DROGON: A Trajectory Prediction Model based on Intention-Conditioned Behavior Reasoning

Chiho Choi[1]* **Srikanth Malla**[1] **Abhishek Patil**[1]† **Joon Hee Choi**[2]*
[1]Honda Research Institute, USA [2]Sungkyunkwan University, Korea
{cchoi, smalla}@honda-ri.com patilnabhi@gmail.com jhchoi2019@skku.edu

Abstract: We propose a Deep RObust Goal-Oriented trajectory prediction Network (DROGON) for accurate vehicle trajectory prediction by considering behavioral intentions of vehicles in traffic scenes. Our main insight is that the behavior (*i.e.*, motion) of drivers can be reasoned from their high level possible goals (*i.e.*, intention) on the road. To succeed in such behavior reasoning, we build a conditional prediction model to forecast goal-oriented trajectories with the following stages: (i) *relational inference* where we encode relational interactions of vehicles using the perceptual context; (ii) *intention estimation* to compute the probability distributions of intentional goals based on the inferred relations; and (iii) *behavior reasoning* where we reason about the behaviors of vehicles as trajectories conditioned on the intentions. To this end, we extend the proposed framework to the pedestrian trajectory prediction task, showing the potential applicability toward general trajectory prediction.

Keywords: Trajectory prediction, Intention and behavior, Intersection dataset

1 Introduction

Forecasting participants' trajectories has gained huge attention in recent years. Extensive research has focused on developing robotic systems for safe navigation in indoor and outdoor environments. In understanding human interaction, studies in [1, 2, 3, 4] have advanced our knowledge on pedestrian movement and social behaviors in crowded environments. Recent breakthroughs in automated driving technologies call for such research in the transportation domain. However, current knowledge on pedestrian behavior cannot be directly applied to predicting vehicle trajectories for the following reasons: (i) current models are based on human movement and thus may not directly be applicable to vehicles with faster speed; and (ii) road layouts, which can provide informative motion cues particularly in driving scenes, have been rarely considered in the literature. There have been some research efforts in the transportation community. However, they largely focus on highway scenarios [5, 6], relative trajectory of vehicles respective to ego-motion [7, 8, 9], and ego-motion prediction [10]. Therefore, a robust solution still does not exist for driving environments.

Inspired by a study in [11] on the causation between intentions (*i.e.*, cause) and behaviors (*i.e.*, effect) of humans, we propose a trajectory prediction model using a relationship between future destinations (intentions) and intermediate locations (behaviors) of drivers. In real driving situations, humans are capable of estimating the intention of others based on prior interactions, which corresponds to the potential destination in future. Then, we subsequently anticipate intermediate paths with respect to the intention. From this viewpoint, automated driving or advanced driving assistance systems should be able to address the following questions: (i) Can they learn to estimate intentions and react to interactions with other vehicles using sensory data?; (ii) If so, how can the systems predict accurate trajectories under conditions of uncertain knowledge in a physically plausible manner?

Our framework, DROGON, is designed to address these questions. We infer relational interactions of vehicles with each other and with an environment. Based on this inference, we build a conditional probabilistic prediction model to forecast vehicle's goal-oriented trajectories. That is, we first

*Corresponding author.

†Work done while employed at Honda Research Institute USA. Now at Hilti Group.

4th Conference on Robot Learning (CoRL 2020), Cambridge MA, USA.

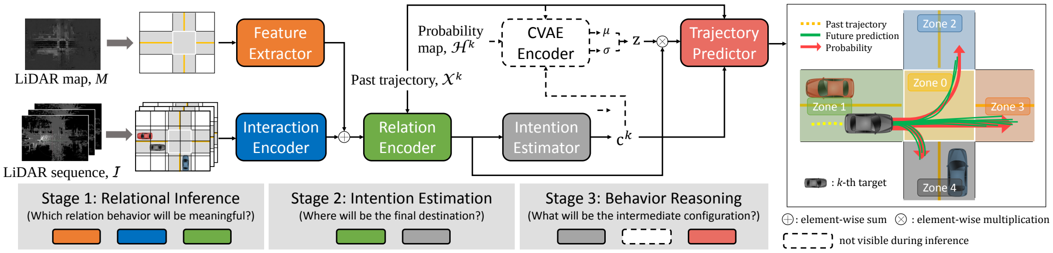

LiDAR map, M
LiDAR sequence, I
Feature Extractor
Interaction Encoder
Relation Encoder
Past trajectory, $\mathcal{X}^k$
Probability map, $\mathcal{H}^k$
CVAE Encoder
μ
σ
z
Intention Estimator
c^k
Trajectory Predictor
Stage 1: Relational Inference
(Which relation behavior will be meaningful?)
Stage 2: Intention Estimation
(Where will be the final destination?)
Stage 3: Behavior Reasoning
(What will be the intermediate configuration?)
Past trajectory
Future prediction
Probability
Zone 0
Zone 1
Zone 2
Zone 3
Zone 4
: k-th target
⊕: element-wise sum
⊗ : element-wise multiplication
not visible during inference

Figure 1: The proposed framework consists of 3 steps: 1) Infer the relational interactions of vehicles; 2) Estimate the probability distributions of intentional goals; 3) Conditionally reason about the goal-oriented behavior as trajectories being sampled from the estimated distributions.

estimate the probability distributions of intentions (*i.e.*, potential destinations of vehicles as zones). Conditioned on the probabilities of formerly estimated intentions (*e.g.*, 5 zones at a four-way intersection), we then predict the multi-modal trajectories of vehicles as illustrated in Fig. 1.

