios at varying levels of specificity according to the structure of the roads, models' performance, and data properties for fine-grained ranking of prediction models; 2) A set of new bias-free metrics for measuring diversity, by incorporating the characteristics of a given scenario, and admissibility, by considering the structure of roads and kinematic compliancy, motivated by real-world driving constraints; 3) Using the proposed benchmark, we conduct extensive experimentation on a representative set of the prediction models using the large scale Argoverse dataset. We show that the proposed benchmark can produce a more accurate ranking of the models and serve as a means of characterizing their behavior. We further present ablation studies to highlight contributions of different elements that are used to compute the proposed metrics[1].

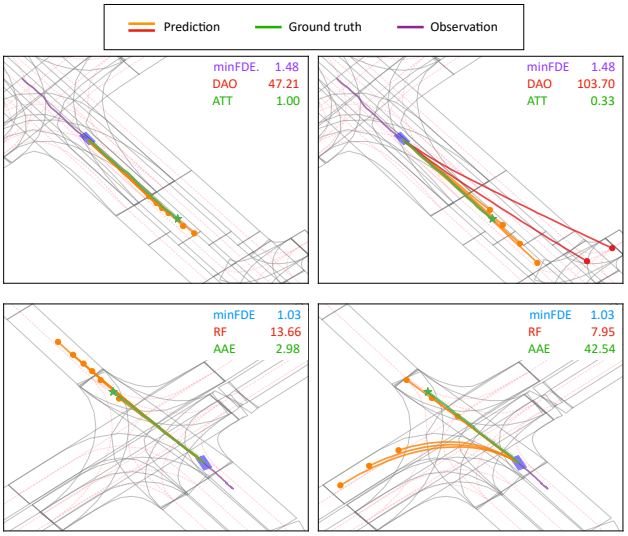

Fig. 1: Examples of two scenarios with two different prediction sets where diversity and admissibility of the trajectories are not reflected in minimum final displacement error (minFDE). In the top two predictions, minFDEs are the same, even though in the right prediction, the red trajectories are inadmissible. In the bottom row, minFDE of both cases are the same, but the right example has a more diverse trajectory.

I. Introduction

Trajectory prediction is one of the main components of autonomous driving systems for estimating the behavior of other road users for safe motion planning. One of the challenges in prediction is that the future behaviors of road users are uncertain due to various hidden factors, such as unknown intentions. Thus, trajectory prediction models typically produce multimodal outputs to capture all possibilities.

Existing trajectory prediction benchmarks [1]–[3] consist of driving data and distance-based metrics to evaluate the accuracy of the models against the ground truth. The datasets, however, are typically biased towards more common scenarios, such as cruising. Consequently, when metrics are averaged over the entire data, the benchmark tends to favor the models that perform better on common cases and provides little insight into the differences among the models and how they behave in rarer situations. This would limit the ability to observe the effects of design choices on particular driving scenarios.

Another shortcoming of existing benchmarks is the over-reliance on distance-based metrics as they are insufficient for assessing the diversity and admissibility of generated trajectories. For the former, one requires to measure the relative positioning of multimodal outputs, therefore the diversity is not effectively captured by measuring error against single-biased ground truths. For admissibility, understanding of the road structure and dynamical constraints is important. Neither of these factors, however, are adequately perceivable when merely computing distance error. The aforementioned shortcomings point to the need for a better measure of goodness of predicted trajectories (see Fig. 1).

In the literature, there are a number of metrics proposed for measuring diversity and admissibility of trajectories. However, the existing diversity-based metrics [4], [5] are based on trajectory end-points, therefore they do not distinguish between lateral (following different routes) and longitudinal (different driving behavior, e.g. speed up/down) diversities, which have different meanings in real-world driving. Furthermore, existing admissibility metrics [4], [6] are heavily biased towards the length of the trajectories, meaning that the

[1]University of Toronto. Work done during internship at Huawei.
[2]Noah's Ark Laboratory, Huawei Technologies Canada.
*Corresponding author Mozhgan.Pourkeshavarz@huawei.com
[1]Code available at https://github.com/huawei-noah/SMARTS/tree/CRITERIA-latest/papers/CRITERIA

longer the generated trajectories are the better metric values get. Additionally, since the entire drivable area is considered valid by these metrics, they verify trajectories even though they fall on invalid parts of the map, such as lanes going in the opposite direction of the vehicle (see Fig. 1).

