Large Scale Interactive Motion Forecasting for Autonomous Driving : The WAYMO OPEN MOTION DATASET

Scott Ettinger [1], Shuyang Cheng [1], Benjamin Caine [2], Chenxi Liu [1], Hang Zhao [1], Sabeek Pradhan [1], Yuning Chai [1], Ben Sapp [1], Charles Qi [1], Yin Zhou [1], Zoey Yang [1], Aurélien Chouard [1], Pei Sun [1], Jiquan Ngiam [2], Vijay Vasudevan [2], Alexander McCauley [1], Jonathon Shlens [2], Dragomir Anguelov [1]
[1] Waymo LLC, [2] Google Brain

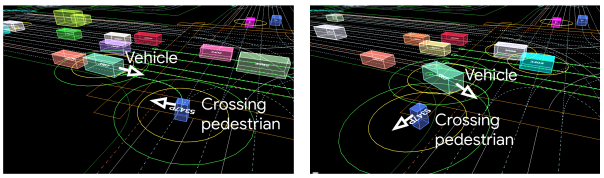
Abstract

As autonomous driving systems mature, motion forecasting has received increasing attention as a critical requirement for planning. Of particular importance are interactive situations such as merges, unprotected turns, etc., where predicting individual object motion is not sufficient. Joint predictions of multiple objects are required for effective route planning. There has been a critical need for high-quality motion data that is rich in both interactions and annotation to develop motion planning models. In this work, we introduce the most diverse interactive motion dataset to our knowledge, and provide specific labels for interacting objects suitable for developing joint prediction models. With over 100,000 scenes, each 20 seconds long at 10 Hz, our new dataset contains more than 570 hours of unique data over 1750 km of roadways. It was collected by mining for interesting interactions between vehicles, pedestrians, and cyclists across six cities within the United States. We use a high-accuracy 3D auto-labeling system to generate high quality 3D bounding boxes for each road agent, and provide corresponding high definition 3D maps for each scene. Furthermore, we introduce a new set of metrics that provides a comprehensive evaluation of both single agent and joint agent interaction motion forecasting models. Finally, we provide strong baseline models for individual-agent prediction and joint-prediction. We hope that this new large-scale interactive motion dataset will provide new opportunities for advancing motion forecasting models.

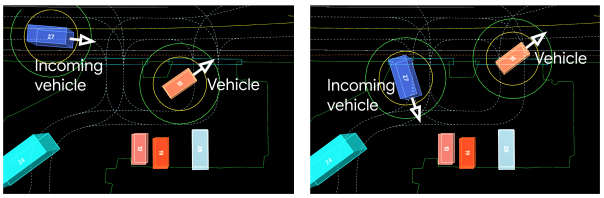
1. Introduction

Motion forecasting has received increasing attention as a critical requirement for planning in autonomous driving systems [8, 14, 40, 36, 28, 34]. Due to the complexity of scenes that autonomous systems need to safely handle, predicting object motion in the scene is a difficult task, suitable for machine learning models. Building effective motion forecasting models requires large amounts of high quality real world data. Creating a dataset for motion forecasting is complicated by the fact that the distribution of real world data is highly imbalanced [4, 18, 32, 38]; in the common case, vehicles drive straight at a constant velocity. In order to develop effective models, a dataset must contain and measure performance on a wide range of behaviors and trajectory shapes for different object types that an autonomous system will encounter in operation.

We argue that critical situations (e.g., merges, lane changes, and unprotected turns) require the joint prediction of a set of multiple **interacting** objects, not just a single object. An example of a pedestrian and vehicle interacting is illustrated in Figure 1a where a vehicle waits for a pedestrian to fully cross the street before turning. In Figure 1b,



(a) A vehicle waits for a pedestrian to fully cross the crosswalk before commencing a turn.



(b) A vehicle accelerates onto the street only after the incoming vehicle turns.

Figure 1: **Examples of interactions between agents in a scene in the WAYMO OPEN MOTION DATASET.** Each example highlights how predicting the joint behavior of agents aids in predicting likely future scenarios. Solid and dashed lines indicate the road graph and associated lanes. Each numeral indicates a unique agent in the scene.

the orange vehicle accelerates into the street only after ensuring the incoming blue vehicle's intention is to decelerate and turn off of the street. Most existing datasets have focused on single agent representation, but there has been considerably less work on interaction modeling at a large scale, which motivates this work.

The goal of this work is to provide a large scale, diverse dataset with specific annotations for interacting objects to promote the development of models to jointly predict interactive behaviors. In addition, we aim to supply object behaviors over a wide range of road geometries, and thus provide a large set of annotated interactions over a diverse set of locations. To generate such a set, we develop criteria for mining interactive behavior over a large corpus of driving data. We explicitly annotate groups of interacting objects in both training and validation/test data to enable development of models that jointly predict the motion of multiple agents as well as individual prediction models.

We aim to provide high quality object tracking data to reduce uncertainty due to perception noise. The cost of hand labeling a dataset of the required size is prohibitive. Instead we use a state-of-the-art automatic labeling system [26] to provide high quality detection and tracking data of objects in the scenes. In contrast with many datasets which provide tracking from on-board autonomous systems, the off-board automatic labeling system provides higher accuracy as it is not constrained to run in real time. These high quality tracks allow us to focus on understanding the complexity of object behavior, rather than on dealing with perception noise.

