Beyond In-Distribution Performance: A Cross-Dataset Study of Trajectory Prediction Robustness

Yue Yao[1,2], Daniel Goehring[2], Joerg Reichardt[1]

Abstract:

We study the Out-of-Distribution (OoD) generalization ability of three SotA trajectory prediction models with comparable In-Distribution (ID) performance but different model designs. We investigate the influence of inductive bias, size of training data and data augmentation strategy by training the models on Argoverse 2 (A2) and testing on Waymo Open Motion (WO) and vice versa. We find that the smallest model with highest inductive bias exhibits the best OoD generalization across different augmentation strategies when trained on the smaller A2 dataset and tested on the large WO dataset. In the converse setting, training all models on the larger WO dataset and testing on the smaller A2 dataset, we find that all models generalize poorly, even though the model with the highest inductive bias still exhibits the best generalization ability. We discuss possible reasons for this surprising finding and draw conclusions about the design and test of trajectory prediction models and benchmarks.

Keywords: trajectory prediction, out-of-distribution test, evaluation metrics

1 Introduction

The robustness of trajectory prediction is essential for practical applications in autonomous driving. The advancement of trajectory prediction models is catalyzed through public motion datasets and associated competitions, such as Argoverse 2 (A2) [1], and Waymo Open Motion (WO) [2]. These competitions establish standardized metrics and test protocols and score predictions on test data that is withheld from all competitors and hosted on protected evaluation servers only. This is intended to objectively compare the generalization ability of models to unseen data.

However, these withheld test examples still share similarities with the training samples, such as sensor setup, map representation, post-processing, geographic, and scenario selection biases employed during dataset creation. Consequently, the test scores reported in each competition are examples of *In-Distribution* (ID) testing. To effectively evaluate model generalization, it is essential to test models on truly *Out-of-Distribution* (OoD) test samples, such as those from different motion datasets.

We investigate model generalization across two large-scale motion datasets [3]: Argoverse 2 (A2) and Waymo Open Motion (WO). The WO dataset, with 576k scenarios, is more than twice the size of A2, which contains 250k scenarios. As model generalization

The authors are with [1]Continental Automotive GmbH (yue.yao@continental.com, joerg.reichardt@continental.com), [2]Dahlem Center for Machine Learning and Robotics, Freie Universitaet Berlin (daniel.goehring@fu-berlin.de)

ability is best illustrated by limiting the amount of training data [4], we first train on the smaller dataset A2 and test on OoD samples from the larger dataset WO [3]. Our results demonstrate that model generalization improves through two key factors: introducing inductive bias via polynomial representation (cf. Section 2.1) and implementing proper data augmentation (cf. Section 2.2).

This naturally raises the question: how does model generalization change when the experimental setup is reversed, i.e., when models are trained on the larger WO dataset and tested on OoD samples from the smaller A2 dataset? Given the increased amount of training data, one would hypothesize that all models will exhibit improved generalization in OoD tests. In this study, we investigate model generalization under this reversed experimental setting to assess whether the hypothesized improvements in robustness are realized.

Our paper is organized as follows: We first detail the two design choices influencing model generalization – data representation and data augmentation – identified in our prior work [3]. From the body of related work, we introduce the two benchmark models along with our proposed model and detail their model designs. We then revisit our dataset homogenization protocol for OoD testing and outline the new experimental settings for training on the WO dataset. Subsequently, we present ID and OoD testing results for all models and analyze the new OoD results with respect to data representation, augmentation strategies and compare them to our expectations. Finally, we conclude our work by arguing for establishing OoD testing as an equally important performance metric alongside traditional ID evaluation.

2 Related Work and Benchmark Models

2.1 Data Representation

The structure and rules of individual datasets and accompanying competitions heavily influence the design choices of trajectory prediction models in the literature. Deep learning-based models often directly ingest data in the format the dataset is stored in. Recent studies employ *sequence-based representation*, e.g., sequences of data points, as the model's input and output [5, 6]. This approach effectively captures diverse information, including agent trajectories and road geometries, while aligning with dataset measurement formats. However, it has notable drawbacks: the representation introduces high redundancy, and its computational cost scales with the length and temporal/spatial resolution of the trajectories and road geometries. Furthermore, the high flexibility of this approach can capture measurement noise and outliers, potentially leading to physically infeasible predictions.

