

# EP-Diffuser: An Efficient Diffusion Model for Traffic Scene Generation and Prediction via Polynomial Representations

Yue Yao , Mohamed-Khalil Bouzidi , Graduate Student Member, IEEE, Daniel Goehring , and Joerg Reichardt 

**Abstract**—As the prediction horizon increases, predicting the future evolution of traffic scenes becomes increasingly difficult due to the multi-modal nature of agent motion. Most state-of-the-art (SotA) prediction models primarily focus on forecasting the most likely future. However, for the safe operation of autonomous vehicles, it is equally important to cover the distribution for plausible motion alternatives. To address this, we introduce EP-Diffuser, a novel parameter-efficient diffusion-based generative model designed to capture the distribution of possible traffic scene evolutions. Conditioned on road layout and agent history, our model acts as a predictor and generates diverse, plausible scene continuations. We benchmark EP-Diffuser against two SotA models in terms of plausibility, diversity, and accuracy of predictions on the Argoverse 2 dataset. Despite its significantly smaller model size, our approach achieves both highly plausible and diverse traffic scene predictions with comparable accuracy. We further evaluate model generalization in an out-of-distribution (OoD) test setting using Waymo Open dataset and show superior robustness of our approach.

**Index Terms**—Autonomous agents, deep learning methods, performance evaluation and benchmarking.

## I. INTRODUCTION

**T**RAFFIC is a complex phenomenon where multiple agents interact in shared space and influence each other's behavior. For autonomous vehicles to integrate safely into such dynamic environments, they must anticipate how traffic scenes will evolve over time. This prediction capability is essential for downstream planning and decision-making processes.

Public motion datasets, such as Argoverse 2 (A2) [1] and Waymo Open (WO) [2], provide real-world traffic scene data and host associated motion prediction competitions to advance research in this field. The evolution of traffic scenes over long time

Received 3 March 2025; accepted 8 July 2025. Date of publication 1 August 2025; date of current version 12 August 2025. This article was recommended for publication by Associate Editor C. Ioan Vasile and Editor L. Pallottino upon evaluation of the reviewers' comments. This work was supported by the German Federal Ministry for Economic Affairs and Energy within the project "NXT GEN AI METHODS". (Corresponding author: Yue Yao.)

Yue Yao, Mohamed-Khalil Bouzidi, and Daniel Goehring are with the Department of Mathematics and Computer Science, Freie Universität Berlin, 14195 Berlin, Germany (e-mail: yue.yao@fu-berlin.de; mohamed-khalil.bouzidi@fu-berlin.de; daniel.goehring@fu-berlin.de).

Joerg Reichardt is with the Continental Automotive GmbH, 30165 Hannover, Germany (e-mail: joerg.reichardt@continental.com).

The code and model checkpoints are available at: <https://github.com/continental/EP-Diffuser>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/LRA.2025.3595042>, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3595042

horizons is governed by an inherently multi-modal probability distribution. Multiple plausible futures exist depending on road topology and agent interactions. However, motion datasets can only record a single observed future sample (the ground truth) per scene. As a result, motion prediction competitions typically frame the problem as a *regression* problem, where models are trained to estimate the most likely outcome based on available ground truth data. Existing approaches can be categorized into two main types:

- *Marginal Prediction*: Forecasting individual agent trajectories without ensuring that they combine into consistent scenes.
- *Joint Prediction*: Modeling multiple interacting agents simultaneously to predict coherent traffic scene continuations, a task we refer to as *traffic scene prediction*.

Regression-based metrics, such as Average Displacement Error (ADE), Final Displacement Error (FDE), and their variants, are widely used in these competitions. These metrics evaluate prediction accuracy by measuring how closely predictions match ground truth trajectories. However, this evaluation approach presents two key limitations illustrated in Fig. 1: First, it does not measure the diversity and coverage of predictions. Second, it fails to account for plausibility and consistency, both of which are essential for safe trajectory planning. Furthermore, focusing solely on the most probable future evolution while ignoring other plausible possibilities may induce overconfident and risky behavior, making it insufficient for planning algorithms, as highlighted in recent trajectory planning studies [3], [4], [5], [6].

Regression-based approaches have attempted to improve prediction diversity using discrete trajectory sets [7] or modified training objectives [8], [9]. In parallel, other studies have explored *generative* modeling frameworks. These include methods based on GANs [10], [11], VAEs [12], and diversity-enhancing sampling methods [13], [14], which initially aim to produce diverse marginal predictions for individual agents.

Building on this trend, recent generative approaches have begun addressing the more challenging task of multi-agent *traffic scene generation*. These models aim to learn the distribution of future traffic scenes under given conditions (i.e., observed agent history and road layout) and enable sampling from the modeled joint distribution. Therefore, they can capture not only the most likely traffic scene evolution but also a range of plausible alternatives.

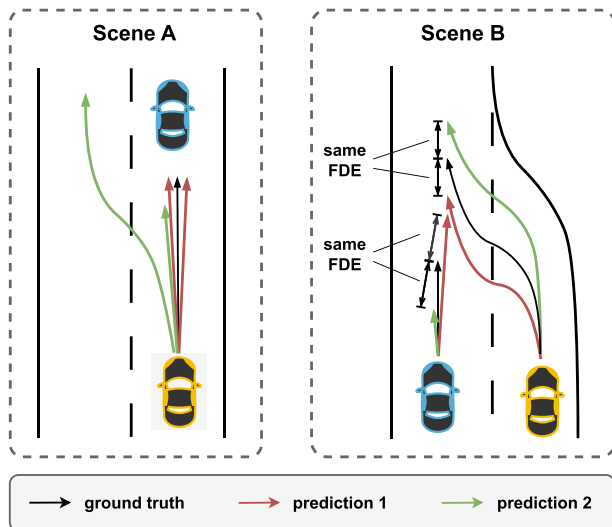


Fig. 1. Two limitations of regression-based metrics. **Scene A:** An example of *multi-modal marginal* prediction. While Prediction 1 yields a lower Final Displacement Error (FDE), it only captures one possible behavior and ignores other plausible maneuvers. **Scene B:** An example of *uni-modal joint* prediction. While both predictions yield the same FDE, Prediction 1 is less plausible due to the agent collision.

