UniTraj: A Unified Framework for Scalable Vehicle Trajectory Prediction

Lan Feng, Mohammadhossein Bahari, Kaouther Messaoud Ben Amor, Éloi Zablocki, Matthieu Cord, Alexandre Alahi

UniTraj: A Unified Framework for Scalable Vehicle Trajectory Prediction

Lan Feng[*,1], Mohammadhossein Bahari[*,1], Kaouther Messaoud Ben Amor[1], Éloi Zablocki[2], Matthieu Cord[2,3], and Alexandre Alahi[1]

[1] EPFL, Switzerland,
firstname.lastname@epfl.ch
[2] Valeo.ai, France,
[3] Sorbonne Université, France

Abstract. Vehicle trajectory prediction has increasingly relied on data-driven solutions, but their ability to scale to different data domains and the impact of larger dataset sizes on their generalization remain under-explored. While these questions can be studied by employing multiple datasets, it is challenging due to several discrepancies, *e.g.,* in data formats, map resolution, and semantic annotation types. To address these challenges, we introduce UniTraj, a comprehensive framework that unifies various datasets, models, and evaluation criteria, presenting new opportunities for the vehicle trajectory prediction field. In particular, using UniTraj, we conduct extensive experiments and find that model performance significantly drops when transferred to other datasets. However, enlarging data size and diversity can substantially improve performance, leading to a new state-of-the-art result for the nuScenes dataset. We provide insights into dataset characteristics to explain these findings. The code can be found here: https://github.com/vita-epfl/UniTraj.

Keywords: Vehicle trajectory prediction · Multi-dataset framework · Domain generalization

1 Introduction

Predicting the trajectories of surrounding vehicles is essential for ensuring the safety and collision avoidance of autonomous driving systems. With the advent of deep learning, researchers have turned to data-driven solutions to tackle this prediction task. However, while these models can achieve high accuracy, they are heavily reliant on the specific data domain used for training.

An autonomous driving system may encounter various situations such as diverse geographical locations. These various situations introduce data domain shifts, which can significantly impact the performance of the prediction models. Consequently, it is essential to study the performance of the models across

[*] Equal contribution as the first authors.

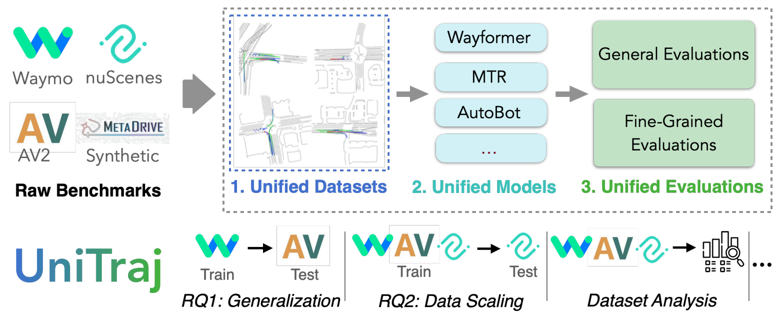


Fig. 1: UniTraj framework. The framework unifies various datasets, forming the largest vehicle trajectory prediction dataset available. It also includes multiple state-of-the-art prediction models and various evaluation strategies, making it suitable for trajectory prediction experimentation. The framework enables the study of diverse Research Questions, including (RQ1) the generalization of trajectory prediction models across different domains and (RQ2) the impact of data size on prediction performance.

diverse domains, such as datasets and cities. However, despite the importance of the question, the generalization of models to different domains has not been adequately studied yet. Therefore, our first Research Question (RQ1) is to investigate the performance drop of trajectory prediction models when transferred to new domains.

A potential solution to improve the generalization ability of prediction models is to scale up the sizes of the datasets to cover a broader spectrum of driving scenarios. While there is a trend in extending datasets' sizes [8,10,14,46], the impact of dataset size on the performance of trajectory prediction models remains largely unexplored. Our second research question (RQ2) is then to study the impact of increasing dataset sizes on the performance of the prediction models.

Exploring these two research questions involves leveraging multiple trajectory prediction datasets. Firstly, these datasets provide diverse domains, allowing for a thorough examination of model generalization across different domains (RQ1). Secondly, combining these datasets creates a much larger dataset, enabling an exploration of the asymptotic limits of data scaling (RQ2). However, significant challenges exist when attempting to leverage multiple datasets. (1) Each of these datasets has a unique data format, posing practical difficulties for researchers utilizing multiple datasets. (2) Each of the datasets undergoes collection and annotation through distinct strategies, with semi-automatic pre-annotations and manual curations [7,8,14]. This leads to multiple discrepancies such as variations in resolution, sampling rates, and types of semantic annotations. (3) Comparing model performance across datasets is not straightforward due to varying dataset settings (e.g., prediction horizons) and evaluation metrics (e.g., mAP metric is used in WOMD [14] and brier-FDE metric in Argoverse 2 [46]). In short, while each of the datasets contributes to the progress in the field, they have been developed independently, without considering harmonization with existing ones.

As a result, many trajectory prediction studies train and evaluate their models using a single dataset [2,3,5,9,13,18,20,26,30,32,42].