Some previous works explored the possibility of using *polynomial representations* for predictions [7, 8]. Su et al. [8] highlight the temporal continuity of this representation, i.e., the ability to provide arbitrary temporal resolution. Reichardt [9] argues for using polynomial representations to integrate trajectory tracking and prediction into a filtering problem. Polynomial representations restrict the kind of trajectories that can be represented and introduce bias into prediction systems. This limited flexibility is generally associated with greater computational efficiency, smaller model capacity, and hence better generalization.

In our previous study, we proposed our model EP with polynomial inputs and outputs as key innovations [3]. Our EP model demonstrates improved robustness against OoD samples compared to sequence-based SotA models.

2.2 Data Augmentation

In competition settings, one or more agents in a scenario are typically designated as *focal agents*, and predictions are scored exclusively for these agents. However, training the model exclusively with the focal agent's behavior fails to exploit all available data. To address this, predicting the future motion of non-focal agents is a typical augmentation strategy for training. As there are many more non-focal agents than focal agents, another important design decision involves determining how to balance the contributions of focal and non-focal agent data.

We select two open-sourced and thoroughly documented SotA models: Forecast-MAE [5] (FMAE) and QCNet [6], with nearly 1900k and 7600k parameters, respectively. As summarized in Table 1, both models employ sequence-based representation but employ different strategies in dealing with non-focal agents:

Heterogeneous Augmentation: FMAE follows the prediction competition protocol and prioritizes focal agent prediction. Thus, agent history and map information are computed within the *focal agent's* coordinate frame. Compared to the *multi-modal* prediction of the focal agent, FMAE only outputs *uni-modal* prediction for non-focal agents. The error of focal agent predictions is weighed higher than for non-focal agents in loss function.

Homogeneous Augmentation: QCNet does not focus on the selected focal agent and proposes a more generalized approach. It encodes the information of agents and map elements in each agent's *individual* coordinate frame. It outputs *multi-modal* predictions for focal and non-focal agents alike. This approach ensures a consistent prediction task for all agents and enhanced model generalization. The prediction error of focal and non-focal agent predictions is weighed equally in the loss functions.

Figure 1 illustrates the augmentation strategies of our benchmark models FMAE and QCNet.

Following our previous study [3], we introduce the third augmentation strategy for comparison: **no augmentation**. This can simply be achieved by removing the loss of non-focal agents in each model, thus limiting the model to learn from focal agent behavior

Table 1: Model variants under study based on augmentation strategies and data representations.

augmentation strategy	input and output representation	
	sequence-based	polynomial-based (ours)
heterogeneous	FMAE [5]	EP-F
homogeneous	QCNet [6]	EP-Q
w/o augmentation	FMAE-noAug QCNet-noAug	EP-noAug

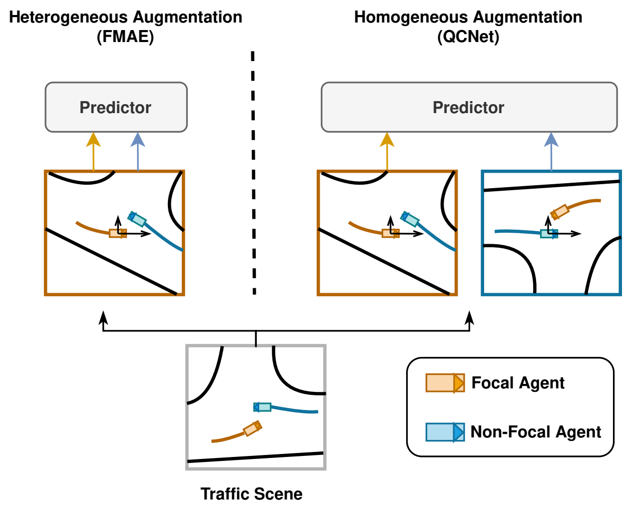

Figure 1: The two augmentation strategies for non-focal agent data employed in benchmark models. **Left**: FMAE employs heterogeneous augmentation, representing information in the focal agent's coordinate frame and only making uni-modal predictions for non-focal agents. **Right**: QCNet employs homogeneous augmentation, encoding information in each agent's individual coordinate frame and making multi-modal predictions for both focal and non-focal agents alike.

only. We denote the benchmark models without augmentation as "FMAE-noAug" and "QCNet-noAug".