To this end, we propose a new benChmarking paRadIgm for evaluaTing trajEctoRy predIction Approaches (CRITERIA) with a goal of providing a more accurate characterization of trajectory prediction models, and consequently better design choices. We propose a new approach for extracting different driving scenarios. In detail, we categorize driving scenarios on different levels of specificity based on agreement among models' performance as well as characteristics of the scene and driving task. This would provide a fine-grained categorized scenarios for better understanding of models' behavior under different conditions.

Furthermore, we propose a novel set of bias-free metrics for accurately measuring the performance of trajectory prediction models. Our metrics present a new insight into the diversity and admissibility of trajectories by defining road and kinematic compliancy for normal driving behavior, which are vital aspects of real-world driving. Our metrics help characterize models' behavior more effectively and provide a different, yet more accurate ranking scheme than the existing metrics.

In summary, our **main contributions** are as follows: We present a new benchmarking paradigm named CRITERIA, to evaluate trajectory prediction models. For this purpose, 1) we present a new scheme that relies on agreement among prediction models as well as driving scenes and tasks to extract driving scenarios with different levels of specificity. 2) We propose a new set of metrics to assess diversity and admissibility of trajectories while mitigating biases in the existing metrics with the ability to define conformity at different levels, including road and kinematic compliancy. 3) We conduct extensive experimentation to highlight the effectiveness of the proposed benchmarking paradigm for ranking and characterizing prediction models. 4) Via ablation studies, we show the contribution of different elements of our admissibility metric and their corresponding values on overall ranking of the models.

II. Related work

A. Trajectory Prediction

An extensive body of literature is dedicated to trajectory prediction in the context of autonomous driving [7]–[15]. Existing methods investigate various representations and approaches to learning scene context, including rasterized images [16]–[18], point-clouds [19] processed with convolutional neural networks, and vectorized representations processed using different architectures, such as graph neural networks [20]–[23] or transformer architectures[24]–[27] often combined with sophisticated fusion mechanisms [28]. To address the inherent uncertainty in predicting future trajectories, many models resort to generating multimodal trajectories in order to cover the space of possibilities. There are different approaches to achieving this goal, including the use of generative adversarial networks [29] [30], conditional variational autoencoders [31], or training sampling networks [5], [32]. Some methods also use trajectory endpoints [22] [33] or targets [22], [33]–[35] to model possible future intentions of vehicles to encourage more diverse predictions. Regardless of the method of choice, there is a need for an effective mechanism to evaluate the quality of generated trajectories both in terms of diversity and admissibility, i.e. compliancy to road structure or dynamical limitations.

B. Evaluation Metrics

1) Distance-based Metrics: Distance-based metrics are the most commonly used metrics in the domain of trajectory prediction. They are often computed based on the Euclidean distance between the ground truth and the generated trajectories and reported as average displacement error (ADE) or/and final displacement error (FDE) [1]. In the multimodal setting, the mode with a minimum error (e.g. minADE, minFDE) with respect to the ground truth is selected as the reported value. An extended version of these metrics also accounts for the probability of the selected mode, taking the form of brier-minADE/FDE as in [1]. Miss Rate (MR) [36] [37] is another common metric in this category that corresponds to the proportion of trajectories whose Euclidean prediction error surpasses a predetermined threshold. Mean average Precision (mAP) [38] is a variation of MR that also considers positional uncertainty in predictions.

2) Admissibility Metrics: Admissibility metrics determine whether trajectories satisfy map constraints, e.g. road boundaries. Some of the metrics in this category include Driveable Area Compliance (DAC) and off-road rate [1], which measure the ratio of trajectories that fall within the driveable area. Both of these metrics, however, consider the entire scene as the drivable area and do not penalize the model if the predictions go outside of the valid lanes, e.g. the ones in the opposite direction. Another metric is referred to as Driveable Area Occupancy (DAO) [4], which measures the proportion of pixels occupied by the trajectories in the drivable area. Similar to other metrics, DAO is highly biased towards the length of trajectories and it does not constrain the extent of the drivable area. This means, the longer trajectories get the higher (better) the value of the metric.

3) Diversity Metrics: As the name implies, these metrics are designed to measure how well the generated trajectories cover future possibilities. Existing diversity metrics, such as Ratio of avgFDE to minFDE (RF) [4], minimum average self-distance (minASD), and minimum final self-distance (minFSD) [5] depend on the distances between generated trajectories. For instance, RF measures the spread of predicted trajectories using their final positions [4], which is highly dependent on the single-biased ground truth. MinASD and minFSD are calculated using the minimum average and final distance between all pairs of predicted trajectories [5]. This means that, first, these metrics do not distinguish between lateral and longitudinal diversity, which is fundamentally different in real-world driving, and second, they are biased towards the length of the trajectories.