Evaluation of interactive prediction models requires metrics formulated for joint predictions as motivated by recent work [33, 6, 34, 28]. In Section 4, we discuss existing work on generalizing metrics to the joint prediction case. We also propose a novel mean Average Precision (mAP) metric to capture the performance of models across different object types, prediction time scales, and trajectory shape buckets (*e.g.*, u-turns, left turns). This method is inspired by metrics used in the object detection literature and overcomes limitations in currently adopted metrics. We discuss how this metric attempts to address issues with existing metrics.

We name our large-scale interactive motion dataset: WAYMO OPEN MOTION DATASET. It will be made publicly available to the research community, and we hope it will provide new directions and opportunities in developing motion forecasting models. We summarize the contributions of our work as follows:

- We release a large-scale dataset for motion forecasting research with specifically labeled interactive behaviors. The data is derived from high quality perception output across a large array of diverse scenes with rich annotations from multiple cities.
- We provide novel metrics for motion prediction analysis along with challenging benchmarks for both the marginal and joint prediction cases.

	Lyft	NuSc	Argo	Inter	Ours
# unique tracks	53.4 m [§]	4.3 k	11.7 m [‡]	40 k	7.64 m
Avg track length	1.8 s [§]	-	2.48 s [‡]	19.8 s [*]	7.04 s [††]
Time horizon	5 s	6 s	3 s	3 s	8 s
# segments	170k	1k	324k	-	104k
Segment duration	25 s	20 s	5 s	-	20 s
Total time	1118 h	5.5 h	320 h	16.5 h [*]	574 h
Unique roadways	10 km	-	290 km	-	1750 km [††]
Sampling rate	10 Hz	2 Hz	10 Hz	10 Hz	10 Hz
# cities covered	1	2	2	6 [*]	6
# object types	3	1 [†]	1 [‡]	1	3
Boxes	2D	3D	None	2D	3D
3D maps			✓		✓
Offline perception				✓	✓
Interactions				✓	✓
Traffic signal states	✓				✓

Table 1: **Comparison of popular behavior prediction and motion forecasting datasets.** Specifically, we compare Lyft Level 5 [19], NuScenes [4], Argoverse [9], Interactions [39], and our dataset across multiple dimensions. # object types measures the number of types of objects to predict the motion trajectory. Dashed line "-" indicates that data is not available or not applicable. [§] Lyft Level 5 number of unique tracks and average track length are determined through private correspondence. [†] nuScenes [4] provides annotations for 23 objects types (stationary vehicles are removed), but only the vehicle is predicted. [‡] Argoverse [9] provides annotations for 15 object types (Appendix B) but only vehicle is predicted. The number of unique tracks is determined through private correspondence. The average track length is estimated from data. [*] Interactions [39] gathered data from 4 countries including 6 cities (the last statistic is collected through personal communication) and the entire dataset is not divided into segments. The average track length is estimated from data. [††] Our average track length is computed on the 20s segments of the training split. Our total unique roadway distance is calculated by hashing our autonomous vehicle poses as UTM coordinates into 25 meter voxels and counting the number of non-zero voxels.

2. Related Work

Motion forecasting datasets Several existing public datasets have been developed with the primary goal of motion forecasting in real-world urban driving environments, compared in Table 1. The datasets vary in size measured in number of scenes, total time, total miles, number of tracked objects, and number of distinct time segments. While Lyft Level 5 [19] has the most hours of data and NuScenes [4] has rich object taxonomy, they were not collected to capture a wide diversity of complex and interactive driving scenarios. Argoverse [9] was collected for interesting behaviors by biasing sampling towards certain observed behaviors (*e.g.*, lane changes, turns) and road features (*e.g.*, intersec-

tions). The INTERACTION dataset [39] manually selected a small set of specific driving locations (e.g., roundabouts), and times of day (e.g., rush hour) to obtain a dataset with high interaction complexity. We explain our own methodology for collecting interactions in Section 3.1.

Another salient dataset attribute is the time horizon for prediction. Our dataset's forecasting horizon is 8 seconds into the future, considerably longer than others (3 or 5 seconds), as we believe that long term forecasting is necessary for safe and human-like planning, and is intrinsically more difficult. Finally, most datasets are auto-labeled with industry-grade, *onboard* 3D perception stacks, employing LiDAR's, cameras, and/or radar, and provided as-is with noisy state estimates and tracking errors. One exception is the INTERACTION dataset [39] which collects data from drone footage, which is then post-processed offline with detection, tracking and track smoothing. We also put considerable effort into creating high quality state estimates and 3D tracks by employing an offboard 3D detection and tracking pipeline, as discussed in Section 3.3.

We consider perception datasets (e.g., KITTI [15], Waymo Open Dataset [32]) outside of the scope of this discussion as they do not contain enough motion data to build sufficiently complex models. Generating synthetic data [29] is another line of research, but by collecting real-world data, the behaviors have no realism concerns, and are therefore less susceptible to domain adaptation and transfer. We also note there are a host of other motion forecasting datasets which, while popular, are orders of magnitude smaller, have $O(10)$ unique locations, and/or are not focused on driving environment, for example the Stanford Drone Dataset [30], NGSIM [10], ETH [24], UCY [21], Town Center [2].