To evaluate the impact of data representation independently of augmentation strategies, we implement all three augmentation modes for our proposed model, EP. These include: "EP-noAug" (no augmentation), "EP-F" (heterogeneous augmentation), and "EP-Q" (homogeneous augmentation).

3 Dataset Homogenization

To work around inconsistencies in data formats and prediction tasks between datasets, we inherit the homogenization protocol from the one proposed in prior work [3] with a different *prediction horizon*. This adopts the following settings for both datasets and details are summarized in Table 2:

- History Length: We set the history length to 5 seconds (50 steps) as in A2.
- Prediction Horizon: All models are trained with 4.1s prediction due to the limited recording length of training data in WO.
- Map Information: We exclude boundary information and the label of lane segments in junctions due to the information's absence in WO. Only lane segments and crosswalks are considered map elements in homogenized data due to their availability in both datasets.
- Focal Agent: We take the same focal agent as labeled in A2. From WO, only the first fully observed, non-ego agent in the list of focal agents is chosen. As the list of focal

agents is unordered, this corresponds to randomly sampling a single fully observed focal agent.

Table 2: Datasets comparison and homogenization protocol for OoD testing.

		A2 [1]	WO [2]	Homogenized Dataset
scenario length		11s	9.1s	9.1s
sampling rate		10 Hz	10 Hz	10 Hz
history length		5s	1.1s	5s
prediction horizon		6s	8s	4.1s
map information	lane center	yes	yes	yes
	lane boundary	yes	not for junction lanes	no
	junction lanes labeled	yes	no	no
prediction target	# focal agents	1	up to 8	1
	ego included	no	yes	no
	fully observed	yes	not guaranteed	yes

Based on the focal agent's selection criterion above, we have 472 235 training samples and 42 465 validation samples from WO. For the A2 dataset, we have 199908 and 24 988 valid samples from training and validation split, respectively.

4 Experimental Results

4.1 Experiment Setup and Metrics

All models are trained on WO from scratch using the original hyperparameters. The WO dataset features a more complex map structure, with more lanes and map points than A2. This leads to "out of memory" issue on our GPU cluster with 192 GB of RAM when training QCNet with its default settings. To fit QCNet on the available training hardware, we had to limit the scenario complexity for QCNet to process up to 50 agents and 80 map elements closest to the ego vehicle per scene.

For ID testing, we use the official benchmark metrics, including minimum Average Displacement Error ($minADE_K$) and minimum Final Displacement Error ($minFDE_K$), for evaluation. The metric $minADE_K$ calculates the average Euclidean distance between the ground-truth trajectory and the best of K predicted trajectories across all future time steps. The $minFDE_K$ measures the final-step error, focusing on long-term accuracy.

For OoD testing, we also propose $\Delta minADE_K$ and $\Delta minFDE_K$ as the difference of displacement error between ID and OoD testing to measure model robustness. In accordance with standard practice, K is selected as 1 and 6.

4.2 ID Testing Results

We present our new ID results in the upper section of Table 3 and reference the ID results from our previous study in the lower section for comparison. In the upper section, all models are trained on the homogenized WO training split and tested on the homogenized WO validation split. As in prior work, the QCNet results serve as the reference for relative comparisons, i.e., as 100%.

As highlighted in [10], the WO dataset features a significantly larger volume of data and higher variance, presenting challenges for small models for training and inference. However, with the much smaller model size (345k), both EP-F and EP-Q outperform FMAE and achieve comparable ID performance to QCNet in $minADE_1$ and $minFDE_1$ for the new ID results. Despite limiting the scenario complexity for QCNet, it still achieves the best ID performance in 3 out of 4 metrics.

Also, note that all models perform better on WO than on A2 in absolute error under all augmentation settings in ID tests, except the $minFDE_6$ of FMAE increase from 0.792m to 0.829m. The better ID performance on WO can be attributed to the larger size of the WO dataset, the relative simplicity of WO scenarios, or a combination of both factors.