It is not feasible to demonstrate DROGON using the existing vehicle trajectory datasets [12, 13, 14, 15, 16] since they do not provide the intentional destinations of individual agents. Even adding intention labels to these datasets is not a trivial task since we should understand road topology of the scene at each time step with respect to the motion direction and orientation of all agents. Thus, the zone label cannot be automatically added to any existing datasets. Moreover, these datasets are restricted in their capacity to discover functional interactions between vehicles, in terms of the dataset size, motion diversity, and duration. Therefore, we created Honda Intersection Dataset (HID)[3], a large-scale vehicle trajectory prediction dataset to investigate the goal-oriented behavior, containing highly interactive scenarios at four-way intersections in the San Francisco Bay Area.

Furthermore, we extend DROGON to predict pedestrians' motion and demonstrate the applicability of our framework for general-purpose trajectory prediction. Unlike vehicles, the interactive environment of pedestrians would be hypothesized as a nearly open space as they move on paved / unpaved roads, sidewalks, grasslands, etc. We thus relax the assumption of intentional destinations to include various types of regions. We evaluate the extended model using the pedestrian trajectory benchmark datasets [17, 18, 19] to demonstrate the generalizability of the DROGON framework.

2 Related Work

Social interaction modeling Following the pioneering work [20, 21], there has been an explosion of research that has applied social interaction models to data-driven systems. Such models are basically trained using recurrent neural networks to make use of sequential attributes of human movements. In [1], a social pooling layer is introduced to model interactions of neighboring individuals, and [2, 3, 7, 8, 22, 23] improves its performance by using more efficient structure or adding supplemental cues. Recently, the recurrent operation is directly applied to interaction modeling. In [4], the relative importance of each person is captured using the attention mechanism, considering interactions between all humans. It is extended in [15] with an assumption that the same types of road users show similar motion patterns. Although their predictions are acceptable in many cases, these approaches may fail in complex scenes without the perceptual consideration of the surrounding environment such as road structures or layouts.

Scene context as an additional modality Scene context of an interacting environment has been presented in [24] in addition to their social model. However, their restriction of the interaction boundary to local surroundings often causes failures toward far future prediction. [25, 26, 27, 28] subsequently extends local scene context through additional global scale image features. Also, [29] analyzes local scene context from a global perspective and encodes relational behavior of all agents. Motivated by their relational inference from the perceptual observation, we design a novel framework on top of relation-level behavior understanding.

Goal-oriented trajectory prediction The future motion of the target agent has been conditioned on the *external* hypotheses such as the possible action of the neighboring agent in [30] or the potential motion of the ego-agent with respect to others in [8]. Although the proposed approach shares the similar aims for conditioning the trajectory, we directly explore the *internal* intention of the target.

[3] https://usa.honda-ri.com/hid

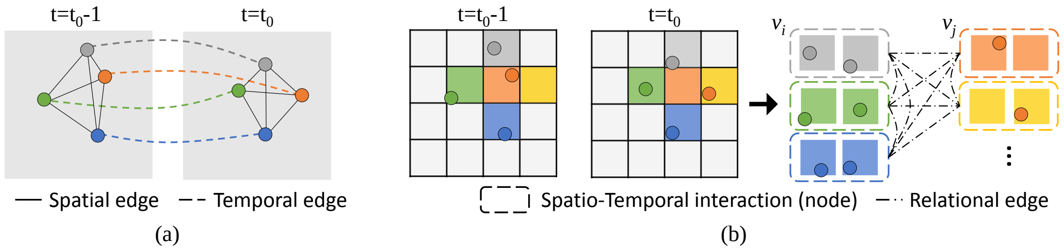


Figure 2: Illustration of different types of graph models to encode spatio-temporal interactions. (a) A node represents the state of each road user, whereas (b) it is a visual encoding of spatio-temporal interactions captured from each region of the discretized grid between adjacent frames.

Recently, a predefined trajectory set is used to represent the driver intention in [31], which is restricted to generate trajectories outside of the anchor set. In contrast, we generate situation-aware behaviors (i.e., motion configuration) with help of the proposed generative pipeline conditioned on the driver's intentional goal (i.e., future destination).