III. CRITERIA

Characterizing existing models and analyzing their behavior under different conditions can provide new insights for future research directions and help improve trajectory prediction models. In this regard, we propose a new benChmarking paRadIgm for evaluaTing trajEctoRy predIction Approaches (CRITERIA). More superficially, we present a new method for extracting scenarios in a systematic and meaningful way (Section III-B) and a set of novel metrics for measuring diversity and admissibility of trajectory prediction models (Sections III-C and III-D).

A. Problem Formulation

Trajectory prediction task can be viewed as an optimization problem. At time step t, let consider the past trajectory of the i-th agent as a set of 2D coordinates in bird's eye view over some observation horizon L time steps $X_i = \{(x_i, y_i)^{t-L+1}, \cdots, (x_i, y_i)^t\}$. Then, the objective is to predict future trajectories $Y_i = \{(x_i, y_i)^{t+1}, \cdots, (x_i, y_i)^{t+T}\}$, where T is the prediction horizon. In this formulation, additional contextual information, such as road lane boundaries and links are used. In the multimodal setting, prediction models output K different trajectories (referred to as modes) per each vehicle. For simplicity, in the remainder of this paper, we refer to $t-L+1$ and $t+T$ as L and T, respectively.

B. Scenario Extraction

Autonomous driving datasets consist of a variety of scenarios with different levels of complexity, such as simple cruising and turning at intersections. These datasets, however, are biased towards more common scenarios, such as cruising. As a result, evaluating prediction models by averaging the performance over the entire datasets tells little about the strengths and weaknesses of the models and also favors the models that perform better on the common cases.

Trajectory prediction models have different architectural designs and often behave differently when applied to the same samples. For instance, the same models may generate different distributions, and trajectories with different lengths, curvature, etc. As a result, it is essential to extract driving scenarios in a meaningful way to better reflect such differences among the models. To this end, we propose a method to extract driving scenarios based on three criteria, namely the road structure, models' performance, and data properties.

1) Road Structure: We select two criteria that impose two different driving behavior, namely straight roads that impose cruising behavior and intersections that require turning behavior. For each time step in the ground truth trajectories, we obtain all lane segments intersecting with the current route with a radius of 100 meters from the current position of the vehicle. We tag the scenario as a turn if the adjacent lane segments contain at least one intersection with either a right or left turn, otherwise, it is considered a cruising scenario.

2) Model Performance: We determine the difficulty of scenarios by aggregating models' accuracy computed based on the average minFDE of all models that are being evaluated on each scenario. We categorize the scenarios into three groups: hard, medium, and easy, with α_1, α_2, and α_3 percent each, where we set them empirically.

3) Data Properties: The last factor we consider is the length of trajectories as they implicitly reflect different driving styles. For example, future trajectories indicate slowing down behavior to yield or stop at an intersection. We empirically identify a threshold value β based on which we categorize trajectories into short and long groups.

C. Admissibility Metric

We propose Admissibility Triad Test (ATT) metric comprised of three tests motivated by the admissibility definition in real-world driving scenarios. The trajectory is considered admissible if it passes all tests, otherwise, it is counted as inadmissible. The final metric is computed based on the ratio of admissible trajectories. The tests are described below.

1) Road Boundary Compliancy Test: This test determines whether a trajectory is compliant to the road boundaries. For this purpose, we examine the positions of trajectory points at each time step (point-wise) and if a point falls within the drivable area, it passes the test.

2) Road Boundary Alignment Test: This test identifies whether the trajectory orientation is aligned with the lane driving direction. Let consider some trajectory τ_i. We first find the orientation of the trajectory based on its last three time steps. Then, we retrieve the lanes from the HD map where the trajectory's last three time steps are situated. Subsequently, we calculate the orientations τ_l of these lanes, and then calculate the orientation variances $\Delta\Theta$ between τ_l and τ_i for the most recent three time steps. Next, we obtain the confidence level between the trajectory's orientation and that of the lanes as below:

$$\mathcal{C} = max(0, 1 - \frac{\Delta\Theta}{\pi}). \tag{1}$$

The maximum confidence value is then compared with $\text{threshold}_{\text{lac}}$. We set the $\text{threshold}_{\text{lac}} = 0.5$, which indicates that the orientation difference should be less than $90°$. The trajectory passes the test if the maximum confidence of the final three time steps is more than $\text{threshold}_{\text{lac}}$, meaning that it is aligned with the road boundaries.

3) Kinematic Compliancy Test: This test determines if a trajectory's longitudinal acceleration is within an admissible range. We begin by calculating the longitudinal acceleration based on the average initial acceleration and final acceleration of predicted trajectories. Then, by defining an admissible range for normal driving behaviors [39], the trajectories pass the test if their longitudinal acceleration falls within the admissible ranges.