However, evaluating generative models is non-trivial, as it requires assessing the modeled distribution rather than the matching to a single ground truth trajectory. Many recent generative studies choose to inherit evaluation metrics from regression-based approaches, despite fundamental differences in modeling objectives [15], [16], [17]. This may bias models towards a narrow range of plausible future evolutions.

Another fundamental challenge is assessing model generalization. Prior studies typically evaluate model performance using a test split from the same dataset used for training. While motion datasets attempt to ensure disjoint splits between training and testing, these subsets still share underlying biases, such as recurring road layouts, traffic flow patterns, and artifacts introduced during data collection and pre-processing. Generative models are known to exhibit memorization – the ability to produce near-replicas of training data [18], [19]. Consequently, models tested solely on these similar samples may appear to perform well by leveraging their memorizing capability instead of learning robust, transferable representations of traffic patterns.

Hence, it is essential to rigorously assess generalization ability also for generative models under out-of-distribution (OoD) conditions, where the model cannot rely on memorized examples. Prior studies highlight notable distribution shifts across real-world motion datasets, such as variations in road layouts, traffic densities, and agent behaviors [20], [21]. These shifts present an opportunity to test whether generative models can extrapolate to unseen traffic scenes rather than merely recalling training patterns.

With these considerations in mind, we present Everything Polynomial Diffuser (EP-Diffuser), a generative model for traffic scene generation conditioned on road layout and observed agent history, thereby performing joint prediction. Unlike traditional models that incorporate sequence-based data – such

as lists of trajectory observations or map points – or those that use polynomials solely as outputs [22], our model employs polynomial representations for both map elements and trajectories on the model’s input and output sides. To comprehensively evaluate our approach, we extend beyond regression-based metrics and benchmark EP-Diffuser against two state-of-the-art (SotA) models from the perspectives of plausibility, coverage, and accuracy. Specifically, we incorporate Waymo’s “Sim Agents” metrics [23] for evaluating plausibility and assess performance under OoD scenarios from the WO dataset. Our results demonstrate that the polynomial representation enhances the efficiency of the denoising process, temporal consistency in generated agent kinematics, and generalization in OoD test settings. Our contributions are summarized as follows:

- We propose a novel diffusion model for generating diverse and highly realistic traffic scenes on the Argoverse 2 Motion dataset by using polynomial representations.
- We compare our model with two SotA models, highlighting a significant disconnect between regression-based metrics and the plausibility of predicted traffic scenes.
- We demonstrate superior generalization capabilities of our approach under out-of-distribution (OoD) conditions.

Our paper is organized as follows: We first review recent traffic scene prediction and generation models, along with their evaluation metrics. We then introduce two benchmark models and the “Sim Agents” metrics, highlighting key differences from regression-based metrics. Next, we present our diffusion-based approach with constrained parametric representations. We evaluate our model against benchmarks, analyzing plausibility, coverage, and regression-based performance. Finally, we extend our evaluation to OoD scenes to assess generalization beyond the training data.

## II. RELATED WORK AND PRELIMINARIES

### A. Traffic Scene Prediction and Benchmark Models

Benchmark datasets and associated prediction competitions have significantly shaped research in traffic scene prediction by framing it as a regression task, evaluating the most likely predicted traffic scenes. Recent studies have followed this competition framework and implemented regression-based deep learning models for traffic scene prediction [24], [25], [26]. Although these models can output multiple modes for potential traffic scenes, they are primarily scored and ranked using regression-based metrics such as minimum ADE (minADE) and minimum FDE (minFDE). These variants of ADE and FDE are tailored to multi-modal predictors and measure the minimum displacement error among all predicted modes.

Regression-based approaches have significantly influenced the development of generative models in the field [16], [17], [27]. Notably, many diffusion-based models have incorporated regression model backbones to output initial predictions that closely align with ground truth data [16], [28]. Although this design choice effectively optimizes for competition results, it may not fully capture the inherent uncertainty of real-world traffic.

TABLE I  
SUMMARY OF MODELS UNDER STUDY

model	FMAE-MA [24]	OptTrajDiff [16]	EP-Diffuser (ours)
input & output representation	sequence	sequence	polynomial
model type	regression	diffusion	diffusion
# output samples	6	inf	inf
# model parameters [million]	1.9	12.5	3.0

Cross-dataset testing introduces a notable distribution shift [20], [21] and is well suited for evaluating OoD generalization. We focus on models trained on the smaller A2 dataset, enabling a more rigorous assessment of their robustness when applied to scenes from the larger WO dataset. While many open-source models exist for the multi-modal marginal prediction task, there are far fewer open-source benchmark models specifically addressing multi-modal traffic scene (joint) prediction with documented performance and reproducible results. This limitation narrows the pool of suitable benchmark models for our experiment.

As representatives of the two model classes, we select two recently open-sourced and thoroughly documented SotA models as benchmarks: Forecast-MAE-multiagent (FMAE-MA) [24] and OptTrajDiff [16]. As summarized in Table I, both models use sequence-based representations but follow different methodological approaches. FMAE-MA follows a regression-based approach and predicts 6 distinct modes of future traffic scenes, with a relatively lightweight architecture of 1.9 million parameters. OptTrajDiff adopts a diffusion-based approach and incorporates QCNNet [31] as its regression backbone, with a total of 12.5 million parameters.