To tackle these challenges, we introduce 'UniTraj', a comprehensive vehicle trajectory prediction framework. UniTraj seamlessly integrates and unifies multiple data sources (including nuScenes [7], Argoverse 2 [46] and Waymo Open Motion Dataset - WOMD [14]), models (including AutoBot [16], MTR [42], and Wayformer [31]), and evaluations. UniTraj not only serves as a solution to tackle our research questions but also provides a comprehensive and flexible platform for the community. First, it is designed for the effortless inclusion of new datasets by proposing a unified data structure compatible with various datasets. Second, Unitraj supports and simplifies the integration of new methods by providing numerous essential data processing and loss functions relevant to the trajectory prediction task. Lastly, UniTraj offers unified evaluation metrics, as well as diverse and insightful evaluation approaches, such as analyzing performance on the long-tail data instances as well as different clusters of data samples to allow a more in-depth understanding of model behavior. Figure 1 shows the overview of the framework.

We conduct extensive experiments using the UniTraj framework to shed light on our two research questions. Our findings reveal a large performance drop when transitioning between data sources, alongside variations in the generalization abilities induced by different datasets (RQ1). We also show that scaling up dataset size and diversity can enhance model performance significantly without any architectural modifications, leading us to rank 1st in the nuScenes public leaderboard. This is accomplished by training models on all existing datasets in the framework. This unified dataset forms the largest public data one can use to train a vehicle trajectory prediction model, with more than 2M samples, 1337 hours of data, and 15 different cities. Finally, by providing an in-depth analysis of the datasets, we offer a more comprehensive understanding of their characteristics. Our analysis reveals that the datasets' generalization capabilities are not only attributed to their size, but also their intrinsic diversity. We believe that the framework opens up new opportunities in the trajectory prediction field, and we will release the framework to foster further advancements. In summary, our contributions are as follows:

- We introduce UniTraj, a comprehensive open-source framework for vehicle trajectory prediction, integrating various datasets, models, and evaluations. It offers a unified platform for comprehensive research in this field.
- We investigate models' generalization across different datasets and cities and provide insight into the characteristics of datasets on which models acquire better generalization capacities.
- We explore the data scaling impact on model performance employing the largest collection of datasets currently available, and establish a new state-of-the-art model on the nuScenes dataset.
- Finally, we provide an in-depth comparative analysis of the datasets, shedding light on our experimental findings.

2 Previous work

Trajectory prediction datasets. Many academic and industrial laboratories have paved the way for research development by open-sourcing real-world driving datasets [6–8, 10, 14, 17, 29, 33, 46, 50]. Notably, Argoverse [10] was among the pioneers in releasing the lane graph information, nuScenes [7] expanded the variety of scenes, Waymo [14] enriched their dataset with fine-grained information, and recently, Argoverse 2 [46] released the largest data in terms of unique roadways. While these datasets contribute to field developments, they have been developed in isolation without considering harmonization with former datasets. Thus, there exist multiple challenges in combining them due to various incompatibilities. This work addresses the challenges through a unified framework.

Trajectory prediction benchmarks. Multi-dataset benchmarks have already been explored in various domains such as object detection [52], semantic segmentation [22, 41], and pose prediction [36]. In the field of trajectory prediction, such benchmarks have primarily been developed for human trajectory prediction [1, 34, 37]. Notably, Trajnet++ [23] provides an interaction-centric benchmark by categorizing trajectories based on the presence of an interaction. trajdata [19] is a unified interface to multiple human trajectory datasets incorporating scene context into the inputs. A related work for the task of vehicle trajectory planning is ScenarioNet [25], a simulator aggregating multiple real-world datasets into a unified format and providing a planning development and evaluation framework. To the best of our knowledge, we are the first to propose an open-source framework for vehicle trajectory prediction. Our framework is not limited to including multiple datasets; it also integrates a variety of trajectory prediction models and evaluation methodologies, thereby providing a comprehensive resource for advancing research and development in the vehicle trajectory prediction task.

Generalization of trajectory prediction models. The discrepancies in data formats in vehicle trajectory prediction datasets hinder research on cross-dataset generalization, leading to limited studies in this area. In [49], one dataset is divided into different domains to explore model generalization. Authors in [40] propose an epistemic uncertainty estimation approach and perform cross-dataset evaluation. In [15], the authors studied cross-dataset generalization of models and showed a performance gap between datasets. However, they provide limited insights into the sources of the generalization gap. Moreover, their code is not publicly available. Previous works also investigated some generalization aspects of trajectory prediction models when they deal with new scenes and cities [11, 24, 27], new agent types [24], using perception outputs instead of curated annotations [45, 47, 51], and facing adversarial situations [3, 35, 38]. In this work, we conduct more extensive and in-depth cross-dataset, and cross-city analyses as well as multi-dataset training. Moreover, we provide insights into the dataset characteristics, explaining the findings. We also release an open-source framework to facilitate this line of research.

Table 1: Summary of the discrepancies in data features. The table shows the features for each dataset and the unified version of the features. Most of the unified features are flexible and can be chosen by the user.

		Argoverse2	WOMD	nuScenes	**UniTraj**
Coordinate frame		Scene-centric	Scene-centric	Scene-centric	Agent-centric
Time length	Past	5 sec	1 sec	2 sec	[0 - 8] sec
	Future	6 sec	8 sec	6 sec	[1 - 8] sec
Agent features	Annotations	velocity, heading	velocity, heading	velocity, heading	velocity, heading
	Other info	—	bounding box size	—	acceleration
	Coordinates	2D	3D	2D	2D
Map features	Range	∼200m	∼200m	∼500m	[0 - 500] m
	Resolution	0.2m∼2m	∼0.5m	∼1m	[0.2 - 2] m
	Coordinates	2D	3D	2D	2D

3 UniTraj framework

The UniTraj framework, illustrated in Figure 1, consists of three main components. The first component unifies the format and features of various datasets (see Section 3.1). The second component adapts trajectory prediction models to the unified data format, facilitating their training (see Section 3.2). The final component consists of a comprehensive and shared evaluation process for the models (see Section 3.3). The integration of the components allows for diverse experimentation, such as cross-dataset training, evaluation, and dataset analysis.