Table 3: Result of In-Distribution (ID) testing of QCNet, FMAE, EP-Q and EP-F evaluated with a 4.1-second prediction horizon. For clarity, the results of models without augmentation are not reported. **Upper Section**: ID result of models trained on the homogenized WO training set and tested on the homogenized WO validation set. **Lower Section**: Referenced ID result of models trained on the homogenized A2 training set and tested on the homogenized A2 validation set [3].

Training	Models	minADE_1 [m]	minFDE_1 [m]	minADE_6 [m]	minFDE_6 [m]
WO	QCNet [6]	0.820 (100.0%)	2.171 (100.0%)	0.344 (100.0%)	0.696 (100.0%)
	FMAE [5] w/o pre-train	0.889 (108.4%)	2.318 (106.8%)	0.374 (108.7%)	0.829 (119.1%)
	EP-Q (ours) [3]	0.821 (100.1%)	2.155 (99.3%)	0.359 (104.4%)	0.802 (115.2%)
	EP-F (ours) [3]	0.831 (101.3%)	2.171 (100.0%)	0.370 (107.6%)	0.825 (118.5%)
A2	QCNet [6]	0.902 (100.0%)	2.301 (100.0%)	0.400 (100.0%)	0.701 (100.0%)
	FMAE [5] w/o pre-train	0.989 (109.4%)	2.459 (106.9%)	0.423 (105.8%)	0.792 (113.0%)
	EP-Q (ours) [3]	1.161 (128.7%)	2.830 (123.0%)	0.491 (122.8%)	0.922 (131.5%)
	EP-F (ours) [3]	1.082 (120.0%)	2.555 (111.0%)	0.476 (119.0%)	0.865 (123.4%)