Jointly consistent multi-agent forecasting Most existing models output independent future distributions per object in a scene, *e.g.* [1, 3, 7, 5, 8, 12, 11, 14, 17, 20, 22, 25, 40]. This is encouraged by the popular metrics, which only measure quality on a per-object level, and by datasets that only require predicting one agent per scene. An important note is that these methods *do* model interactions between objects to achieve better performance, but explicitly modeling joint futures is much less common. There are a few exceptions which model jointly-consistent futures: Precog [28] and MFP [34] employ models which roll out trajectory samples timestep-by-timestep, where each agent's next step sample conditions on all other agents' current and past steps. In contrast, ILVM [6] (also used by TrafficSim [33]), samples from a latent variable from which multiple steps of future joint samples from all agents are decoded, without explicit conditioning on each step of rollout. These works all measure a stricter version of distance error metrics, reporting the per-agent error of the best *joint* configuration. It is important to note that none of the datasets in Table 1 provide such joint metrics in their release, in contrast to our WAYMO OPEN MOTION DATASET.

3. Dataset

The dataset provides high quality object tracks generated using an offboard perception system (described in Section 3.3) along with both static and dynamic map features to provide context for the road environment. Object track states are sampled at 10Hz. Each state includes the object's bounding box (3D center point, heading, length, width, and height), and the object's velocity vector.[1] Due to sensor range or occlusion, measurements of an object's state may not exist at some time steps. A valid flag is provided to indicate which time steps have valid measurements. Map data is provided as a set of polylines and polygons created from curves sampled at a resolution of 0.5 meters. Static map feature types include lane centers, lane boundary lines, road edges, stop signs, crosswalks, and speed bumps. Traffic signal states and the lanes they control are included. In addition to the geometry data, map features also contain additional data specific to each feature type e.g. lane boundaries have a field to indicate if they are a broken white boundary, a double yellow boundary, etc.

Starting with 20 second segments that are specifically mined from interactions as described in 3.1, we create 9.1 second (91 steps at 10Hz) scenes, splitting the data into a 70% training, 15% validation, and 15% test set. We derive two versions of the validation and test sets which we refer to as the standard and interactive versions. The standard validation and test sets provide up to 8 objects to predict in each scene. Selection is biased to require objects that do not follow a constant velocity model or straight paths. The interactive versions of the validation and test sets focus on the interactive portion of the segment and require only the 2 mined interactive objects to be predicted. The original 20 second segments are also provided for research requiring longer time frames.

3.1. Mining for interesting scenarios

We mine for interesting scenarios by first hand-crafting semantic predicates involving agents' relationships—*e.g.*, "agent A changed lanes at time t", and "agents A and B crossed paths with a time gap t and relative heading difference θ". These predicates can be composed to retrieve more complex queries in an efficient SQL and relational database framework on an overall data corpus orders of magnitude larger than the resulting curated WAYMO OPEN MOTION DATASET.

With this framework, we specifically mined for the following pairwise interaction scenarios: merges, lane

[1] Raw videos and sensor data are not part of the release, as including them would increase the dataset to an impractical size (hundreds of TBs).

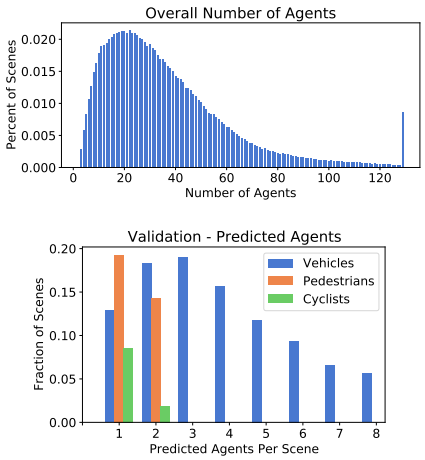


Figure 2: **Our dataset contains many agents including pedestrians and cyclists.** Top: 46% of scenes have more than 32 agents, and 11% of scenes have more than 64 agents. Bottom: In the standard validation set, 33.5% of scenes require at least one pedestrian to be predicted, and 10.4% of scenes require at least one cyclist to be predicted.

changes, unprotected turns, intersection left turns, intersection right turns, pedestrian-vehicle interactions, cyclist-vehicle interactions, interactions with close proximity, and interactions with high accelerations. The pair of interacting objects is annotated within the dataset in each scenario, and the interaction happens close to the 10s mark of the 20s clip.

3.2. Dataset statistics

In contrast with many existing datasets that provide a limited number of agents per scene or agent types, we provide more diverse scenes in terms of the number of agents and types of agents, reflecting many complicated real world driving scenarios like city driving and busy intersections. We show the distribution of number of agents per scene (Figure 2, top). All scenes have at least one vehicle, 57% of scenes have at least one pedestrian (with 20% having four or more), and 16% of scenes have at least one cyclist.