3.1 Unified data

Two types of discrepancies are found across trajectory forecasting datasets: *data formats* and *data features*. The former amounts to differences in the way data is structured and organized, while the latter stems from differences in the characteristics of the data itself, such as spatio-temporal resolution, range, and agent and map annotation taxonomy. In this section, we present solutions to tackle both types of discrepancies.

Unified data format: To address the issue of different data formats used in trajectory prediction datasets, such as TFRecord in WOMD [14] and Apache Parquet in Argoverse 2 [46], we utilize ScenarioNet [25]. ScenarioNet was initially designed for traffic scenario simulation and modeling, but we repurposed it for the vehicle trajectory prediction task. It provides a unified scenario description format containing HD maps and detailed object annotations, which simplifies the process of decoding the dataset with different formats. ScenarioNet currently supports converting WOMD, nuScenes, and nuPlan, and we extend its support to Argoverse 2 for our research. This reduces the need for multiple versions of preprocessing code to extract information from raw datasets and create batched data for the training of prediction models.

Unified data features: Despite the data being converted into a unified format, significant discrepancies persist across the datasets, affecting model performance. For example, the scenarios are 11 seconds long in Argoverse 2, while they are 9 seconds in WOMD; or the precision of map annotations are 1 meters in nuScenes while they are 0.5 meters in WOMD. Therefore, we aim to harmonize these discrepancies and minimize their impact on the model's performance. Table 1 summarizes the discrepancies and the unified features. Our data processing approach involves specific harmonizations, including the following:

- **Coordinate Frame.** Recent trajectory prediction models predominantly utilize vectorized, agent-centric data as input [16, 31, 42, 43, 53]. Our data processing pipeline is designed to transform scene-level raw data into this format. It processes traffic scenarios, which consist of multiple trajectories, and selects trajectories designated as training samples within the datasets. The pipeline then converts the entire scenario into the coordinate frames of these selected agents, and normalizes the input accordingly.
- **Time Length.** The trajectories in different datasets are with the same frequency of 10Hz but with a duration ranging from 8 to 20 seconds. To standardize this aspect, we truncate all trajectories to a uniform length of 8 seconds. Within this duration, UniTraj provides the option to flexibly determine a unified length of past and future trajectories for all datasets.
- **Agent Features.** Among the datasets, WOMD provides the most detailed agent information, including 3D coordinates, velocity, heading, and bounding box size, whereas nuScenes lacks certain data, like bounding box size. We standardize inputs across datasets by using 2D coordinates, velocity, and heading. Our data processing also introduces supplementary features, such as one-hot encoding of agent type and time steps of trajectories, and acceleration. These elements are combined to create a rich, unified input for the model.
- **Map Features.** Datasets differ in HD map resolution. We normalize this by using linear interpolation to standardize the distance between consecutive points to 0.5 meters, with an option for further downsampling to adjust map resolution. Additionally, we enrich the data with each lane point's direction and one-hot encode lane types. Our experiments utilize semantic map classes such as center lanes, road lines, crosswalks, speed bumps, and stop signs.

Our framework allows for customization of specific features through predefined parameters for focused single-dataset research, while still providing a standardized data format across all datasets. The data processing module supports various parameters, such as the length of historical and future trajectories, number of points per lane, map resolution, types of surrounding agents and lines, and masked attributes. Thanks to its modular structure, our data processing pipeline enables easy integration of new processing methodologies and models. The framework is equipped with multi-processing and caching mechanisms for efficient processing. Our framework currently includes four large-scale, real-world datasets with over 1337 hours of driving data from 15 cities.

3.2 Unified models

Trajectory prediction models are often implemented in different pipelines, making direct comparisons challenging and fairness hard to ensure. We integrate three recent trajectory prediction models within the UniTraj framework. These models were chosen based on their state-of-the-art results on various benchmarks, indicating the research value of their designs. We include:

- **AutoBot** [16] is a recent transformer-based model with competitive results on multiple existing datasets. It is based on equivariant feature learning to learn the joint distribution of trajectories with multi-head attention blocks.
- **MTR** [42] ranked first on the WOMD Challenge in 2022. It operates by integrating global intention priors with local movement refinement. It uses a limited number of adaptable motion query pairs, allowing precise trajectory prediction and improvement for different motion types.
- **Wayformer** [31] is a transformer-based model, featuring a multi-axis encoder. It employs latent queries that facilitate the combination of multi-dimensional inputs. We re-implement the model, as the original code has not been released.

The models' capacities cover a large range (1.5M parameters for AutoBot, 60.1M for MTR, and 16.5M for Wayformer), enabling research on model size and scaling.

Integrating new models: The flexibility of UniTraj's data processing pipeline greatly simplifies the integration of new models. Furthermore, we provide a standardized output format, enabling seamless use of UniTraj's evaluation and logging tools.

3.3 Unified evaluation

In trajectory prediction, various metrics have been proposed to evaluate the models, yet there is no consensus about them. As a result, each dataset provides a different set of evaluation metrics, making it challenging to compare performances across datasets. For example, WOMD employs mean average precision (mAP) metric [14] while Argoverse 2 uses brier minimum Final Displacement Error (brier-minFDE) [46]. Our framework provides a unified set of metrics to allow comprehensive and consistent evaluation across different datasets. To this end, we employ two sets of metrics: general and fine-grained evaluation metrics.