### B. Sim Agents Metrics

In contrast to regression-based tasks, “Sim Agents” frames traffic scene prediction as a multi-agent generative task, emphasizing the importance of capturing the diversity and realism of traffic behaviors [23]. Rather than focusing solely on minimizing prediction errors, Sim Agents metrics evaluate models based on their ability to produce plausible and consistent traffic scenes using several complementary metrics:

- *Agent Kinematic Metrics*: Evaluate the kinematic properties of individual agents, such as speed, acceleration, and adherence to realistic motion patterns.
- *Agent Interaction Metrics*: Measure the quality of interactions between agents, ensuring that predicted behaviors reflect realistic social dynamics and comply with traffic rules.
- *Map Adherence Metrics*: Assess whether predicted trajectories conform to road layouts, lane boundaries, and other map-related constraints.
- *Realism Meta Metric*: An aggregated score that combines all above evaluations into a holistic measure of scene realism and consistency.

These metrics are calculated by comparing the distribution approximated from 32 predicted samples against real-world data, encouraging models to replicate the variability and interaction patterns observed in actual traffic scenes. All results are normalized scores in the interval [0,1], with 1 indicating the highest score.

In this work, we use the “realism meta” metric as the primary measure of predicted scene plausibility.

### III. DATA REPRESENTATION AND MODEL

EP-Diffuser extends our previous marginal prediction model, EP [20], by introducing a generative diffusion framework and enabling multi-agent joint trajectory generation. Unlike other diffusion models that use sequence-based representations [15], [16], [28], our approach integrates polynomial representations for map elements and trajectories.

This polynomial design offers three key benefits. First, it provides a low-dimensional, truncated Taylor-like approximation that captures key motion patterns without modeling complex high-order residuals. This improves the efficiency of the diffusion-denoising process and simplifies the overall learning task. Second, since polynomial basis functions possess well-defined derivatives, training with positional loss implicitly influences kinematic terms such as velocity and acceleration. This yields smooth and physically plausible trajectories when appropriately regularized with moderate polynomial degrees. Third, it enhances model generalization by mitigating dataset-specific bias and noise commonly present in sequence-based representations, as highlighted in [20]. Unlike data-dependent dimensionality reduction methods such as Principal Component Analysis (PCA) used in [17], polynomial basis functions are data-independent and support better generalization across datasets.

In the subsequent sections, we first describe our approach for representing diverse data types using polynomial representations, followed by the implementation details of our model.

#### A. Data as Polynomials

Different types of polynomials of the same degree – such as monomial, Bernstein, and Chebyshev polynomials – can be linearly transformed into one another [32], offering the same expressiveness without affecting the model’s capacity. In this work, we employ Bernstein polynomials to represent agent histories, future trajectories, and map geometry, though our approach is not restricted to this basis. The parameters of Bernstein polynomials have spatial semantics as *control points*, making them particularly convenient for coordinate transformations such as translation, rotation, and scaling.

The A2 dataset segments each 11-second recording into a 5-second history and a 6-second future. The choice of polynomial degrees for agent trajectories has been extensively analyzed in our prior work [33], which examines the trade-off between representation fidelity and polynomial degree across various agent types and trajectory lengths in both the A2 and WO datasets. Based on the suggestions in [33], we represent different data types as follows:

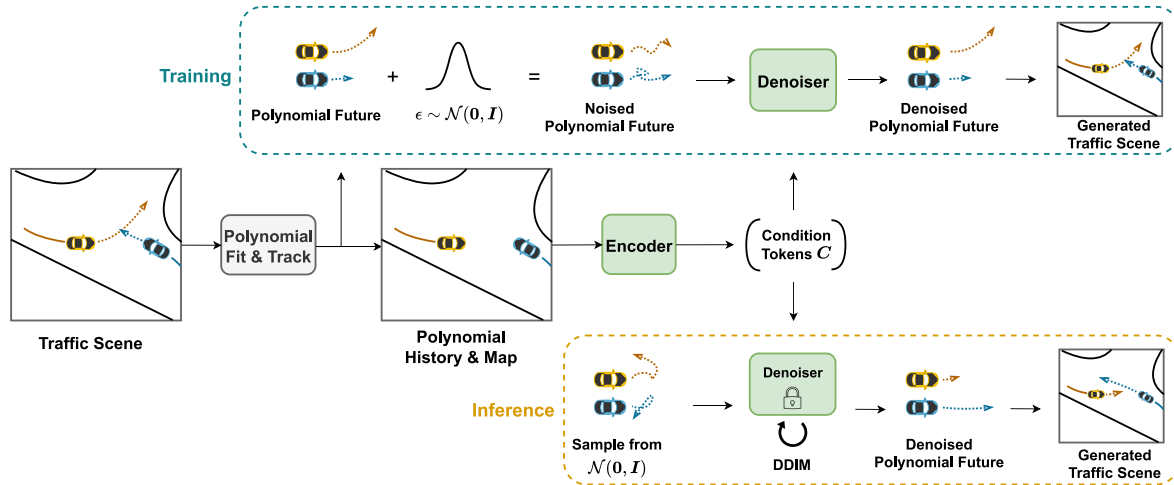


Fig. 2. Overview of EP-Diffuser for traffic scene generation. For clear presentation, we adopt a similar visual layout to [17], while the technical implementation extends our previous EP model [20]. The traffic scene comprised of agent history and map elements as degree 5 and 3 polynomials, respectively, is encoded via a transformer encoder [29] into a set of condition tokens  $\mathcal{C}$ . The ground truth (GT) future trajectories are represented as polynomials of degree 6. During **training**, a random set of noise is sampled i.i.d. from a standard normal distribution and added to the parameters of the ground truth future trajectory. The denoiser, while attending to the condition tokens, jointly predicts the denoised polynomial parameters of trajectories corresponding to each agent. During **inference**, a set of trajectory parameters for each agent is initially sampled from a standard normal distribution, and iteratively denoised using a DDIM schedule [30] to produce plausible future trajectories.