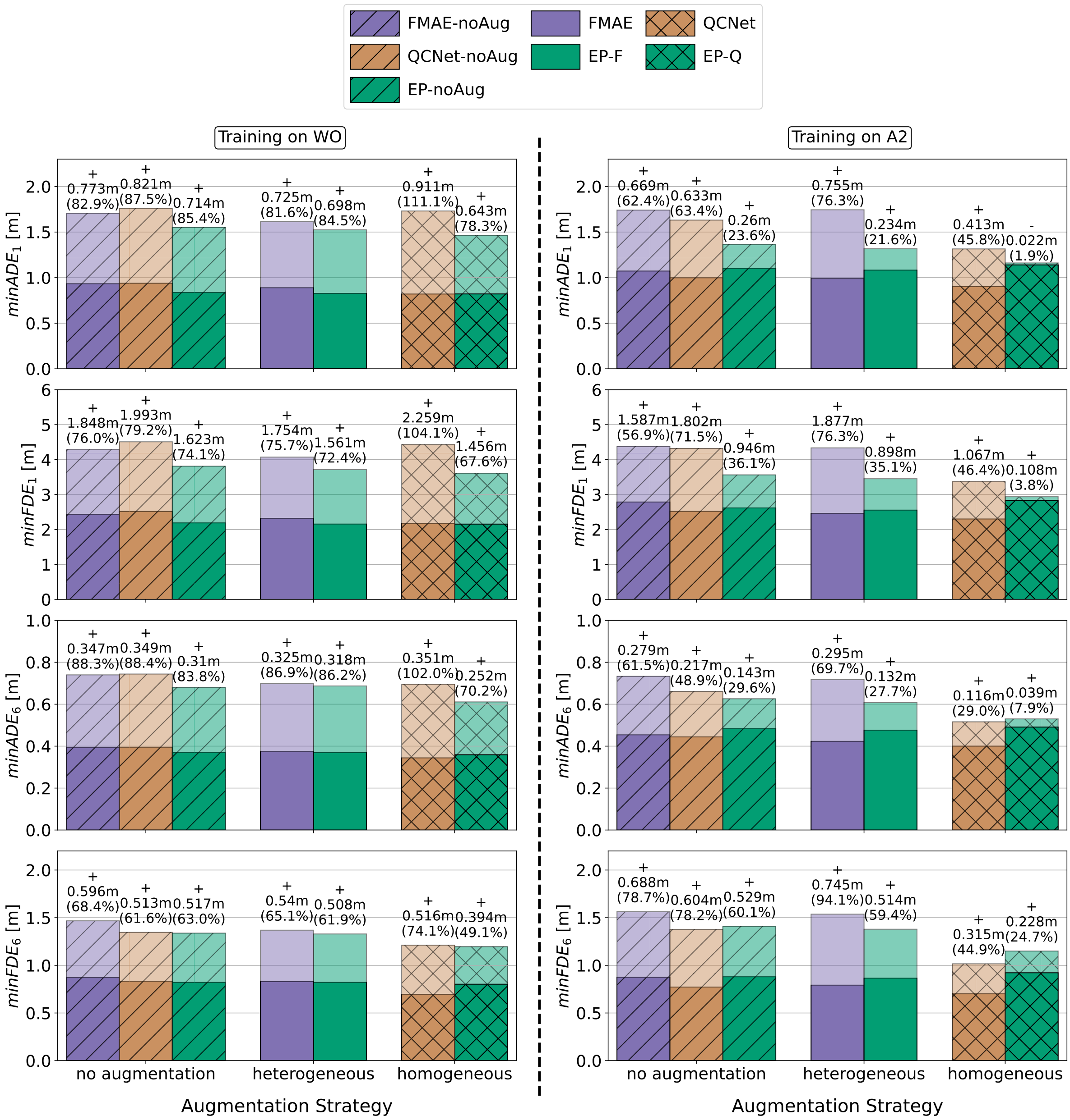

Figure 2: The OoD testing results of FMAE, QCNet, EP and their variants. We indicate the *absolute* and *relative* difference in displacement error between ID and OoD results. The *solid* bars represent ID results reported in Table 3, while the *transparent* bars indicate the increase in displacement error during OoD testing. **Left**: Models trained on the homogenized WO training set and tested on the homogenized WO (ID) and A2 (OoD) validation sets. **Right**: Models trained on the homogenized A2 training set and tested on the homogenized A2 (ID) and WO (OoD) validation sets [3].

4.3 OoD Testing Results

Figure 2 compares our new OoD results (left) with our previous OoD results (right). In both cases, we evaluate model robustness and generalization using the absolute and relative increases in prediction error, e.g. $\Delta minADE_K$ and $\Delta minFDE_K$. We will present the

OoD results from three perspectives: (i) augmentation strategy, (ii) data representation, and (iii) comparison with our expectations in Section 1.

4.3.1 Augmentation Strategy

Our previous study (right side of Figure 2, originally) showed that heterogeneous augmentation yields only marginal improvements in robustness for OoD testing. In our new experiments (left side of Figure 2, originally), this trend continues for both EP-F and FMAE, indicating consistent results with respect to heterogeneous augmentation.

Based on our prior findings, we expect models employing homogeneous augmentation to exhibit a notable improvement in generalization. The results for EP-Q align with this expectation, as $\Delta minADE_6$ decreases from +0.310m (83.8%) to +0.252m (70.2%) compared to EP-noAug. However, QCNet shows contrasting results with worse robustness compared to QCNet-noAug, with $\Delta minADE_6$ increasing from +0.349m (88.4%) to +0.351m (102.0%). The performance of QCNet can be attributed to the limited scenario complexity, as discussed in Section 4.1.

4.3.2 Data Representation

The advantage of using polynomial representation for model generalization remains evident for EP-Q, with a significant decrease in $\Delta minFDE_6$ from +0.516m (74.1%) to +0.394m (49.1%) compared to QCNet. Moreover, EP-Q achieves the lowest prediction error in the new OoD testing across all benchmarks.

In contrast, while EP-F and EP-noAug still demonstrate improvement in robustness compared to sequence-based models, the gains are only marginal. This is reflected in the $\Delta minFDE_6$ between EP-F (+0.508m (61.9%)) and FMAE (+0.540m (65.1%)).