Our dataset contains rich interactions between vehicles, pedestrians, and cyclists, and the users of this dataset must be able to accurately predict the trajectories of *all three classes*, which is not the case in previous datasets [9, 4, 39]. We show the frequency of scenes in which we ask the model to predict each class in the validation set (Figure 2, bottom). Notably, 38.3% of scenes in the validation set require the model to predict more than one type of agent (e.g. a vehicle and a pedestrian or cyclist), and 4.9% of scenes require a model to predict trajectories for all three classes. Finally, in the interactive validation set, where we task the model with predicting the joint future trajectories of two interacting agents, 77.5% of scenes involve two interacting vehicles, 14.9% of scenes involve a vehicle interacting with a

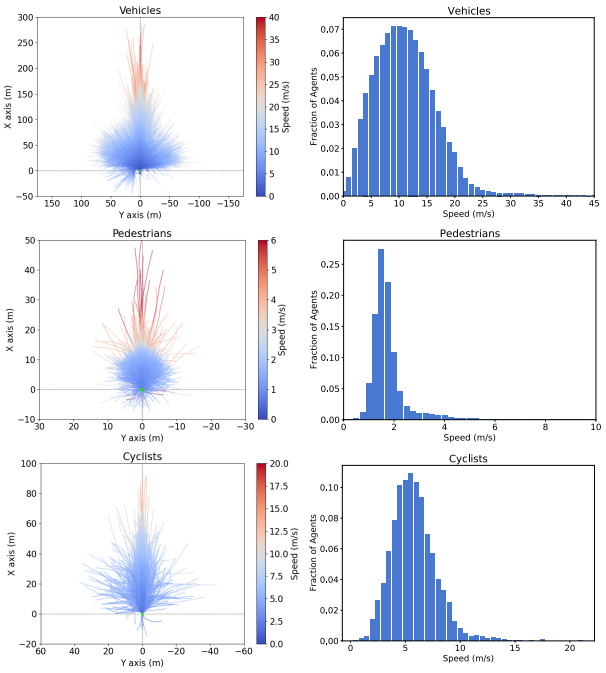


Figure 3: **Agents selected to be predicted have diverse trajectories.** Left: Ground truth trajectory of each predicted agent in a frame of reference where all agents start at the origin with heading pointing along the positive X axis (*pointing up*). Right: Distribution of maximum speeds achieved by all of the agents along their 9 second trajectory. Plots depict variety in trajectory shapes and speed profiles.

pedestrian, and 7.6% of scenes involve a vehicle interacting with a cyclist.

Finally, a motion forecasting dataset should contain diverse scenarios, trajectories, and agent interactions. Table 1 shows that we gather data across a large range of roadways. Figure 3 visualizes the future ground-truth trajectories and maximum speeds of agents we task the models with predicting. These agents represent a wide range of trajectory shapes, speeds, and behaviors, which we believe accurately captures the many different behavioral modes for each class.

3.3. Offboard perception system

Modern motion forecasting systems require a large amount of training data to imitate human maneuvers in complex real-world scenarios. Recently released datasets for motion forecasting [9, 18, 4] are orders of magnitude larger than popular 3D perception datasets [4, 19, 32, 15]. However, manually annotating datasets at such large scales not only incurs exorbitant cost but it also takes tremendous amount of time [26, 37]. Constrained by the high cost, most existing motion forecasting datasets [9, 18] directly employ onboard perception output as groundtruth for trajectory prediction. But limited by the onboard perception

system performance, such annotated 3D objects tracks may have a high degree of state estimation error, lack temporal kinematic consistency or under-/over-segment tracks.

In this work, we aim to alleviate the perception quality bottleneck in existing motion datasets captured by autonomous vehicles and propose using the recently introduced offboard algorithms [26, 37] to automatically generate high-quality motion labels, allowing motion forecasting algorithms to focus on the subtle dynamics and interactions of agents instead of overcoming the noise generated by a constrained, onboard perception system. Compared to the onboard counterpart, offboard perception has two major advantages: 1) it can afford much more powerful models running on the ample computational resources; and 2) it can maximally aggregate complementary information from different views by exploiting the full point cloud sequence including both history and future. Thanks to those advantages, the offboard perception system has shown superior perception accuracy compared to onboard detectors [26] and we have further validated its quality in Section 5.3.

4. Metrics

To measure the accuracy of motion predictions we use a suite of five metrics, which we extend to handle joint predictions over multiple agents as proposed by a few related works [34, 6, 28]. Several common metrics report a minimum error within a trajectory set; when generalized, the joint metric analog constrains the minimum over the best joint configuration of trajectories from a group of agents.

We report standard trajectory-set distance error metrics minADE, minFDE, and Miss Rate (MR), with a custom definition of a match explained below. We also report overlap rate (OR) to measure frequency of predicted tracks' extents overlapping with others'. Finally, inspired by the detection literature, we propose an Average Precision (AP) metric according to the defined MR to measure the precision and recall performance of models across different confidence values. We then account for imbalanced data by reporting mean AP (mAP) over different semantic trajectory motion types.