D. Diversity metrics

We consider diversity as the spread of predicted trajectories in lateral and longitudinal directions; hence, we propose two metrics to capture both aspects.

1) AAE: Average Angular Expansion: AAE determines lateral diversity by computing the differences between predicted trajectories' angles. It captures the lateral variation of trajectories independently of their lengths. Mathematically speaking, given K trajectories in a multimodal setting, for each pair of trajectories $(i, j), i \neq j$, we calculate the angle between the first and last time steps as follows:

$$AE_{i,j} = \angle\,[\vec{v}_i, \quad \vec{v}_j] \tag{2}$$

where $\vec{v}$ is a trajectory vector comprised of the first and last points $\vec{v} = \langle \hat{x}^T - \hat{x}^t, \hat{y}^T - \hat{y}^t \rangle$ of each trajectory. The angle between two trajectories is calculated once $AE_{ij} = AE_{ji}$. Then, we compute AAE by averaging over all trajectory pairs' angular expansion.

2) AMV: Average Magnitude Variation: The AMV metric measures longitudinal diversity based on the rate of change (speed-up/down driving behavior) of trajectories. In this manner, however, it is important to exclude scores resulted from exaggerated longitudinal changes, i.e. trajectories should be kinematically feasible besides contributing to the diversity. In this regard, we first run the kinematic compliancy test mentioned in Section III-C.3 and clip the trajectories at their maximum acceptable length according to the test. Then, we compute longitudinal diversity as follows: given K trajectories in a multimodal setting, for each pair of trajectories $(i, j), i \neq j$, we calculate the magnitude of the difference between the two trajectories per each time step and then average over the entire prediction horizon T,

$$MV_{i,j} = \sum_{t=1}^{t=T} |\|\vec{v}_i^{\,t}\| - \|\vec{v}_j^{\,t}\|| \tag{3}$$

where $\|\vec{v}^{\,t}\|$ is the magnitude of the trajectory vector $\vec{v} = \langle \hat{x}^t - \hat{x}^{t-1}, \hat{y}^t - \hat{y}^{t-1} \rangle$ starting at time step $t-1$ and ending at t that is passed by kinematic the compliancy test. The magnitude variation between two trajectories is calculated once $MV_{ij} = MV_{ji}$, and then AMV is calculated by taking an average over all trajectory pairs' magnitude variances.

IV. Experiments

A. Experimental Setup

Dataset. We conduct our evaluations on the Argoverse dataset [6], where the task is to predict 3 seconds of future trajectories given 2 seconds of past observations. This dataset consists of more than $300K$ real-world driving sequences, which are split into train (205K), validation (39K), and test (78K) sets without geographical overlap, along with corresponding HD maps.

Metrics. We report on accuracy metrics, minADE [1] and minFDE [1]. For diversity and admissibility, in addition to the proposed metrics AAE, AMV and ATT, we report on RF [4], minASD/minFSD [5], DAO [4], and DAC [1].

Models. We select the models with different architectural designs to better reflect the characteristics of each approach in different scenarios. These models are: *TNT* [33] as a target-based model, *LaneGCN* [28] as a graph-based method that uses road information, *FTGN* [40] as a graph-based model that relies on the entire scene, *MMTransformer* [27] as a fully transformer-based model, and finally *HiVT* [25] as the recent SOTA model which utilizes an innovative combination of transformers and recurrent networks. For all models, except TNT[2], we use the official code released by the authors.

TABLE I: Number of scenarios in each category, where C1, C2, and C3 indicate the road structure, model performance, and data characteristics criteria, respectively.

C1	C2	C3	Number
Hard	Turn	short	2230
		long	1124
	Cruising	short	248
		long	345
Middle	Turn	short	8424
		long	4179
	Cruising	short	1565
		long	3594
Easy	Turn	short	5106
		long	4581
	Cruising	short	2544
		long	5532

Parameter setup. We split the Argoverse validation set into 12 groups according to the criteria discussed in Sec. III-B. The statistics are reported in Table I. We empirically set α_1, α_2, and α_3 as 10%, 45%, and 45%, respectively. In the kinematic compliancy test, we set the range of longitudinal acceleration for normal driving as $(-2m/s^2$ - $1.47m/s^2)$. For the scenario extraction, we set β, which determines whether trajectories are short or long, as $28.8m$.