- **Agent History:** Following the Akaike Information Criterion (AIC) [34] from the study [33], we represent 5-second history trajectories of vehicles, cyclists, and pedestrians in A2 using optimal 5-degree polynomials. We also use the 5-degree polynomial for the ego vehicle. We track the control points of agent history with the method proposed in [32] and incorporate the observation noise models in [33].
- **Agent Future:** We model 6-second future trajectories as 6-degree polynomials – one degree higher than suggested by AIC in [33] to better capture complex trajectories. Bayesian Regression is applied to fit agent future trajectories, following priors and observation noise models from [33].
- **Map:** Map elements, such as lane segments and crosswalks, are represented with 3-degree polynomials, aligning with OpenDRIVE [35] standards. We fit the sample points of map elements via the total-least-squares method by Borges-Pastva [36].

## B. Model Architecture

The pipeline of EP-Diffuser is illustrated in Fig. 2. The model employs an encoder-denoiser architecture, with implementation details provided in Appendix A. To mitigate variability in sampling procedures and accelerate the inference process, we adopt the Denoising Diffusion Implicit Models (DDIM) method [30] with 10 denoising steps – consistent with the configuration of OptTrajDiff (details in Appendix A-A).

## IV. EXPERIMENTAL RESULTS

### A. Experiment Setup

1) **In-Distribution Training and Testing on A2:** Following the Argoverse 2 Motion Prediction Competition setup, models are

required to output 6-second predictions given 5-second history. All models are trained with this setup from scratch with their original hyperparameters on the A2 training split containing 199,908 scenarios.

The “Sim Agents” evaluation requires 32 modeled traffic scene samples for metric computation. For generative models, we randomly sample 32 generated traffic scenes. For FMAE-MA, which by design outputs 6 different predictions, we use an ensemble of 6 independently trained models, each initialized with a different seed, predicting 36 traffic scenes in total. From these, we select the 32 predictions with the highest predicted probabilities for evaluation.

2) **Out-of-Distribution Testing on WO:** Cross-dataset testing presents challenges due to inconsistencies in data formats and prediction tasks. To address this, we adopt the homogenization protocol from [20] to enable cross-dataset evaluation between the A2 and WO datasets. This protocol aligns WO’s prediction task with the A2’s competition setup by incorporating a 5-second history. Since WO recordings are shorter (9.1 seconds), we evaluate only the first 4.1 seconds of the 6-second predictions. Additionally, models are restricted to only considering lane centers and crosswalks as available map information due to their availability across both datasets.

For OoD testing, we apply the same sampling strategies as in A2. All three models are trained on the homogenized A2 training split and tested on OoD samples from the homogenized WO validation split. Results are reported based on the 4.1-second prediction.

3) **Test Data Selection:** Due to the computational cost of “Sim Agents” metric calculations, we subsample the validation data from both the A2 and WO datasets under two distinct settings:

TABLE II  
RESULTS OF “SIM AGENTS” METRICS WITH STANDARD DEVIATION OVER TESTED SAMPLES

train / test	model	realism meta $\uparrow$	kinematic metrics $\uparrow$	interactive metrics $\uparrow$	map metrics $\uparrow$	minADE [m] $\downarrow$	coverage [m]	# model params [million]	
R20P	A2 / A2	FMAE-MA [24]	0.749 $\pm$ 0.117	0.459 $\pm$ 0.181	0.786 $\pm$ 0.149	0.867 $\pm$ 0.187	0.333 $\pm$ 0.295	2.000 $\pm$ 3.672	11.4 (6 $\times$ 1.9)
		OptTrajDiff [16]	0.795 $\pm$ 0.091	0.564 $\pm$ 0.166	0.805 $\pm$ 0.117	<b>0.915</b> $\pm$ 0.121	<b>0.325</b> $\pm$ 0.277	4.447 $\pm$ 6.818	12.5
		Seq-Diffuser (ablation)	0.748 $\pm$ 0.108	0.414 $\pm$ 0.195	0.789 $\pm$ 0.135	0.885 $\pm$ 0.161	0.331 $\pm$ 0.255	1.718 $\pm$ 3.714	3.1
		EP-Diffuser (ours)	<b>0.809</b> $\pm$ 0.087	<b>0.632</b> $\pm$ 0.149	<b>0.808</b> $\pm$ 0.105	0.913 $\pm$ 0.129	0.398 $\pm$ 0.340	4.178 $\pm$ 7.127	3.0
	A2 / WO	FMAE-MA [24]	0.709 $\pm$ 0.102	0.309 $\pm$ 0.111	0.800 $\pm$ 0.125	0.819 $\pm$ 0.191	0.363 $\pm$ 0.203	3.560 $\pm$ 4.733	11.4 (6 $\times$ 1.9)
		OptTrajDiff [16]	0.768 $\pm$ 0.068	0.420 $\pm$ 0.107	0.827 $\pm$ 0.087	0.890 $\pm$ 0.114	<b>0.310</b> $\pm$ 0.227	3.496 $\pm$ 4.692	12.5
		Seq-Diffuser (ablation)	0.688 $\pm$ 0.079	0.212 $\pm$ 0.093	0.801 $\pm$ 0.090	0.814 $\pm$ 0.154	0.458 $\pm$ 0.286	2.446 $\pm$ 4.949	3.1
		EP-Diffuser (ours)	<b>0.788</b> $\pm$ 0.063	<b>0.491</b> $\pm$ 0.133	<b>0.834</b> $\pm$ 0.073	<b>0.900</b> $\pm$ 0.096	0.348 $\pm$ 0.307	3.224 $\pm$ 5.603	3.0
C500	A2 / A2	FMAE-MA [24]	0.636 $\pm$ 0.125	0.320 $\pm$ 0.118	0.679 $\pm$ 0.173	0.760 $\pm$ 0.241	0.479 $\pm$ 0.371	2.780 $\pm$ 4.155	11.4 (6 $\times$ 1.9)
		OptTrajDiff [16]	0.709 $\pm$ 0.094	0.459 $\pm$ 0.119	<b>0.717</b> $\pm$ 0.128	<b>0.841</b> $\pm$ 0.176	<b>0.449</b> $\pm$ 0.329	5.361 $\pm$ 7.240	12.5
		Seq-Diffuser (ablation)	0.634 $\pm$ 0.111	0.279 $\pm$ 0.111	0.674 $\pm$ 0.153	0.786 $\pm$ 0.212	0.467 $\pm$ 0.298	2.020 $\pm$ 4.188	3.1
		EP-Diffuser (ours)	<b>0.713</b> $\pm$ 0.094	<b>0.507</b> $\pm$ 0.116	0.707 $\pm$ 0.124	0.838 $\pm$ 0.178	0.546 $\pm$ 0.315	4.500 $\pm$ 7.075	3.0
	A2 / WO	FMAE-MA [24]	0.630 $\pm$ 0.123	0.271 $\pm$ 0.079	0.716 $\pm$ 0.159	0.723 $\pm$ 0.238	0.426 $\pm$ 0.229	3.855 $\pm$ 5.060	11.4 (6 $\times$ 1.9)
		OptTrajDiff [16]	0.721 $\pm$ 0.085	0.403 $\pm$ 0.092	0.771 $\pm$ 0.106	0.839 $\pm$ 0.163	<b>0.357</b> $\pm$ 0.201	3.189 $\pm$ 4.527	12.5
		Seq-Diffuser (ablation)	0.630 $\pm$ 0.098	0.210 $\pm$ 0.069	0.739 $\pm$ 0.126	0.730 $\pm$ 0.196	0.448 $\pm$ 0.206	1.980 $\pm$ 4.327	3.1
		EP-Diffuser (ours)	<b>0.742</b> $\pm$ 0.072	<b>0.456</b> $\pm$ 0.099	<b>0.782</b> $\pm$ 0.090	<b>0.854</b> $\pm$ 0.134	0.372 $\pm$ 0.175	2.786 $\pm$ 5.244	3.0