4.3.3 Comparison with Expectations

Comparing the new and previous OoD results, visualized in the left and right parts of Figure 2, respectively, our empirical observations diverge significantly from our expectations outlined in Section 1. For instance, FMAE does not exhibit any improvement in robustness, as indicated by the nearly identical $\Delta minADE_1$ across both settings. Moreover, QCNet and EP variants exhibit significantly poorer generalization under the new settings when trained on WO. Notably, EP-Q's $\Delta minADE_1$ increases from -0.022m (1.9%) in previous OoD results to +0.643m (78.3%) under the new settings. Drawing from our experimental observations and previous work on dataset characteristics, we provide two potential reasons:

> **Complexity of Prediction Task**: Creators of both datasets mined their recordings to identify challenging scenarios and focal agents for prediction [1, 2]. Agent's motion becomes particularly difficult to predict when they accelerate, brake, or turn due to interactions with other agents and map - scenarios that simple models such as constant velocity cannot easily capture. While we maintain the history length of A2 for homogenized datasets, we increase the history length of WO from 1.1 seconds to 5 seconds. This adjustment potentially moves the challenging behaviors into the historical data, effectively making the prediction task less complex for WO. As a result, models trained on WO with these less challenging prediction tasks tend to

generalize poorly when tested on A2, which features more demanding prediction scenarios.

Noise Level in Datasets: Based on our previous study of the noise characteristics across both datasets [11], WO features lower noise levels compared to A2. Homogeneous augmentation and polynomial representation can be effective for training on a noisy dataset, such as A2, enhancing robustness with cleaner OoD test data (WO). Conversely, when the test data exhibits higher noise levels than the training data, both methods have only a minimal improvement in model robustness.

To validate these hypotheses, further research is required, such as including controlled experiments with known noise levels or identifying focal agent behavior. However, both hypotheses are closely tied to dataset properties, making it challenging to isolate and control variables for a definitive explanation. We leave this as an open question for future investigation.

5 Conclusion and Future Directions

Our work emphasizes the importance of evaluating trajectory prediction models beyond In-Distribution (ID) performance, with a focus on Out-of-Distribution (OoD) robustness for autonomous driving. We systematically investigate the ID and OoD performance of three state-of-the-art prediction models and demonstrate the improved model robustness of our proposed EP-Q model with polynomial data representation and homogeneous data augmentation.

However, our findings show that model robustness is influenced by factors beyond model architecture and design choices. Contrary to our initial hypothesis, models trained on the larger Waymo Open Motion dataset showed reduced robustness when tested on the smaller Argoverse 2 dataset. This surprising result highlights the complex relationship between dataset properties, model design choices, and generalization performance. Our work outlines two potential factors for future investigation: the complexity of the prediction task and dataset noise levels. Our results outline the importance of OoD testing for prediction models and emphasize that improving OoD robustness requires a deeper understanding of both model design and dataset properties, rather than simply increasing training data volume.

ACKNOWLEDGMENT

Authors would like to thank Shengchao Yan, Wolfram Burgard for contributions to our previous study, and Andreas Philipp for many useful discussions. We also acknowledge funding from the German Federal Ministry for Economic Affairs and Climate Action within the project “NXT GEN AI METHODS - Generative Methoden fuer Perzeption, Praediktion und Planung”.

Literature

[1] 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," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[2] 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*, pp. 9710–9719, 2021.

[3] Y. Yao, S. Yan, D. Goehring, W. Burgard, and J. Reichardt, "Improving out-of-distribution generalization of trajectory prediction for autonomous driving via polynomial representations," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 488–495, IEEE, 2024.

[4] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *CoRR*, vol. abs/1712.00409, 2017.

[5] J. Cheng, X. Mei, and M. Liu, "Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8679–8689, 2023.

[6] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17863–17873, 2023.

[7] T. Buhet, E. Wirbel, A. Bursuc, and X. Perrotton, "Plop: Probabilistic polynomial objects trajectory planning for autonomous driving," *CoRR*, vol. abs/2003.08744, 2020.

[8] Z. Su, C. Wang, H. Cui, N. Djuric, C. Vallespi-Gonzalez, and D. Bradley, "Temporally-continuous probabilistic prediction using polynomial trajectory parameterization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3837–3843, IEEE, 2021.

[9] J. Reichardt, "Trajectories as markov-states for long term traffic scene prediction," in *14-th UniDAS FAS-Workshop*, p. 14, 2022.

[10] L. Feng, M. Bahari, K. M. B. Amor, É. Zablocki, M. Cord, and A. Alahi, "Unitraj: A unified framework for scalable vehicle trajectory prediction," in *European Conference on Computer Vision*, pp. 106–123, Springer, 2025.

[11] Y. Yao, D. Goehring, and J. Reichardt, "An empirical bayes analysis of object trajectory representation models," in *IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 902–909, IEEE, 2023.