B. Accuracy, Diversity, and Admissibility

We begin by providing an overview of the models and their rankings using the existing and proposed metrics. We report the results over the entire validation set, marked as overall and a challenging subset of the data comprising of hardest scenarios, i.e. hard turns with long trajectories. The results are summarised in Table II. Note that the proposed metrics in the case of overall are weighted according to the degree of difficulty of samples to provide a more representative overview of the performance.

Our first observation of the results is that the best model in terms of accuracy, HiVT, is not the most diverse, nor admissible one. In fact, on RF and DAO, this model ranks last. TNT, on the other hand, exhibits an opposite behavior by ranking high on diversity and admissibility while having the lowest accuracy.

The proposed diversity metrics show an alternative ranking. Thanks to removing length bias, we can see that MM-Transformer achieves the highest magnitude diversity. This new ranking suggests that TNT tends to generate longer trajectories, compared to other models, hence being favored in diversity metrics in general. This model, however, achieves a second best score in terms of angular diversity owing to target selection based on road topology.

For admissibility, the top models' rankings are similar based on DAO and ATT. This is generally expected as TNT

[2]The implementation used is from `https://github.com/Henry1iu/TNT-Trajectory-Prediction`

TABLE II: Results on the Argoverse validation set and the most challenging scenarios.

		Accuracy		Diversity					Admissibility		
	Models	minFDE↓	minADE↓	RF↑	minFSD↑	minASD↑	AAE↑	AMV↑	DAO↑	DAC↑	ATT↑
Overall	TNT [33]	1.73	0.95	3.98	**10.83**	**2.79**	13.77	6.50	**86.53**	0.9906	**0.861**
	LaneGCN [28]	1.08	0.71	4.41	3.79	0.95	11.43	6.64	73.82	0.9917	0.830
	HiVT [25]	**0.96**	**0.66**	3.64	0.40	0.12	9.77	6.17	70.18	**0.9919**	0.812
	FTGN [40]	1.08	0.73	4.14	2.51	0.68	**14.23**	6.90	72.43	0.9908	0.816
	MMTrans.[24]	1.08	0.71	**4.64**	4.51	1.14	12.14	**7.03**	74.58	0.9902	0.821
Challenging	TNT [33]	7.38	3.54	1.94	10.90	2.88	8.63	7.05	**79.26**	0.9634	**0.798**
	LaneGCN [28]	4.51	2.40	2.68	13.48	3.46	6.73	8.90	62.53	0.9662	0.734
	HiVT [25]	4.34	**2.20**	2.61	1.35	0.41	5.31	8.59	60.20	0.9655	0.706
	FTGN [40]	4.80	2.46	2.62	7.81	2.25	**12.05**	**9.66**	58.29	0.9552	0.709
	MMTrans.[24]	**4.30**	2.30	**2.93**	**16.10**	**4.12**	7.35	9.35	65.10	**0.9729**	0.731

tends to constrain trajectory endpoints based on center-lines of the lanes, hence producing more compliant trajectories. For other models, the changes in ranking suggest that they are not successful at one or more of the tests proposed as part of ATT (more on this in Sec. IV-E).

In challenging scenarios, we see a significant degradation across all metrics for all models. The degree of change, however, is different. For instance, LaneGCN and FTGN that were sharing the same spot with MMTransformer, have more performance degradation compared to this model. The relative differences across other metrics remain stable, with the exception of TNT whose performance on the existing diversity metrics drops drastically due to significantly worse accuracy and HiVT which changes its first position in DAC with MMTransformer. Despite being the most accurate model, HiVT continues being at the bottom of the diversity and admissibility rankings across all metrics.

C. Scenario-based Analysis

To get a better sense of performance differences, we take a look at a finer breakdown of scenarios. For this purpose, we use the criteria described in Sec. I and split the samples into 12 categories (see Table I for statistics of each category). For each model under a given scenario, we report the results using the proposed metrics.

As shown in Table III, TNT appears to be a more successful method in terms of diversity, however, with a different degrees of success. In general, in hard turn scenarios, TNT ranks second last and last in AMV, suggesting that this model does not effectively model vehicles' dynamics.

The biggest change in the ranking can be seen in terms of the admissibility metric, ATT. HiVT maintains the first place throughout easy scenarios and performs reasonably in middle scenarios. However, in Hard scenarios, HiVT is placed last in all cases except cruising long where the performance gap between all models is relatively smaller compared to the gaps in metrics in other scenarios. TNT's admissibility generally tends to be lagging in cruising cases, especially with longer horizons. The fluctuations in the performance of the models highlight their limitations under different conditions and the impact of certain scenarios' properties on their performance.