The best value for each metric across models is highlighted in bold. **Upper Section:** Evaluated on the randomly selected 20% subset from Argoverse 2 and Waymo validation sets. **Lower Section:** Evaluated on the 500 most challenging scenes from Argoverse 2 and Waymo validation sets. **In-Distribution Testing:** Performed on the Argoverse 2 validation set with a 6-second prediction horizon. **Out-of-Distribution Testing:** Performed on the homogenized Waymo validation set with a 4.1-second prediction horizon.

- *20% Random Subsample:* We randomly select 20% of the validation set from each dataset, corresponding to 5,000 samples from A2 and 8,400 samples from WO. This setting aims to reflect the model’s overall performance across typical scenes presented by datasets. We refer to this subset as **R20P** in the remainder of the paper.
- *500 Most Challenging Scenes:* We identify and select the 500 most difficult traffic scenes in each validation set based on the largest deviations between ground truth trajectories and those generated by a constant velocity model, as measured by the realism meta metric. This setting emphasizes model performance under more complex and demanding conditions. We refer to this subset as **C500** in the remainder of the paper.

4) *Coverage Metric:* Inspired by the inter-policy diversity metric in [37], we compute the average pairwise distance between the final positions of the 32 sampled trajectories to measure the trajectory diversity. This metric serves as an **auxiliary** indicator and is not used as a standalone measure of performance – since high coverage is only meaningful when accompanied by high plausibility.

## B. Comparison With SotA Models on A2

Table II (“A2/A2” sections) summarizes the in-distribution results of all models tested on the A2 validation set across both the R20P and C500 subsets. Fig. 3 visualizes multiple samples generated by EP-Diffuser.

We observe a reversed ranking when evaluating prediction plausibility and *minADE*. For example, on R20P, despite its smaller model size (3 million parameters) and the largest *minADE* (0.398m), EP-Diffuser achieves the highest realism meta score at 0.809 and performs comparably to OptTrajDiff in terms of agent interaction and map adherence,

while also maintaining high coverage (4.178m). Notably, EP-Diffuser demonstrates superior agent kinematics (0.632), outperforming OptTrajDiff (0.564) and FMAE-MA (0.456).

As an illustrative example, Fig. 4 visualizes the vehicle kinematics during a left-turn maneuver in A2. EP-Diffuser produces the most plausible agent kinematics according to the measures in [38].

Fig. 5 reports the realism meta scores and inference time of diffusion-based models over a range of DDIM denoising steps, with FMAE-MA as the regression-based baseline for comparison. Across all denoising step configurations, EP-Diffuser consistently outperforms OptTrajDiff in both metrics, achieving higher realism scores with lower inference time. While EP-Diffuser’s inference time exceeds FMAE-MA’s, it reaches peak performance with 5 denoising steps in just 55.3ms, making it viable for real-time applications.

## C. OoD Testing on WO

In the OoD Setting, the model trained on A2 is asked to generate traffic scene continuations for independent scenes taken from the WO dataset – thus removing any sort of shared bias between training and test data. The OoD testing results are presented in the “A2/WO” sections in Table II.

In the OoD setting, EP-Diffuser maintains the top scores in realism meta and agent kinematics. Additionally, it also demonstrates improved performance by achieving the best agent interactions and map adherence scores on both subsets. Furthermore, EP-Diffuser outperforms FMAE-MA in *minADE* and closely matches OptTrajDiff (0.372m vs. 0.357m on C500, respectively) with high coverage (2.786 m). These results demonstrate EP-Diffuser’s ability to learn robust, transferable representations from training data, highlighting its enhanced generalization beyond dataset-specific patterns.