For each evaluated example scene $\mathbf{e}$, a model makes K possibly joint predictions $S_k, k \in 1 \dots K$. Each S_k contains a scalar confidence c_k, and a trajectory $\mathbf{s}^k = \{s_{a,t}\}_{t=1:T,a=1:A}$ for T future time steps for A agents. Similarly, the ground truth is denoted as $\hat{\mathbf{s}} = \{\hat{s}_{a,t}\}$. The individual object prediction task becomes a special case of this formulation where each joint prediction contains only a single agent $A = 1$.

minADE. The minimum Average Displacement Error computes the L2 norm between $\hat{\mathbf{s}}$ and the closest joint prediction: $\frac{1}{TA} \min_k \sum_a \sum_t ||\hat{s}_{a,t} - s^k_{a,t}||_2$.

minFDE. The minimum Final Displacement Error is equivalent to evaluating the minADE at a single time step T: $\frac{1}{A} \min_k \sum_a ||\hat{s}_{a,T} - s^k_{a,T}||_2$

Overlap rate (OR). The overlap rate is computed by taking the highest confidence joint prediction from each multimodal joint prediction. If any of the A agents in the jointly predicted trajectories overlap at any time with any other objects that were visible at the prediction time step (compared at each time step up to T) or with any of the jointly predicted trajectories, it is considered a single overlap. The overlap rate is computed as the total number of overlaps divided by the total number of predictions. See the supplementary material for details. The overlap is calculated using box intersection, with headings inferred from consecutive waypoint position differences.

Miss rate (MR). A binary match/miss indicator function $\textsc{IsMatch}(\hat{s}_t, s_t)$ is assigned to each sample waypoint at a time t. The average over the dataset creates the miss rate at that time step. A single distance threshold to determine IsMatch is insufficient: we want a stricter criteria for slower moving and closer-in-time predictions, and also different criteria for lateral deviation (*e.g.* wrong lane) versus longitudinal (*e.g.* wrong speed profile). We define it as:

$$IsMatch(\hat{s}_t, s_t) = \mathbb{1}[x^k_t < \lambda^{lon}] \cdot \mathbb{1}[y^k_t < \lambda^{lat}] \quad (1)$$
$$[x^k_t, y^k_t] \coloneqq (\hat{s}_t - s^k_t) \cdot \mathbf{R}_t$$

where $\mathbf{R}_t$ is a 2D rotation matrix defined by the ground truth heading of the agent at timestamp t. The parameters λ^{lon} and λ^{lat} are longitudinal and lateral thresholds that vary with time and velocity. Since agents can have different speeds at time 0, we scale these thresholds by their speed so that we do not over-penalize faster agents: $\lambda^{lon} = \lambda^{lon}_t \gamma(v_x)$ and $\lambda^{lat} = \lambda^{lat}_t \gamma(v_y)$, where $\gamma(v) = (\max(0, \min(1, (v - v_L)/(v_H - v_L)))/2 + 0.5$. We set v_H to 11 m/s and v_L to 1.4 m/s. The time dependent thresholds are as follows:

	λ^{lat}_t	λ^{lon}_t
T=3 seconds	1	2
T=5 seconds	1.8	3.6
T=8 seconds	3	6

For a particular joint configuration, a miss is assigned for time t if any of the trajectories don't match their ground truth trajectory: $\text{MR}_t = \min_k \vee_a \neg IsMatch(\hat{s}_t, s^k_{a,t})$.

Mean average precision (mAP). The Average Precision computes the area under the precision-recall curve by applying confidence score thresholds c_k across a validation set, and using the definition of Miss Rate above to define true positives, false positives, *etc*. Consistent with object detection mAP metrics [23], only one true positive is allowed for each object and is assigned to the highest confidence prediction, the others are counted as false positives. Further inspired by object detection literature [13], we seek an overall metric balanced over semantic buckets, some of

which may be much more infrequent (*e.g.*, u-turns), so report the mean AP over different driving behaviors. The final mAP metric averages over eight different ground truth trajectory shapes: straight, straight-left, straight-right, left, right, left u-turn, right u-turn, and stationary.

5. Experiments

In this section, we evaluate various baseline models on the WAYMO OPEN MOTION DATASET to investigate the importance of rich map annotations (*e.g.* 3D road graph, traffic signal states), interaction context, and joint modeling (Section 5.1). We then compare the standard validation and interactive validation datasets on conditional behavior prediction metrics to show that the interactive validation dataset is both more challenging and more interactive (Section 5.2). Furthermore, we show that our offboard perception system achieves a similar accuracy and perception noise reduction to human labels (Section 5.3). Finally, to provide insight on the performance measurement of motion prediction tasks, we empirically analyze minADE vs. mAP on their ability to reflect the quality of confidence score calibration (Section 5.4). We explicitly do not compare results with existing datasets as differences in the data (e.g. perception noise) can dramatically affect metrics results.

5.1. Baseline model performance

In this section, we evaluate several baseline models on the proposed dataset. First, we consider a *Constant Velocity* model in which we assume the agent will maintain its velocity at the `current` timestamp for all `future` steps. Second, we consider a family of deep-learned models using various encoders, with a base architecture of an LSTM to encode a 1-second history of observed state [16, 1]; this includes agents' positions, velocity, and 3D bounding boxes. In order to measure the importance of particular additional features, we selectively provide additional information:

- Road graph (`rg`): Encode the 3D map information with polylines following [14].
- Traffic signals (`ts`): Encode the traffic signal states with an LSTM encoder as an additional feature.
- High-order interactions (`hi`): Model the high-order interactions between agents with a global interaction graph following [14].

In experiments, combinations of these encodings are concatenated together to create an embedding per-agent. Note that the model is heavily based on the architecture reported in [36], which was one of the top entries on Argoverse and should be considered close to state-of-the-art. We decode K=6 trajectories for output using another MLP with min-of-k loss [12, 35]. See the supplementary material for details.