General evaluations: The most common metrics in the literature are the ones that provide an overall score based on accuracy measures by comparing the output with the ground truth in different aspects. We include the following three general metrics in the framework:
1) Minimum Average / Final Displacement Error (minADE/minFDE): It represents the minimum average/final displacement error between the predictions and the ground truth. The minimum is computed over the 6 modes of the output.

2) Miss Rate (MR): It is defined as the ratio of the samples with minFDE exceeding 2 meters, and is useful where up to 2 meters deviation is acceptable.
3) Brier Minimum Final Displacement Error (brier-minFDE): While the previous metrics focus on covering the ground truth, they do not account for the probability assigned to each predicted trajectory. The brier-minFDE metric addresses this by adding a penalty term, $(1-p)^2$, to the minFDE where p corresponds to the probability of the trajectory that best matches the ground truth.

Fine-grained evaluations: We also provide two fine-grained evaluations.

(1) Trajectory types. Datasets usually exhibit a significant prevalence of 'straight' trajectories, resulting in heavily imbalanced datasets. Besides, we argue that rare trajectory types can sometimes be the more safety-critical ones. Therefore, it is critical to specifically access prediction performances on rare situations and trajectory types. To address this, the UniTraj framework enables the stratification of evaluation metrics based on trajectory types. In practice, we adopt the trajectory taxonomy defined in the WOMD challenge [14], to categorize trajectories into the following groups: 'stationary', 'straight', 'straight left', 'straight right', 'left-turn', 'right-turn', 'left u-turn', 'right u-turn'. While the use of this taxonomy provides valuable insights, its scope has limitations as it does not account for variations in motion dynamics. For instance, it does not differentiate between straight accelerating and decelerating trajectories, both of which are categorized as 'straight'. Consequently, we additionally use the notion of 'Kalman difficulty' introduced below.

(2) Kalman difficulty. Some situations are more challenging to forecast than others, typically when the future is not a simple extrapolation of the past and when contextual factors play a significant role. The context encloses various elements such as map data, social interactions, or input signals coming from perception. Moreover, previous works [4, 28, 44] observe that these complex scenarios, while critical, are much less frequent than scenarios that are easier to forecast. To specifically evaluate the performance over critical cases, and reduce evaluation noise coming from the large number of simple scenarios, the authors in [28] propose to filter them as the ones with a high mismatch between their ground truth and predictions from a Kalman filter [21]. We follow this idea as it offers a simple method to evaluate how challenging a situation is. Accordingly, UniTraj stratifies evaluation metrics based on *Kalman difficulty* that we define as the FDE between the ground-truth trajectory and the prediction of a linear Kalman filter.

4 Experiments

The UniTraj framework opens up new opportunities for research and experimentation. This section presents experiments highlighting these opportunities, focusing on cross-domain (i.e., cross-dataset and cross-city) generalization (RQ1) in Section 4.1, and data scaling impact (RQ2) for trajectory prediction models in Section 4.2. We provide fine-grained dataset analyses and discussions in Sec-

tion 4.3. Additional experiments in the appendix, such as continual learning and synthetic-to-real transfer, further demonstrate the framework's research utility.
Experimental settings: We replicate the model configurations and hyper-parameters from their original implementations. Throughout the experiments, we have limited the training and validation samples to vehicle trajectories. The map range extends to a 100m radius with a spatial resolution of 0.5m. The temporal parameters are set to 2 seconds of historical trajectories and 6 second future trajectories. For our multi-dataset training experiments, we utilize WOMD [14], Argoverse 2 [46], and nuScenes [7] datasets. Since the nuPlan [8] dataset is oriented towards planning tasks and lacks an official training/validation set for prediction tasks, we exclusively use it for the cross-city generalization studies due to its large number of samples for different cities. We only report the results with the brier-minFDE metric and leave other metrics for the appendix.

4.1 Generalization evaluation

Generalization to new domains is a crucial challenge for data-driven models, necessitating diverse data for comprehensive evaluation. The UniTraj framework enables the exploration of model generalization across various datasets and cities.

Cross-dataset evaluation: To assess the generalization capabilities of models, we train models on each individual dataset and evaluate their performance on all other available datasets. The findings are presented in Table 2. Analyzing the data in different columns of the table, the first observation is that all models' performances decline significantly when models are tested on other datasets. This is a consistent trend across all of the three model architectures, and all of the considered datasets. For instance, the second column under MTR reports the performance evaluated on the validation set of Argoverse 2. It indicates that MTR achieves its peak performance when it is trained on the training set of Argoverse 2 itself, while models trained on nuScenes and WOMD exhibit significantly lower performances.

With a more detailed investigation, we can also compare the generalization capabilities of different datasets. For instance, considering the same column, the model trained on WOMD outperforms the one trained on nuScenes when evaluated on the Argoverse2 dataset. By making similar comparisons across other columns, we establish a generalization order: models trained on WOMD data exhibit the highest generalization ability, followed by those trained on Argoverse2, and then nuScenes. This order remains consistent across all models. The superior generalization of models trained on WOMD can be attributed to both the larger number of data samples and the greater variety present in the WOMD dataset. We provide a more detailed explanation in Section 4.3, where we discuss the specific characteristics of each dataset and their influence on model performance and generalization.

Cross-city evaluation: Despite our care to standardize data formats and align features among datasets, certain fundamental discrepancies may persist caused

Table 2: Cross-dataset generalization and multi-dataset training experiments. Training and validation are across multiple datasets. Rows indicate the training data of the model, columns indicate the evaluation data. 'All' designates the combination of the three considered datasets. The study is conducted for three models (AutoBot [16], MTR [42], and Wayformer [31]). '∗' indicates our internal implementation of the model. We report the brier-minFDE (↓) metric.