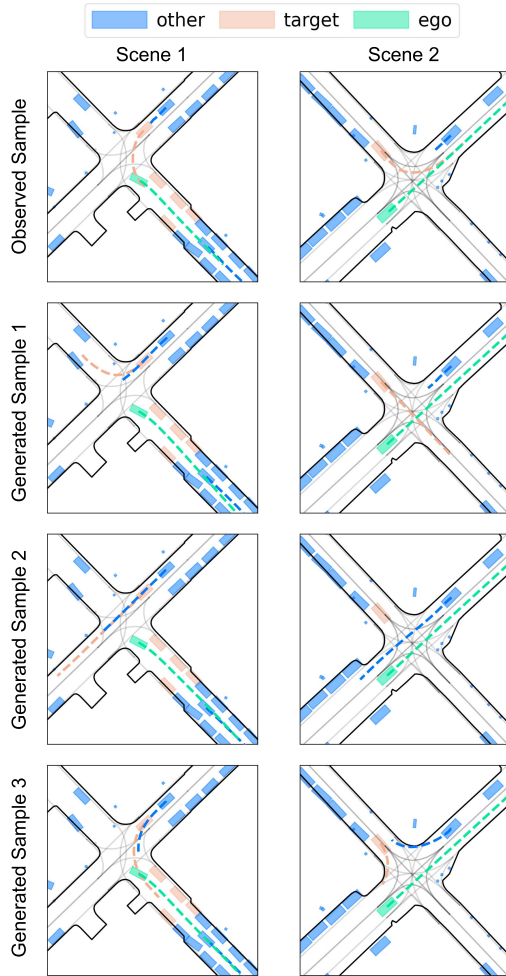


Fig. 3. Observed sample (ground truth) and generated traffic scene samples from EP-Diffuser for two traffic scenes in Argoverse 2, demonstrating EP-Diffuser’s capability to generate diverse and plausible traffic scenes. **Dashed lines** represent the future trajectories of selected highly interactive agents.

#### D. Ablation Study

To evaluate the impact of polynomial representations, we introduce a variant, **Seq-Diffuser**, which replaces polynomial inputs and outputs in EP-Diffuser with sequence-based representations. Only minor modifications to the model architecture are required to accommodate the increased dimensionality of the sequence-based representation. The training and diffusion settings are identical for both EP-Diffuser and Seq-Diffuser. Comparative results are presented in Table II.

While Seq-Diffuser achieves better in-distribution accuracy with lower *minADE* (e.g. 0.331 m vs. 0.398 m on R20P), it exhibits a notable degradation in both the plausibility and diversity of the generated traffic scenes across both subsets. Moreover, Seq-Diffuser generalizes poorly, as reflected by its lower performance in plausibility and higher OoD *minADE* (e.g. 0.458 m vs. 0.348 m on R20P). It also incurs a higher inference time, as shown in Fig. 5.

These results indicate that, although sequence-based representations may capture finer-grained details beneficial for pointwise accuracy, polynomial representations lead to better

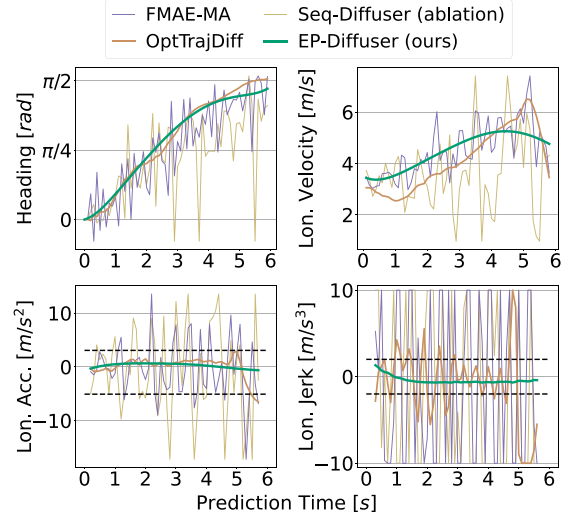


Fig. 4. Predicted heading, longitudinal velocity, acceleration, and jerk for a vehicle’s left turn maneuver over a 6-second time horizon in A2. For clarity, the results of Seq-Diffuser, as well as the jerk outputs of FMAE-MA and OptTrajDiff, are clipped. **Dashed lines** indicate the ranges of longitudinal acceleration and jerk for aggressive human drivers based on [38], suggesting that the agent kinematics of EP-Diffuser are the most plausible.

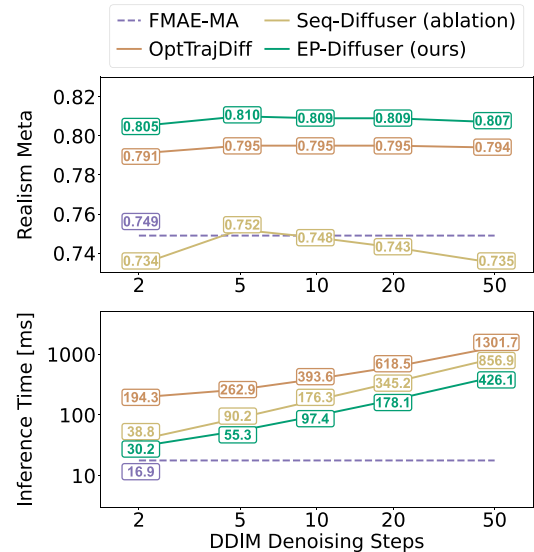


Fig. 5. Realism meta score on A2 R20P and inference time for diffusion models across different DDIM denoising steps, with FMAE-MA as the regression-based baseline for comparison. Both the denoising steps and inference time are shown on *log scale*. Inference time is measured by predicting 6 samples of an Argoverse 2 traffic scene with 50 agents and 150 map elements, using a single A10 G GPU.

plausibility, diversity, and OoD generalization in multi-agent scene generation.

#### V. CONCLUSION

Traffic scene prediction constitutes a fundamental challenge in autonomous driving due to its inherently multi-modal nature. Most motion prediction competitions aim for prediction accuracy, encouraging models to focus on reproducing observed behaviors rather than capturing the diversity of plausible future scene evolutions. In this work, we introduced EP-Diffuser,

a novel diffusion-based framework that leverages polynomial representations to efficiently model agent trajectories and road geometry. Through extensive experiments on Argoverse 2 and Waymo Open datasets, we demonstrated that EP-Diffuser not only improves the plausibility of generated traffic scenes but also generalizes well to out-of-distribution (OoD) environments. Notably, EP-Diffuser’s computational efficiency addresses a fundamental limitation of diffusion models, making it viable for real-time applications.