In Table 2 and 3, we report the marginal metrics on the standard validation/test set and joint metrics on the interactive validation/test set, respectively. Specifically, minADE, miss rate, and mAP at 8s are chosen to be the representatives, and we break down the metrics across 3 object types. The constant velocity model performs quite poorly, e.g., achieving double digit minADE on vehicles. This shows that our dataset contains nontrivial trajectories.

We then investigate the importance of encoding 3D map information, traffic signal states, and high-order interactions between agents. Intuitively, they should all benefit motion forecasting, and this is indeed supported by the experimental results. For example, on the standard validation set (Table 2) for vehicle trajectory prediction, minADE improves from 2.63 to 1.34 and mAP improves from 0.07 to 0.23 when incrementally adding more information in this order. The same trend holds for pedestrian and cyclist as well.

We only evaluate joint metrics on the interactive sets. Since making joint predictions is a relatively new practice, there are no mature, established baselines. In Table 3, we reuse the models trained to make K marginal predictions; but when evaluating on the 2 interactive agents, we select the top K among the K^2 possibilities based on the product of predicted probabilities, as described in [6]. The overall low performance in Table 3 can be attributed to at least 3 factors: the higher difficulty level of the mined interactive agents; the requirement to make good predictions for *both* agents as dictated by the joint version of the metrics; the fact that the predictions are post-hoc manipulations rather than the result of true joint training.

We have argued the importance of jointly predicting interactive behaviors. In Table 4 we provide direct comparison between a base LSTM (without `rg`, `ts`, or `hi`) trained to make marginal or joint predictions for the 2 interactive agents. In the joint prediction model, the neural features for the 2 interactive agents are concatenated with each other to provide the minimal necessary context; the sum of their individual distances to the ground truth (while matching the pairs of trajectories jointly) are used for training; the confidence score are jointly predicted for each pair of trajectories to ensure consistency. When evaluated on the interactive set using joint metrics, this joint model performs favorably against its marginal counterpart. We hope this preliminary experiment can motivate further development of joint models on our dataset, especially the interactive set.

5.2. Quantifying interactivity

Following [36], we use Conditional Behavior Prediction (CBP) to quantify the interactivity in our dataset. [36] introduces a model that can produce either unconditional predictions or predictions conditioned on a "query trajectory" for one of the agents in the scene. If two agents are not interacting, then one's actions have no effect on the other, so knowledge of that agent's future should not change predictions for the other agent. Thus, [36] defines the *degree of influence* agent A has on agent B as the KL divergence

Set	Model	rg	ts	hi	Vehicle minADE ↓	Vehicle MR ↓	Vehicle mAP ↑	Pedestrian minADE ↓	Pedestrian MR ↓	Pedestrian mAP ↑	Cyclist minADE ↓	Cyclist MR ↓	Cyclist mAP ↑
Standard Validation	Const. Vel.				11.0	0.95	0.02	1.55	0.60	0.07	4.17	0.82	0.02
	LSTM				2.63	0.67	0.07	0.73	0.22	0.15	1.86	0.60	0.07
				✓	1.67	0.40	0.16	0.74	0.18	0.18	1.50	0.40	0.12
		✓			1.54	0.32	0.19	0.66	0.14	**0.23**	1.36	0.31	0.17
		✓		✓	1.36	0.26	0.22	**0.63**	0.14	**0.23**	1.29	0.30	0.18
		✓	✓		1.52	0.31	0.18	0.65	0.15	0.20	1.34	0.33	0.15
		✓	✓	✓	**1.34**	**0.25**	**0.23**	**0.63**	**0.13**	**0.23**	**1.26**	**0.29**	**0.21**
Standard Test	Const. Vel.				11.0	0.95	0.02	1.58	0.60	0.06	4.12	0.83	0.03
	LSTM	✓	✓	✓	**1.34**	**0.24**	**0.24**	**0.64**	**0.13**	**0.22**	**1.29**	**0.28**	**0.20**

Table 2: **Marginal metrics on the standard validation and test set.** All metrics computed at 8s. `rg` stands for road graph information. `ts` stands for traffic signal states information. `hi` stands for high-order interactions between agents' features. The constant velocity baseline employs $K = 1$ predicted trajectories; all other models employ $K = 6$.

Set	Model	rg	ts	hi	Vehicle minADE ↓	Vehicle MR ↓	Vehicle mAP ↑	Pedestrian minADE ↓	Pedestrian MR ↓	Pedestrian mAP ↑	Cyclist minADE ↓	Cyclist MR ↓	Cyclist mAP ↑
Interactive Validation	Const. Vel.				10.3	0.98	0.00	3.62	1.00	0.00	6.35	1.00	0.00
	LSTM				4.16	0.88	0.01	2.45	0.93	0.02	4.00	0.98	0.00
				✓	2.89	0.75	0.06	**2.22**	0.93	0.01	3.75	0.94	0.01
		✓			2.94	0.75	0.04	2.39	**0.86**	0.06	3.30	0.88	0.02
		✓		✓	2.45	**0.66**	0.06	**2.22**	**0.86**	0.03	**3.02**	**0.83**	**0.03**
		✓	✓		2.92	0.75	0.04	2.69	0.93	**0.10**	3.24	0.89	0.01
		✓	✓	✓	**2.42**	**0.66**	**0.08**	2.73	1.00	0.00	3.16	**0.83**	0.01
Interactive Test	Const. Vel.				10.3	0.98	0.01	4.56	1.00	**0.00**	6.21	1.00	0.00
	LSTM	✓	✓	✓	**2.46**	**0.67**	**0.08**	**2.47**	**0.89**	**0.00**	**2.96**	**0.89**	**0.01**

Table 3: **Joint metrics on the interactive validation and test set.** See Table 2 for abbreviations and details. Note that these metrics indicate that the interactive split is systematically more challenging.