		MTR [42]			Wayformer * [31]			AutoBot [16]		
		← **Evaluation** →								
↓ **Training**	#trajs	nuScenes	Argoverse 2	WOMD	nuScenes	Argoverse 2	WOMD	nuScenes	Argoverse 2	WOMD
nuScenes	32k	**2.86**	4.50	7.38	**3.06**	4.68	7.16	**3.36**	4.48	6.89
Argoverse 2	180k	3.72	**2.08**	4.68	3.69	**2.38**	4.80	4.35	**2.51**	4.43
WOMD	1800k	3.10	3.63	**2.13**	3.12	3.60	**2.10**	3.73	3.23	**2.47**
All	2012k	**2.27**	**1.99**	**2.13**	**2.32**	**2.12**	**2.09**	**3.07**	2.54	**2.47**

by the data collection and annotation processes. For instance, annotation noises could still exist across datasets. To control for this potential residual discrepancy, we explore the generalization of AutoBot when the city is changed inside a single dataset. Similar to the previous experiment, we train AutoBot on each city and evaluate it on the rest of cities. We employed nuPlan [8] data for this experiment due to the large number of samples existing in diverse cities and selected 10K samples from each city. The results are shown in Table 3. It shows that the performance of AutoBot drops once evaluated on other cities. For instance, the first row shows that the model trained on Pittsburgh has the best performance on Pittsburgh (brier-minFDE 2.4) and worse performances on Boston (2.7) and Singapore (3.5). This indicates a clear generalization gap between cities, emphasizing the discrepancies between different environments. Moreover, it can also be observed that the model trained on Singapore performs the worst on average. This is an expected outcome, given that Singapore is a left-hand traffic city, unlike the other ones.

Takeaways: The findings in this section reveal that *state-of-the-art models trained on recent large-scale datasets struggle to generalize to new domains*. As a concrete recommendation, it highlights the importance of geographical diversity in the data collection process for both training and evaluation.

	← **Evaluation** →			
↓ **Training**	Pittsburgh	Boston	Singapore	*Average*
Pittsburgh	**2.4**	2.7	3.5	**2.8**
Boston	4.1	**2.2**	3.4	3.2
Singapore	4.9	3.5	**2.1**	3.5

Table 3: Cross-city generalization experiment. We train and validate AutoBot across multiple cities in the nuPlan dataset and report the brier-minFDE (↓) metric.

4.2 Scaling data to 2M trajectories.

The unified data available in UniTraj forms the largest public data one can use to train a trajectory prediction model. In this section, we explore if we can improve the models' performance by simply scaling the size of the training dataset. Therefore, we combine all the existing real datasets in UniTraj into a single large training set on which we train the considered models. The results of this experiment are presented at the bottom of Table 2 (row 'All'), demonstrating improvements over the model trained solely on a single dataset. While the improvements are not identical for different datasets, they are particularly significant in the case of the nuScenes dataset, making the MTR model trained on combined data rank 1^{st} in the nuScenes leaderboard (shown in Table 4). For instance, training the MTR model on "All" datasets enables it to outperform the model trained on nuScenes and Argoverse2 by a large margin. Moreover, while the performance on WOMD has not been improved, the resulting model performs much better than the model trained on WOMD on other datasets. These improvements are attributed to the relatively larger size and diversity of the combined dataset compared to each individual one. We elaborate on this more in Section 4.3.

The table also shows that certain models benefit more from larger data sizes compared to others. Specifically, when looking at the performance on Argoverse 2 and nuScenes datasets, which benefit from the increased dataset size, models like MTR and Wayformer show more significant improvements than AutoBot. This difference in performance enhancement is attributed to the models' capacity. For instance, MTR has 60.1 M parameters, providing it with a higher capacity to learn from larger datasets, whereas AutoBot, with only 1.5 M parameters, may not be as able to utilize the additional data.

In order to illustrate the impact of data size on the performance of a trajectory predcition model, we gradually increase the number of training samples from 20% to 100% of the combined dataset. We then report the AutoBot model's performance using the average brier-minFDE metric among all three datasets. Figure 2 shows the curve revealing a consistent reduction in the prediction error once the dataset size increases. This highlights the substantial benefits of larger datasets on the model's performance and offers prospects for improved performances with larger data sizes.

Takeaways: The experimental results underscore the potential and need for larger, more diverse datasets in the trajectory prediction field. Such datasets will also push the boundaries of the current performances of the models.

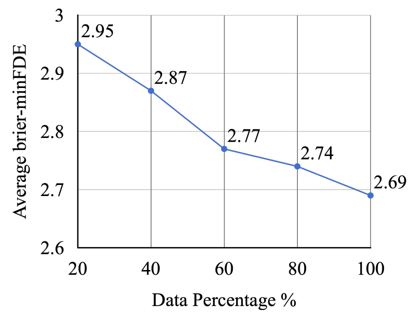
4.3 Analyzing the results

In this section, we aim to explain the findings about generalization gaps and the data scaling impact in Sections 4.1 and 4.2. Thus, we first delve into a comparative analysis of datasets integrated within the UniTraj framework with our fine-grained evaluations. We then employ these insights to explain the findings.

Table 4: nuScenes Leaderboard. We train AutoBot and MTR with all datasets, and evaluate on nuScenes (ranking at the time of submission among public methods)

Method	Ranking (↓)	minADE5 (↓)
MTR-UniTraj	1	0.96
Goal-LBP [48]	2	1.02
CASPNet++ [39]	3	1.16
Socialea [12]	4	1.18
Autobot-Unitraj	11	1.26
Autobot	19	1.37

Fig. 2: Relationship between dataset size and model performance. The prediction error of AutoBot as the combined dataset size increases, varying from 20% to 100% of the total data.