Future work will focus on refining evaluation methodologies to further bridge the gap between prediction accuracy and coverage of plausible alternatives for real-world feasibility. Additionally, expanding EP-Diffuser to incorporate uncertainty-aware decision-making for autonomous vehicles presents an exciting direction for enhancing robustness in motion planning.

APPENDIX A  
IMPLEMENTATION DETAILS

A. Diffusion and Denoising

We perform diffusion-denoising on the polynomial parameters (control points) that represent each agent’s future trajectory. Specifically, we operate on the displacement vectors between consecutive control points of 6-degree Bernstein polynomials representing 2D future trajectories, denoted as  $\delta^{fut} \in \mathbb{R}^{12}$  for each agent.

Following the practice of Denoising Diffusion Probabilistic Models (DDPM) [39], we denote the diffused  $\delta^{fut}$  for the  $i$ -th agent at  $s$ -th diffusion step as  $\delta_{s,i}^{fut} \in \mathbb{R}^{12}$ . Here,  $s = 0$  corresponds to the fitted polynomial parameters without added noise. We can write  $\delta_{s,i}^{fut}$  as a linear combination of the initial parameters  $\delta_{0,i}^{fut}$  and a Gaussian noise  $\epsilon_i$ :

$$\delta_{s,i}^{fut} = \sqrt{\bar{\alpha}_s} \delta_{0,i}^{fut} + \sqrt{1 - \bar{\alpha}_s} \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (1)$$

where  $\bar{\alpha}_s$  is the noise-scheduling parameter at diffusion step  $s$  and controls the diffusion process.

We apply a total of  $S = 1000$  diffusion steps to gradually transition from the data distribution  $q(\delta_{0,i}^{fut})$  to the target prior distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . For the backward denoising process, we adopt Denoising Diffusion Implicit Models (DDIM) [30] – the same procedure as OptTrajDiff – with 10 denoising steps. The future trajectories of agents are iteratively denoised by predicting the added noise  $\hat{\epsilon}_i$  for each agent and subtracting  $\hat{\epsilon}_i$  from  $\delta_{s,i}^{fut}$  at each step.

B. Encoder

We employ the encoder architecture of our EP model [20] and adopt the “query-centric” design of QCNet [31], adapting it to operate on polynomial representations.

For all agents  $\mathcal{A}$  in a scene, we encode the 2D displacement vectors of agent history control points, modeled as 5-degree polynomials, denoted as  $\Delta^{hist} \in \mathbb{R}^{A \times 10}$ , where  $A$  is the number of agents. Additionally, we encode the time window of each agent’s appearance in history, represented as  $TW \in \mathbb{R}^{A \times 2}$ , along with the agent category information. These features

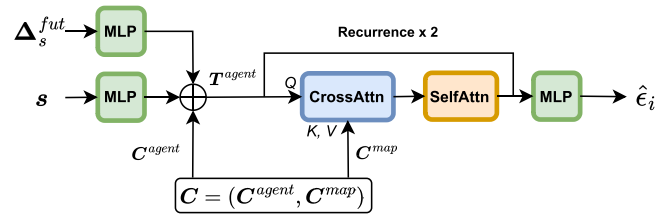


Fig. 6. Denoiser architecture of EP-Diffuser.

TABLE III  
EP-DIFFUSER TRAINING SETUP

hidden dimension $D$	128
$\beta_s$	$s * (\beta_{end} - \beta_{start}) / S + \beta_{start}$ , with $\beta_{end} = 0.2, \beta_{start} = 1e - 5, S = 1000$
optimizer	AdamW
learning rate	5e-4
learning rate schedule	cosine
batch size	32
training / warmup epochs	64 / 10
dropout	0.1

are summarized to form the agent condition tokens  $C^{agent} \in \mathbb{R}^{A \times D}$ , where  $D$  is the hidden dimension.

Similarly, for all map elements  $\mathcal{M}$  in a scene, we encode their control point vectors of 3-degree polynomials, denoted as  $\Delta^{map} \in \mathbb{R}^{M \times 6}$ , along with the corresponding map element categories. These features are summarized to form the map condition tokens  $C^{map} \in \mathbb{R}^{M \times D}$ , where  $M$  is the number of map elements.

C. Denoiser

We visualize the denoiser architecture in Fig. 6. The denoiser processes the vectors between control points of noised 6-degree future trajectories, denoted as  $\Delta_s^{fut} \in \mathbb{R}^{A \times 12}$ , along with the diffusion step indices for each agent  $s \in \mathbb{R}^A$ . These inputs are embedded and summarized with the agent condition tokens  $C^{agent}$  to form the agent tokens  $T^{agent} \in \mathbb{R}^{A \times D}$ . Multiple attention blocks based on Transformer [29] perform the agent-map and agent-agent attentions sequentially to update  $T^{agent}$ . Finally, the updated agent tokens  $T^{agent}$  is decoded to predict the added noise  $\hat{\epsilon}_i$  for each agent.

D. Training Loss

Following DDPM [39], the EP-Diffuser is trained to minimize the mean squared error (MSE) between the added noise  $\epsilon_i$  and predicted noise  $\hat{\epsilon}_i$  averaged across all agents:  $\mathcal{L} = \frac{1}{A} \sum_{i=1}^A \|\epsilon_i - \hat{\epsilon}_i\|_2^2$ .

E. Training Setup

We report the training setup for EP-Diffuser in Table III. The noise scheduling parameter is expressed as  $\bar{\alpha}_s = \prod_k^s \alpha_k$ , where  $\alpha_s = 1 - \beta_s$ .

## F. Post-Processing

We observe that stationary agents in recorded ground truth often exhibit minor positional shifts and unrealistic rotations, which can lead to unnatural behaviors in predicted scenes across all models. To address this, we adopt a lightweight post-processing step inspired by prior work [40]:

- *Stationary Agent Correction*: For each predicted trajectory, if an agent moves less than 1m over the prediction horizon, we classify it as non-moving and retain its last measured position and heading.