Model	Vehicle minADE ↓ 3s	5s	8s	Vehicle mAP ↑ 3s	5s	8s
Marginal	**0.65**	1.66	4.16	0.08	**0.07**	0.01
Joint	**0.65**	**1.59**	**3.81**	**0.10**	0.06	**0.03**

Table 4: **Joint modeling is advantageous on interactive agents.** Numbers are from the interactive validation set.

between the unconditional predictions for B and the predictions for B conditioned on A's ground truth future trajectory.

We apply this framework to our interactive and standard validation datasets, computing the KL divergence between unconditional and conditional predictions for every query agent/target agent pair in the dataset. We find that the KL divergences are much larger in the interactive validation dataset than in the standard validation dataset. In particular, 73% of agent pairs in the interactive dataset have KL divergences greater than 10, and 45% have KL divergences greater than 50; in the standard dataset, these numbers are 48% and 28% respectively. Figure 4 presents a full histogram of the KL divergences between unconditional and conditional prediction for each agent pair. Conditioning on a query agent's future trajectories makes little difference in the standard validation dataset but a large difference in the interactive validation dataset, providing evidence that the interactive dataset contains more cases where multiple agents are interacting with and influencing each other. For details on the CBP model, see the supplementary material.

5.3. Analysis of perception data quality

In this section, we study the quality of our offboard perception system and compare them with two alternatives – human labels and baseline detector boxes. Following [26], we conduct a study on the same five validation set run segments from the Waymo Open Dataset (WOD) re-labeled by extra three independent human labelers. With the duplicate human labels, we can analyze the human label consistency to understand the "background noise" in label accuracy. Instead of comparing detection results in average precision [26], we evaluate the box distance errors (DE) in meters by comparing to the original WOD ground truth boxes.

Figure 5 shows that offboard perception achieves an accuracy and distance error distribution similar to human labels. We also show the distance errors of boxes obtained

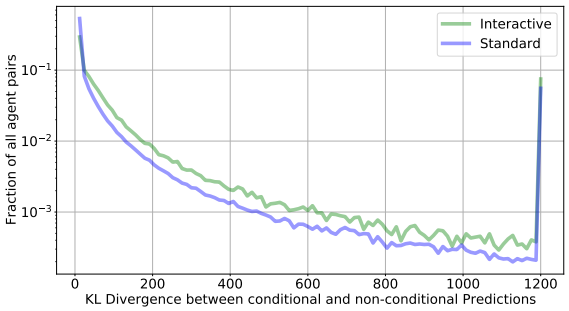


Figure 4: **The interactive split sees much larger improvements from conditional prediction.** Each element in the histogram is one pair of query agent/target agent, and the x axis shows the KL divergence between the unconditional predictions on the target agent and the predictions for the target agent conditioned on the query agent's ground truth future. The higher number of near zero KL divergence examples in the standard set along with the greater number of examples with large KL divergence in the interactive set indicate higher interactivity in the interactive set.

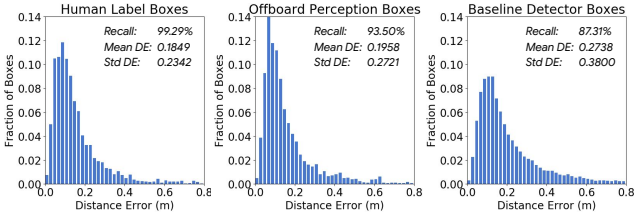


Figure 5: **Distance error statistics of vehicle bounding boxes.** We compare three sets of vehicle bounding boxes with the Waymo Open Dataset (WOD) ground truth boxes on the 5 selected run segments from the val set. The statistics include the histogram of distance errors (capped at 0.8m), the box recall (using a 3D IoU threshold of 0.03), mean distance error and standard deviation (std) of the distance error. Only boxes with at least one point inside are considered. Note that the DE from different boxes are not directly comparable as the recalls are different.

from a baseline detector (Multi-view Fusion [41]) with a Kalman filter-based tracker (the same tracker used in the offboard perception). Using the baseline (onboard) detector leads to a significantly higher mean distance error – this increased perception noise indicates a higher lower-bound minADE that a behavior model can achieve.