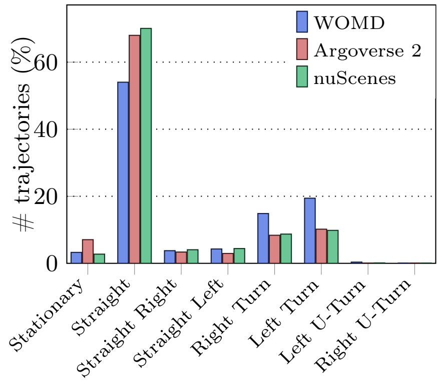


Dataset analysis. We provide an in-depth comparison between datasets to help the understanding of the results presented in previous subsections. Moreover, the insights aid in making informed decisions about selecting the most appropriate datasets and settings for specific research or application needs.

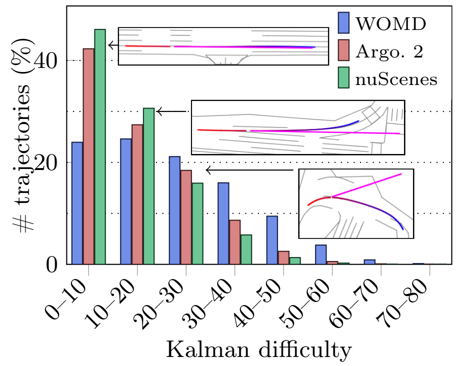
Trajectory type based comparison: The analysis of trajectory types in the WOMD, Argoverse 2, and nuScenes datasets in Figure 3a reveals trajectory type imbalances, primarily featuring a prevalence of straight trajectories, constituting 54% to 68% of all trajectories, with minimal instances of u-turns. WOMD exhibits a notably diverse trajectory mix, with a significant number of left and right turn trajectories, approximately two times more than what is observed in Argoverse 2 and nuScenes. This stands in contrast to Argoverse 2 and nuScenes, which primarily contain straight trajectories over varied turning maneuvers.

Kalman difficulty based comparison: The distribution of sample difficulties within the WOMD, Argoverse 2, and nuScenes datasets, shown in Figure 3b, exhibits a consistent trend where easier scenarios significantly outnumber more challenging ones. WOMD demonstrates a relatively balanced distribution across lower to moderate difficulty levels, with around ∼24% of trajectories falling within the easiest category (Kalman difficulties up to 10), with a consistent presence observed up to a Kalman difficulty of 50.0. In contrast, both nuScenes and Argoverse 2 exhibit a substantial bias towards easier difficulties, comprising approximately ∼42% and ∼46% of samples, respectively, in the lowest difficulty range (Kalman difficulties lower than 10), and show a sharp decrease in proportion with increasing difficulty levels, indicating datasets primarily composed of simpler scenarios which may potentially limit their efficacy in training models for more complex situations.

Explaining the findings in cross-dataset generalization and multi-dataset training experiments: Our cross-dataset generalization experiment in Section 4.1 shows that models do not generalize equally across different datasets.



(a) Distribution of trajectory types across datasets.



(b) Histogram of the Kalman Difficulty of trajectories.

Fig. 3: Figure (a) shows the distribution of trajectory types. It reveals an imbalance across different types with straight being the most common trajectory type in the datasets. Figure (b) shows the histogram of the Kalman Difficulty of trajectories. To give a sense of the Kalman difficulty, we overlay three random examples. The past trajectory, the ground truth, and the Kalman filter prediction are shown in red, blue, and magenta, respectively. The plot displays a clear trend with a notably higher count of simpler scenarios compared to challenging ones. WOMD, in particular, shows a relatively balanced distribution across scenarios.

Table 5: Stratified evaluations per trajectory type. We report brier-minFDE on nuScenes validation data. We compare the performance of two MTR [42] models trained on nuScenes data (nuScenes) and the combined dataset in UniTraj (All).

Traj. Type	Stationary	Straight	Straight right	Straight left	Right u-turn	Right-turn	Left u-turn	Left-turn	All
MTR (nuScenes)	2.15	2.58	4.85	4.26	8.13	4.82	5.17	4.85	2.86
MTR (All)	2.23	2.31	3.13	3.06	2.98	3.53	2.10	2.82	2.27

Notably, models trained on WOMD generalize better to other datasets. Furthermore, models trained on combined datasets exhibit considerable improvements. To understand these phenomena, it's important to delve into the differences between datasets, focusing on two main aspects: size and diversity.

To investigate the impact of dataset size, we replicate the cross-dataset generalization experiments (Table 2), but with control on the dataset size, as we select 30k random samples for each dataset's training set. Table 6 shows the results. The last column shows the generalization hierarchy where again the WOMD generalizes best, followed by Argoverse 2 and then nuScenes. This shows that better cross-dataset generalization is not solely attributed to the size of the datasets. The last row illustrates that for multi-dataset training, there is a considerable improvement for all the datasets, highlighting the pronounced benefit of adding more data in small-scale dataset scenarios.

Table 6: Cross-dataset generalization experiments with identical sample size. We select 30K random samples from every dataset, then train and validate AutoBot across them and report the brier-minFDE metric.