This ensures physically consistent behavior for stationary agents without modifying the model's core predictions. For fairness, we apply this post-processing step uniformly across both our model and benchmark models.

## ACKNOWLEDGMENT

The authors would like to thank Jonas Neuhoefer and Andreas Philipp for many helpful discussions. We also gratefully acknowledge the Gauss Center for Supercomputing for providing compute on JUWELS at JSC.

## REFERENCES

- [1] B. Wilson et al., "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Proc. Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021.
- [2] S. Ettinger et al., "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9710–9719.
- [3] Y. Chen, U. Rosolia, W. Ubellacker, N. Csomay-Shanklin, and A. D. Ames, "Interactive multi-modal motion planning with branch model predictive control," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 5365–5372, Apr. 2022.
- [4] M.-K. Bouzidi, B. Derajic, D. Goehring, and J. Reichardt, "Motion planning under uncertainty: Integrating learning-based multi-modal predictors into branch model predictive control," in *Proc. IEEE 27th Int. Conf. Intell. Transp. Syst.*, 2024, pp. 2592–2598.
- [5] K. A. Mustafa, D. J. Ornia, J. Kober, and J. Alonso-Mora, "RACP: Risk-aware contingency planning with multi-modal predictions," *IEEE Trans. Intell. Veh.*, vol. 10, no. 1, pp. 228–243, Jan. 2025.
- [6] M.-K. Bouzidi et al., "Closing the loop: Motion prediction models beyond open-loop benchmarks," 2025, *arXiv:2505.05638*.
- [7] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "CoverNet: Multimodal behavior prediction using trajectory sets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, 2020, pp. 14074–14083.
- [8] Y. Xu, V. Letzelter, M. Chen, É. Zablocki, and M. Cord, "Annealed winner-takes-all for motion forecasting," 2024, *arXiv:2409.11172*.
- [9] J. Lidard et al., "NashFormer: Leveraging local nash equilibria for semantically diverse trajectory prediction," 2023, *arXiv:2305.17600*.
- [10] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "SocialGAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2018, pp. 2255–2264.
- [11] D. Roy, T. Ishizaka, C. K. Mohan, and A. Fukuda, "Vehicle trajectory prediction at intersections using interaction based generative adversarial networks," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2019, pp. 2318–2323.
- [12] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 683–700.
- [13] X. Huang et al., "DiversityGAN: Diversity-aware vehicle motion prediction via latent semantic sampling," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 5089–5096, Oct. 2020.
- [14] Y. J. Ma, J. P. Inala, D. Jayaraman, and O. Bastani, "Diverse sampling for normalizing flow based trajectory forecasting," 2020, *arXiv:2011.15084*.
- [15] Y. Choi, R. C. Mercurius, S. M. A. Shabestary, and A. Rasouli, "DICE: Diverse diffusion model with scoring for trajectory prediction," in *Proc. IEEE Intell. Veh. Symp.*, 2024, pp. 3023–3029.
- [16] Y. Wang et al., "Optimizing diffusion models for joint trajectory prediction and controllable generation," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 324–341.
- [17] C. Jiang et al., "MotionDiffuser: Controllable multi-agent motion prediction using diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, 2023, pp. 9644–9653.
- [18] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
- [19] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Understanding and mitigating copying in diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 47783–47803.
- [20] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 488–495.
- [21] 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 *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 106–123.
- [22] X. Huang, S. G. McGill, B. C. Williams, L. Fletcher, and G. Rosman, "Uncertainty-aware driver trajectory prediction at urban intersections," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 9718–9724.
- [23] N. Montali et al., "The waymo open sim agents challenge," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 59151–59171.
- [24] J. Cheng, X. Mei, and M. Liu, "Forecast-MAE: Self-supervised pre-training for motion forecasting with masked autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8679–8689.
- [25] Z. Zhou, Z. Wen, J. Wang, Y.-H. Li, and Y.-K. Huang, "QCNext: A next-generation framework for joint multi-agent trajectory prediction," 2023, *arXiv:2306.10508*.
- [26] W. Luo, C. Park, A. Cornman, B. Sapp, and D. Anguelov, "JFP: Joint future prediction with interactive multi-agent modeling for autonomous driving," in *Proc. Conf. Robot Learn.*, 2023, pp. 1457–1467.
- [27] A. Seff et al., "MotionLM: Multi-agent motion forecasting as language modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8579–8590.
- [28] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, "Leapfrog diffusion model for stochastic trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*, 2023, pp. 5517–5526.
- [29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [30] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [31] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17863–17873.
- [32] J. Reichardt, "Trajectories as Markov-states for long term traffic scene prediction," in *Proc. 14th UniDAS FAS-Workshop*, 2022, pp. 21–34.
- [33] Y. Yao, D. Goehring, and J. Reichardt, "An empirical Bayes analysis of object trajectory representation models," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2023, pp. 902–909.
- [34] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Berlin, Germany: Springer, 1998, pp. 199–213.
- [35] Association for standardization of automation and measuring systems (ASAM), "ASAM OpenDRIVE," 2024. Accessed: Nov. 24, 2024. [Online]. Available: <https://www.asam.net/standards/detail/opendrive/>
- [36] C. F. Borges and T. Pastva, "Total least squares fitting of bézier and B-spline curves to ordered data," *Comput. Aided Geom. Des.*, vol. 19, no. 4, pp. 275–289, 2002.
- [37] S. Shiroshita et al., "Behaviorally diverse traffic simulation via reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 2103–2110.
- [38] I. Bae et al., "Self-driving like a human driver instead of a robocar: Personalized comfortable driving experience for autonomous vehicles," 2020, *arXiv:2001.03908*.
- [39] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
- [40] W. Zeng, S. Wang, R. Liao, Y. Chen, B. Yang, and R. Urtasun, "DSDNet: Deep structured self-driving network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 156–172.