D. How to Select the Best Model

As discussed earlier, when evaluating the performance of trajectory prediction models, there are three considerations:

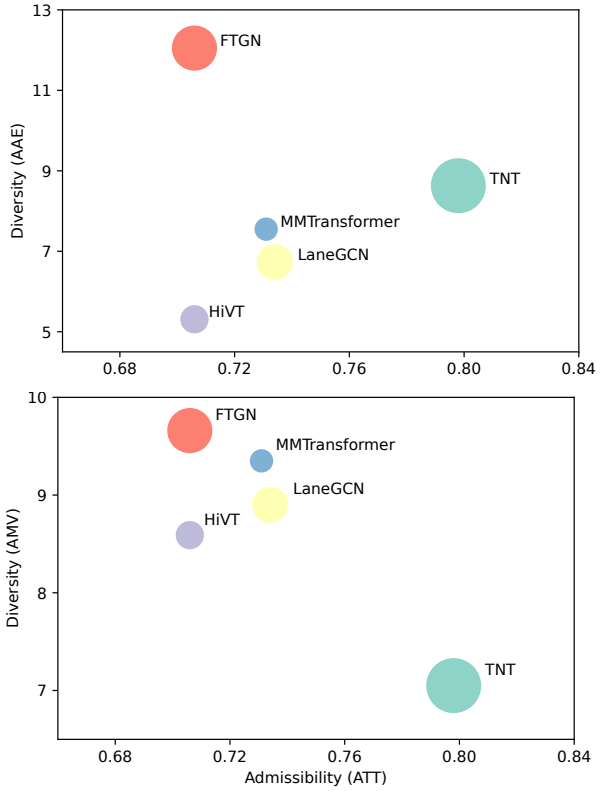

Fig. 2: Performance comparison on the challenging scenarios in terms of diversity, admissibility, and minFDE (represented as the size of circles, so smaller is better).

trajectories should be accurate, diverse and admissible. In Sec. IV-B, we showed that performing good based on one class of metrics does not necessarily translate to a good performance using another class of metrics. This raises the question of how one should select (or design) a model based on the performance? The answer to this question is not intuitive and highly depends on the use case of the model which may put more emphasis on one aspect of the performance, e.g. diversity, over the other, e.g. accuracy. However, for ranking models based on multiple criteria, one can look at how balanced the performance of the models are. For this purpose, we report the results using a three-dimensional representation as illustrated in Fig. 2.

We can see that when considering challenging cases, at two ends of the spectrum, TNT tends to be most successful in terms of admissibility and FTGN in terms of diversity. However, taking into account the accuracy of these models,

TABLE III: Results on the extracted scenarios with the existing and proposed metrics for diversity and admissibility.

		Cruising						Turn					
		Short			Long			Short			Long		
	Models	AAE↑	AMV↑	ATT↑	AAE↑	AMV↑	ATT↑	AAE↑	AMV↑	ATT↑	AAE↑	AMV↑	ATT↑
Hard	TNT [33]	8.71	**6.63**	**0.894**	1.68	**7.75**	0.885	**20.10**	5.83	**0.875**	8.63	7.05	**0.798**
	LaneGCN [28]	**9.95**	5.86	0.842	1.52	6.80	0.883	16.95	5.86	0.849	6.73	8.90	0.734
	HiVT [25]	8.55	5.26	0.818	1.46	5.91	**0.902**	14.87	5.52	0.827	5.31	8.59	0.706
	FTGN [40]	9.18	6.30	0.852	**2.53**	7.00	0.891	19.45	5.91	0.832	**12.05**	**9.66**	0.709
	MMTrans.[24]	6.27	6.29	0.862	1.07	6.84	0.867	18.50	**6.31**	0.834	7.35	9.35	0.730
Middle	TNT [33]	**7.26**	**7.29**	0.916	**1.35**	**8.10**	0.921	**15.93**	**6.45**	0.884	2.72	**7.70**	**0.903**
	LaneGCN [28]	4.66	5.15	0.919	1.15	5.73	0.935	9.58	5.07	**0.902**	1.84	6.31	0.892
	HiVT [25]	2.65	4.22	0.911	0.45	4.32	**0.958**	6.94	4.44	0.900	1.04	5.18	0.897
	FTGN [40]	5.84	5.32	0.914	1.32	5.61	0.940	11.71	5.10	0.894	**3.69**	6.36	0.897
	MMTran.[24]	3.36	5.61	**0.927**	0.62	5.60	0.923	10.89	5.45	0.886	1.50	6.19	0.898
Easy	TNT [33]	**9.09**	**7.77**	0.944	**1.18**	**8.37**	0.940	**10.53**	**6.83**	0.908	1.73	**8.23**	0.936
	LaneGCN [28]	3.36	4.15	0.969	0.87	5.08	0.958	4.63	4.42	0.935	1.13	5.22	0.947
	HiVT [25]	1.87	2.76	**0.971**	0.22	3.36	**0.985**	2.64	3.44	**0.937**	0.41	3.57	**0.972**
	FTGN [40]	4.78	4.09	0.959	0.88	4.82	0.967	6.79	4.44	0.933	**2.03**	4.85	0.964
	MMTran.[24]	3.61	4.71	0.970	0.48	4.99	0.941	5.02	4.86	0.931	0.76	4.95	0.947