	← Evaluation →			
↓ **Training**	nuScenes	Argoverse 2	WOMD	Average
nuScenes	**3.38**	4.48	6.88	4.91
Argoverse 2	4.67	**2.90**	5.07	4.21
WOMD	4.42	4.04	**3.22**	**3.89**
All	**3.25**	**2.80**	**3.13**	**3.06**

Table 7: Fine-grained evaluation Kalman difficulty. We report the brier-minFDE (↓) metric across three chunks of Kalman difficulties on nuScenes validation data. We compare the performance of two MTR [42] models trained on nuScenes data (nuScenes) and the combined dataset in UniTraj (All).

Kalman difficulty	Easy $\in [0, 30[$	Medium $\in [30, 50[$	Hard $\in [50, 100[$
MTR (nuScenes)	2.73	4.52	4.25
MTR (All)	2.23	2.97	4.20

Section 4.3 reveals that the datasets are dissimilar in terms of diversity. Notably, WOMD encompasses the most diverse range of scenarios in comparison to other datasets. This explains the superior generalization of WOMD to other datasets, as the diversity enables models to learn the full spectrum of data distributions more comprehensively. Similarly, the combined dataset provides a more diverse collection of trajectories, leading to enhanced performance for the models. To demonstrate this, we compare the fine-grained evaluations of MTR model trained on nuScenes and the combined dataset. Table 5 shows the per trajectory type performance of the two models where the model trained on full data outperforms in every trajectory type since the full data includes significantly more samples from each trajectory type. We also compare the performances using the Kalman difficulty measure in Table 7. The combined data has considerably more medium-difficulty samples (shown in Figure 3b) leading to significant performance improvements in the medium-range samples. These results highlight the importance of diversity in the data.

5 Conclusions

In conclusion, our study examines two critical research questions essential for advancing the field of vehicle trajectory prediction. We have uncovered that models face significant challenges in generalizing across different domains (RQ1), exhibiting considerable performance drops when encountering new datasets or cities. Additionally, our findings affirm that larger, more diverse datasets significantly boost model performance and generalization capabilities (RQ2), underscoring the importance of data richness. Besides, we release the UniTraj framework as a versatile tool that opens up new opportunities for exploration in trajectory predictions. We believe that this framework will help significantly in advancing research in the field of trajectory prediction.

Acknowledgments: Authors would like to thank Mickaël Chen, Quanyi Li, Ahmad Rahimi, Yihong Xu, Olaf Dunkel and Ahmad Salimi for their valuable discussions and help on preliminary versions of this work. The project was partially funded by Honda R&D Co., Ltd, Valeo Paris, and ANR grant MultiTrans (ANR-21-CE23-0032).

References

1. Javad Amirian, Bingqing Zhang, Francisco Valente Castro, Juan Jose Baldelomar, Jean-Bernard Hayet, and Julien Pettré. Opentraj: Assessing prediction complexity in human trajectories datasets. In *Proceedings of the asian conference on computer vision*, 2020. 4
2. Mohammadhossein Bahari, Ismail Nejjar, and Alexandre Alahi. Injecting knowledge in data-driven vehicle trajectory predictors. *Transportation research part C: emerging technologies*, 128:103010, 2021. 3
3. Mohammadhossein Bahari, Saeed Saadatnejad, Ahmad Rahimi, Mohammad Shaverdikondori, Amir Hossein Shahidzadeh, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Vehicle trajectory prediction works, but not everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17123–17133, 2022. 3, 4
4. Hedi Ben-Younes, Éloi Zablocki, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Raising context awareness in motion forecasting. In *CVPRW*, pages 4408–4417. IEEE, 2022. 8
5. Prarthana Bhattacharyya, Chengjie Huang, and Krzysztof Czarnecki. Ssl-lanes: Self-supervised learning for motion forecasting in autonomous driving. In *Conference on Robot Learning*, pages 1793–1805. PMLR, 2023. 3
6. Julian Bock, Robert Krajewski, Tobias Moers, Steffen Runde, Lennart Vater, and Lutz Eckstein. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1929–1934. IEEE, 2020. 4
7. Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3, 4, 9
8. Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 2, 4, 9, 10
9. Yulong Cao, Chaowei Xiao, Anima Anandkumar, Danfei Xu, and Marco Pavone. Advdo: Realistic adversarial attacks for trajectory prediction. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022. 3
10. Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4
11. Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *ICCV*, pages 9804–9813. IEEE, 2021. 4
12. Jiuyu Chen, Zhongli Wang, Jian Wang, and Baigen Cai. Q-eanet: Implicit social modeling for trajectory prediction via experience-anchored queries. *IET Intelligent Transport Systems*, 2023. 12

13. Pasquale Coscia, Francesco Castaldo, Francesco AN Palmieri, Alexandre Alahi, Silvio Savarese, and Lamberto Ballan. Long-term path prediction in urban scenarios using circular distributions. *Journal on Image and Vision Computing*, 69:81–91, 2018. 3
14. Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin 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*, pages 9710–9719, 2021. 2, 3, 4, 5, 7, 8, 9
15. Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Uncertainty estimation for cross-dataset performance in trajectory prediction. *IEEE Int. Conf. on Robotics and Automation Workshop on Fresh Perspectives on the Future of Autonomous Driving*, 2022. 4
16. Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2022. 3, 6, 7, 10
17. John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021. 4
18. Kai-Chieh Hsu, Karen Leung, Yuxiao Chen, Jaime Fernàndez Fisac, and Marco Pavone. Interpretable trajectory prediction for autonomous vehicles viacounterfactual responsibility. In *IEEE/RSJ Int. Conf. on Intelligent Robots & Systems*, 2023. 3
19. Boris Ivanovic, Guanyu Song, Igor Gilitschenski, and Marco Pavone. trajdata: A unified interface to multiple human trajectory datasets. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, New Orleans, USA, Dec. 2023. 4
20. Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9644–9653, 2023. 3
21. Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 8
22. Dongwan Kim, Yi-Hsuan Tsai, Yumin Suh, Masoud Faraki, Sparsh Garg, Manmohan Chandraker, and Bohyung Han. Learning semantic segmentation from multiple datasets with label shifts. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 4
23. Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 4
24. Parth Kothari, Danya Li, Yuejiang Liu, and Alexandre Alahi. Motion style transfer: Modular low-rank adaptation for deep motion forecasting. In *CoRL*, volume 205 of *Proceedings of Machine Learning Research*, pages 774–784. PMLR, 2022. 4
25. Quanyi Li, Zhenghao Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in Neural Information Processing Systems*, 2023. 4, 5
26. Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Com-*