5.4. Comparing mAP with minADE

While minADE is widely adopted for performance measurement in motion forecasting tasks [9, 8, 14, 40], it fails to measure the quality of confidence score calibration in the trajectory prediction. In contrast, the mAP metric described in Section 4 provides a measurement of the quality of the

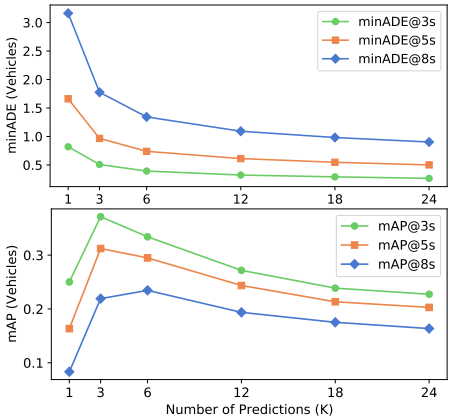


Figure 6: **Comparison of minADE and mAP across increasing numbers of predictions.** Using the best LSTM baseline model in Section 5.1, the minADE (top) artificially improves as one allows for increasing numbers of predictions. Conversely, the mAP (bottom) saturates as the model must produce high quality confidence estimates in addition to accurate trajectories.

confidence score calibration by design. In this section, we perform an analysis of minADE vs. mAP with increasing numbers of predictions at different time steps to show that minADE does not provide a full picture of the model performance while mAP provides more insight.

As shown in Figure 6, minADE artificially improves as the number of predictions increase, while the mAP value peaks at 3 predictions for 3s and 5s, and at 6 predictions for 8s. The minADE scores may improve so long as any of the predictions are good regardless of their confidence score. In contrast, mAP penalizes high confidence false positive predictions and does not continue to improve with the number of predictions. Precision-recall curves for these experiments are shown in the supplementary material.

6. Discussion

In this work we release the WAYMO OPEN MOTION DATASET, a large-scale motion forecasting dataset containing data mined for interactive behaviors across a diverse set of road geometries from multiple cities. The data comes with rich 3D object state and HD map information. Object tracks are generated with a state-of-the-art offboard automatic labeling system which is significantly higher fidelity than typical onboard 3D perception stacks. For evaluation we outline a set of metrics for both per-agent and joint trajectory predictions, including a novel mAP metric to measure performance in a balanced way across driving behaviors. We provide baseline models for both individual and interactive prediction tasks, which we hope provides great opportunities for advancing motion forecasting research.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 3, 6, 11

[2] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011*, pages 3457–3464. IEEE, 2011. 3

[3] Thibault Buhet, Emilie Wirbel, and Xavier Perrotton. Plop: Probabilistic polynomial objects trajectory planning for autonomous driving. *arXiv preprint arXiv:2003.08744*, 2020. 3

[4] 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. 1, 2, 4

[5] Sergio Casas, Cole Gulino, Renjie Liao, and Raquel Urtasun. Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 9491–9497. IEEE, 2020. 3

[6] Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 2, 3, 5, 6, 11

[7] Sergio Casas, Wenjie Luo, and Raquel Urtasun. Intentnet: Learning to predict intention from raw sensor data. In *Conference on Robot Learning*, pages 947–956. PMLR, 2018. 3

[8] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Conference on Robot Learning*, pages 86–99. PMLR, 2020. 1, 3, 8, 11

[9] 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*, pages 8748–8757, 2019. 2, 4, 8, 11

[10] Benjamin Coifman and Lizhe Li. A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset. *Transportation Research Part B: Methodological*, 105:362–377, 2017. 3

[11] Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Jeff Schneider, David Bradley, and Nemanja Djuric. Deep kinematic models for kinematically feasible vehicle trajectory predictions. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10563–10569. IEEE, 2020. 3

[12] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE, 2019. 3, 6

[13] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 5

[14] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. VectorNet: Encoding hd maps and agent dynamics from vectorized representation. In *CVPR*, 2020. 1, 3, 6, 8, 12

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3, 4

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[17] Joey Hong, Benjamin Sapp, and James Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *CVPR*, 2019. 3

[18] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020. 1, 4

[19] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Lyft level 5 perception dataset 2020. https://level5.lyft.com/dataset/, 2019. 2, 4

[20] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017. 3

[21] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007. 3

[22] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556. Springer, 2020. 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[24] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268. IEEE, 2009. 3, 11

[25] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2020. 3

[26] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6134–6144, 2021. 2, 4, 5, 7

[27] Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 772–788, 2018. 11

[28] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2821–2830, 2019. 1, 2, 3, 5, 11

[29] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 3

[30] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565, 2016. 3

[31] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. *arXiv preprint arXiv:2001.03093*, 2020. 11

[32] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1, 3, 4

[33] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[34] Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. In *NeurIPS*, 2019. 1, 2, 3, 5

[35] Luca Anthony Thiede and Pratik Prabhanjan Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9954–9963, 2019. 6

[36] Ekaterina Tolstaya, Reza Mahjourian, Carlton Downey, Balakrishnan Vadarajan, Benjamin Sapp, and Dragomir Anguelov. Identifying driver interactions via conditional behavior prediction. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 1, 6, 12

[37] Bin Yang, Min Bai, Ming Liang, Wenyuan Zeng, and Raquel Urtasun. Auto4d: Learning to label 4d objects from sequential point clouds. *arXiv preprint arXiv:2101.06586*, 2021. 4, 5

[38] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020. 1

[39] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kümmerle, Hendrik Königshof, Christoph Stiller, Arnaud de La Fortelle, and Masayoshi Tomizuka. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv:1910.03088 [cs, eess]*, 2019. 2, 3, 4

[40] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. *arXiv preprint arXiv:2008.08294*, 2020. 1, 3, 8

[41] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pages 923–932, 2020. 8