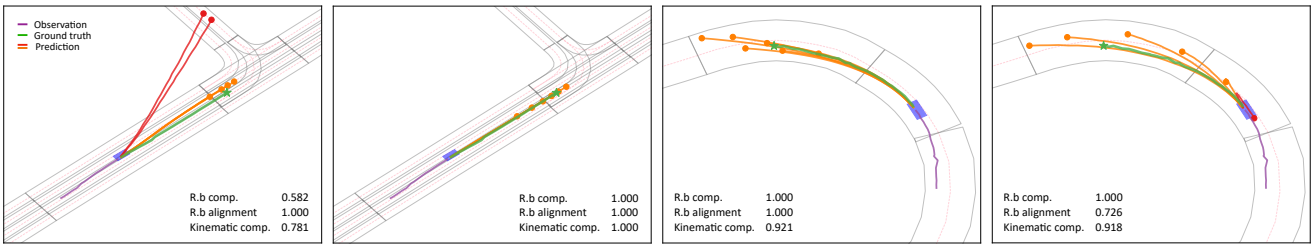

Fig. 3: Qualitative samples of different ATT test scores for predicted trajectories. The red lines indicate inadmissible trajectories. From left to right, the samples are generated by TNT, MMTransformer, TNT and LaneGCN.

TABLE IV: Contributions of the tests in ATT. R.b and comp. stand for road boundary and compliancy, respectively.

models	R.b comp.	Kinematic comp.	R.b alignment
TNT [33]	0.938	0.918	**0.822**
LaneGCN [28]	0.832	0.928	0.542
HiVT [25]	0.807	**0.935**	0.546
FTGN [40]	0.916	0.927	0.561
MMTrans.[24]	**0.977**	0.922	0.628

both tend to lag behind others. Considering all three metrics, MMTransformer and LaneGCN offer a more balanced performance. These models are placed generally in the center areas between diversity and admissibility axes and at the same time have reasonable accuracy.

E. Ablation study on Contributions of Tests in ATT metric

We examine the models in terms of individual ATT's components, namely road boundary, kinematic, and road boundary alignment compliancy tests. As shown in Table IV, the models have different strengths. MMTransformer stands out in terms of road boundary compliancy while HiVT is best in kinematic and TNT in road alignment compliancy tests. In general, road alignment is the weakest point of all models where there is a large gap between TNT and the rest. TNT's superior performance can be attributed to its target selection mechanism which constrains the predicted trajectories to correct lane center-lines.

In terms of other tests, while all models are fairly kinematically compliant, they are not equally successful in road boundary test. In fact, we can see a big difference between the bottom two models, LaneGCN and HiVT and the top two, TNT and MMTransfomer. Such differences can show the effectiveness of map encoding methods used by different models. We illustrate examples of generated trajectories by the models and corresponding tests scores in Fig. 3.

V. Conclusions

In this paper, we presented a new paradigm for evaluating the performance of vehicle trajectory prediction models for autonomous driving. Our proposed approach consists of a new method to extract scenarios from existing autonomous driving datasets based on characteristics of the data as well as performance agreement among prediction models. Additionally, we proposed three new metrics for measuring diversity and admissibility of trajectories by eliminating length and accuracy biases in the existing metrics.

We conducted extensive set of experiments using the proposed evaluation approach on state-of-the-art trajectory prediction models and provided an alternative perspective of their performance. We showed that good performance under one set of criteria does not necessarily translate to good performance across all scenarios or metrics. Additionally, we demonstrated how the newly proposed metrics can be used to characterize the performance of the models. For future work, we will consider a finer evaluation at model architectural level by conducting studies on models' design choices using the proposed CRITERIA evaluation protocol.

REFERENCES

[1] M.-F. Chang, J. W. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, 2019.

[2] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *CVPR*, 2020.

[3] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv:2301.00493*, 2023.

[4] 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 *ECCV*, 2020.

[5] Y. J. Ma, J. P. Inala, D. Jayaraman, and O. Bastani, "Likelihood-based diverse sampling for trajectory forecasting," in *CVPR*, 2021.