puter Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 541–556. Springer, 2020. 3
27. Yuejiang Liu, Riccardo Cadei, Jonas Schweizer, Sherwin Bahmani, and Alexandre Alahi. Towards robust and adaptive motion forecasting: A causal representation perspective. In *CVPR*, pages 17060–17071. IEEE, 2022. 4
28. Osama Makansi, Özgün Çiçek, Yassine Marrakchi, and Thomas Brox. On exposing the challenging long tail in future prediction of traffic actors. In *ICCV*, pages 13127–13137. IEEE, 2021. 8
29. Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021. 4
30. Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. Attention based vehicle trajectory prediction. *IEEE Transactions on Intelligent Vehicles*, 6(1):175–185, 2021. 3
31. Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2980–2987. IEEE, 2023. 3, 6, 7, 10
32. Mozhgan Pourkeshavarz, Changhe Chen, and Amir Rasouli. Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8384–8393, 2023. 3
33. Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory prediction in crowded scenes. In *European Conference on Computer Vision (ECCV)*, volume 2, page 5, 2016. 4
34. Andrey Rudenko, Luigi Palmieri, Wanting Huang, Achim J Lilienthal, and Kai O Arras. The atlas benchmark: an automated evaluation framework for human motion prediction. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 636–643. IEEE, 2022. 4
35. Saeed Saadatnejad, Mohammadhossein Bahari, Pedram Khorsandi, Mohammad Saneian, Seyed-Mohsen Moosavi-Dezfooli, and Alexandre Alahi. Are socially-aware trajectory prediction models really socially-aware? *arXiv preprint arXiv:2108.10879*, 2021. 4
36. Saeed Saadatnejad, Mehrshad Mirmohammadi, Matin Daghyani, Parham Saremi, Yashar Zoroofchi Benisi, Amirhossein Alimohammadi, Zahra Tehraninasab, Taylor Mordan, and Alexandre Alahi. Toward reliable human pose forecasting with uncertainty. *IEEE Robotics and Automation Letters*, 9(5):4447–4454, 2024. 4
37. Amir Sadeghian, Vineet Kosaraju, Agrim Gupta, Silvio Savarese, and Alexandre Alahi. Trajnet: Towards a benchmark for human trajectory prediction. *arXiv preprint*, 2018. 4
38. Jay Sarva, Jingkang Wang, James Tu, Yuwen Xiong, Sivabalan Manivasagam, and Raquel Urtasun. Adv3d: Generating safety-critical 3d objects through closed-loop simulation. *CoRR*, abs/2311.01446, 2023. 4
39. Maximilian Schäfer, Kun Zhao, and Anton Kummert. Caspnet++: Joint multi-agent motion prediction, 2023. 12
40. Wenbo Shao, Yanchao Xu, Jun Li, Chen Lv, Weida Wang, and Hong Wang. How does traffic environment quantitatively affect the autonomous driving prediction? *IEEE Transactions on Intelligent Transportation Systems*, 2023. 4
41. Bowen Shi, Xiaopeng Zhang, Haohang Xu, Wenrui Dai, Junni Zou, Hongkai Xiong, and Qi Tian. Multi-dataset pretraining: A unified model for semantic segmentation. *arXiv preprint arXiv:2106.04121*, 2021. 4

42. Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022. 3, 6, 7, 10, 13, 14
43. Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022. 6
44. Yuning Wang, Pu Zhang, Lei Bai, and Jianru Xue. FEND: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In *CVPR*, pages 1400–1409. IEEE, 2023. 8
45. Xinshuo Weng, Boris Ivanovic, Kris Kitani, and Marco Pavone. Whose track is it anyway? improving robustness to tracking errors with affinity-based trajectory prediction. In *CVPR*, pages 6563–6572. IEEE, 2022. 4
46. Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 2, 3, 4, 5, 7, 9
47. Yihong Xu, Loïck Chambon, Éloi Zablocki, Mickaël Chen, Alexandre Alahi, Matthieu Cord, and Patrick Pérez. Towards motion forecasting with real-world perception inputs: Are end-to-end approaches competitive? In *ICRA*, 2024. 4
48. Zhen Yao, Xin Li, Bo Lang, and Mooi Choo Chuah. Goal-lbp: Goal-based local behavior guided trajectory prediction for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–10, 2023. 12
49. Luyao Ye, Zikang Zhou, and Jianping Wang. Improving the generalizability of trajectory prediction models with frenet-based domain normalization. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 4
50. Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019. 4
51. Pu Zhang, Lei Bai, Jianru Xue, Jianwu Fang, Nanning Zheng, and Wanli Ouyang. Trajectory forecasting from detection with uncertainty-aware motion encoding. *CoRR*, abs/2202.01478, 2022. 4
52. Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7571–7580, 2022. 4
53. Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023. 6