[6] 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 *CVPR*, 2019.

[7] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *ICRA*, 2023.

[8] S. Pini, C. S. Perone, A. Ahuja, A. S. R. Ferreira, M. Niendorf, and S. Zagoruyko, "Safe real-world autonomous driving by learning to predict and plan with a mixture of experts," in *ICRA*, 2023.

[9] A. Cui, S. Casas, K. Wong, S. Suo, and R. Urtasun, "Gorela: Go relative for viewpoint-invariant motion forecasting," in *ICRA*, 2023.

[10] M. Wang, X. Zhu, C. Yu, W. Li, Y. Ma, R. Jin, X. Ren, D. Ren, M. Wang, and W. Yang, "Ganet: Goal area network for motion forecasting," in *ICRA*, 2023.

[11] S. Su, Y. Li, S. He, S. Han, C. Feng, C. Ding, and F. Miao, "Uncertainty quantification of collaborative detection for self-driving," in *ICRA*, 2023.

[12] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "ViP3D: End-to-end visual trajectory prediction via 3d agent queries," in *CVPR*, 2023.

[13] B. Ivanovic, J. Harrison, and M. Pavone, "Expanding the deployment envelope of behavior prediction via adaptive meta-learning," in *ICRA*, 2023.

[14] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *CVPR*, 2023.

[15] K. Jain, V. Chhangani, A. Tiwari, K. M. Krishna, and V. Gandhi, "Ground then navigate: Language-guided navigation in dynamic scenes," in *ICRA*, 2023.

[16] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "MultiPath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," in *CoRL*, 2019.

[17] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*, 2020.

[18] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "GOHOME: Graph-oriented heatmap output for future motion estimation," in *ICRA*, 2022.

[19] M. Ye, T. Cao, and Q. Chen, "TPCN: Temporal point cloud networks for motion forecasting," in *CVPR*, 2021.

[20] 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 *CVPR*, 2020.

[21] J. Mercat, T. Gilles, N. El Zoghby, G. Sandou, D. Beauvois, and G. P. Gil, "Multi-head attention for multi-modal joint vehicle motion forecasting," in *ICRA*, 2020.

[22] J. Gu, C. Sun, and H. Zhao, "DenseTNT: End-to-end trajectory prediction from dense goal sets," in *ICCV*, 2021.

[23] W. Zeng, M. Liang, R. Liao, and R. Urtasun, "LaneRCNN: Distributed representations for graph-centric motion forecasting," in *IROS*, 2021.

[24] Z. Huang, X. Mo, and C. Lv, "Multi-modal motion prediction with transformer-based neural network for autonomous driving," in *ICRA*, 2022.

[25] Z. Zhou, L. Ye, J. Wang, K. Wu, and K. Lu, "HiVT: Hierarchical vector transformer for multi-agent motion prediction," in *CVPR*, 2022.

[26] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D'Souza, S. E. Kahou, F. Heide, and C. Pal, "AutoBot: Latent variable sequential set transformers for joint multi-agent motion prediction," in *ICLR*, 2022.

[27] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, "Multimodal motion prediction with stacked transformers," in *CVPR*, 2021.

[28] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning lane graph representations for motion forecasting," in *ECCV*, 2020.

[29] 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 *CVPR*, 2019.

[30] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *CVPR*, 2019.

[31] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *NeurIPS*, 2015.

[32] I. Bae, J.-H. Park, and H.-G. Jeon, "Non-probability sampling network for stochastic human trajectory prediction," in *CVPR*, 2022.

[33] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid *et al.*, "TNT: Target-driven trajectory prediction," in *CoRL*, 2021.

[34] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi, "LOKI: Long term and key intentions for trajectory prediction," in *ICCV*, 2021.

[35] K. Mangalam, Y. An, H. Girase, and J. Malik, "From goals, waypoints & paths to long term human trajectory forecasting," in *ICCV*, 2021.

[36] R. A. Yeh, A. G. Schwing, J. Huang, and K. Murphy, "Diverse generation for multi-agent sports games," in *CVPR*, 2019.

[37] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *CVPR*, 2017.

[38] 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 *CVPR*, 2021.

[39] I. Bae, J. Moon, J. Jhung, H. Suk, T. Kim, H. Park, J. Cha, J. Kim, D. Kim, and S. Kim, "Self-driving like a human driver instead of a robocar: Personalized comfortable driving experience for autonomous vehicles," *arXiv:2001.03908*, 2022.

[40] G. Aydemir, A. K. Akan, and F. Güney, "Trajectory forecasting on temporal graphs," *arXiv:2207.00255*, 2022.