Trajectory datasets in driving scenes The NGSIM [12, 13] dataset has been widely used in the transportation domain [32, 33] for vehicle trajectory forecast with different congestion levels in highways. However, the motion of vehicles and their interactions are mostly simple. The KITTI [14] dataset includes multi-modal sensor data such as LiDAR point clouds, RGB images, and IMUs with various agent categories. However, its small number of tracklets makes the dataset barely used [24] for the purpose of trajectory forecast. Recently, [15] released the dataset for the trajectory prediction task. It has been collected from urban driving scenarios as a subset of ApolloScape [34]. However, they only provide trajectory information with no corresponding visual data of ApolloScape, which is insufficient to discover visual scene context. Subsequently, the Argoverse [16] forecasting dataset is available together with visual information such as 3D Maps and LiDAR data. Although its total number of motions is larger than KITTI, each segment is 5 sec long, which results in the short-term (2 sec observation and 3 sec prediction) prediction horizon. For the general-purpose driving tasks, nuScenes [35] and Waymo [36] is recently introduced. Since they are not designed for trajectory prediction, the motion complexity is not considered. Therefore, we create a new trajectory dataset with more diverse vehicle motions in highly interactive scenarios, particularly at intersections.

3 Preliminaries

3.1 Spatio-Temporal Interactions

Spatio-temporal interactions between road users have been considered as one of the most important features to understand their social behaviors. In [4, 15], spatio-temporal graph models are introduced with nodes to represent road users and edges to express their interactions with each other. To model spatio-temporal interactions, the spatial edges capture the relative motion of two nodes at each time step, and temporal edges capture the temporal motion of each node between adjacent frames as shown in Fig. 2(a). Recently in [29], spatio-temporal features are visually computed using a convolutional kernel within a receptive field. In the spatio-temporal domain, these features not only contain interactions of road users with each other, but also incorporate their interactions with the environment. We use a similar approach and reformulate the problem with a graph model.

3.2 Relational Graph

In the proposed approach, the traditional definition of a node is extended from an individual road user to a spatio-temporal feature representation obtained by exploiting spatial locality in input images. Thus, the edge captures relational behavior from spatio-temporal interactions of road users. We refer to this edge as 'relational edge' as shown in Fig. 2(b). In this view, we define an undirected and fully connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a finite set of $|\mathcal{V}| = n$ nodes ($n = 25$ is used) and $\mathcal{E}$ is a set of relational edges connecting each pair of nodes. Given τ number of input images, we visually extract a node $v_i \in \mathcal{V}$, where v_i is a d-dimensional vector representing spatio-temporal interactions within the i-th region of the discretized grid. The feature r_{ij} of the relational edge between two nodes (v_i, v_j) first determines whether the given interaction pair has meaningful relations from a spatio-temporal perspective through the function ϕ, and then the function θ is used to identify how their relations r_{ij} can affect the future motion of the target k based on its past motion context q^k: $r_{ij} = \phi(v_{ij}; W^r)$ and $f_{ij}^k = \theta(r_{ij}, q^k; W^f)$, where $v_{ij} = v_i \boxtimes v_j$ is the concatenation of two nodes, W^r denotes the weight parameters of ϕ, W^f is those of θ, and q^k is an m-dimensional feature representation extracted from the past trajectory $\mathcal{X}^k = \{X_{t_0-\tau+1}^k, X_{t_0-\tau+2}^k, ..., X_{t_0}^k\}$ of the k-th agent observed in the given perceptual information. We subsequently collect relational information f_{ij}^k from all pairs and perform element-wise sum to produce a unique relational representation $\mathcal{F}^k = \sum_{i,j} f_{ij}^k$ for the k-th agent.

4 Methodology

We transfer knowledge of spatio-temporal relational inference $\mathcal{F}^k$ to predict the probability of intentional goals as well as goal-oriented trajectories. To accomplish this, we assemble building blocks from (i) *relational inference* to encode relational interactions of vehicles using a relational graph, (ii) *intention estimation* to compute the probability distribution of intentional goals based on the inferred relations from the perceptual context, and (iii) *behavior reasoning* to reason about the goal-oriented behavior of drivers as future locations conditioned on the intentional destinations.

4.1 Problem Definition

Given $\mathrm{X}^k = \{\mathcal{I}, M, \mathcal{X}^k\}$, the proposed framework aims to predict δ number of likelihood heatmaps $\mathcal{H}^k = \{H_{t_0+1}^k, H_{t_0+2}^k, ..., H_{t_0+\delta}^k\}$ for the $k\text{-}th$ target vehicle observed in $\mathcal{I}$, where $\mathcal{I} = \{I_{t_0-\tau+1}, I_{t_0-\tau+2}, ..., I_{t_0}\}$ is τ number of past LiDAR images and M is a top-down LiDAR map with a same coordinate with $\mathcal{I}$. Details are provided in the supplementary material. The future locations $\mathcal{Y}^k = \{Y_{t_0+1}^k, Y_{t_0+2}^k, ..., Y_{t_0+\delta}^k\}$ are found using a coordinate of a point with a maximum likelihood from each heatmap H_t^k.

4.2 Behavior Reasoning for Trajectory Prediction

4.2.1 Conditional Trajectory Prediction

We use a conditional VAE (CVAE) framework to forecast multiple possible trajectories of each vehicle. For given observation c, a latent variable z is sampled from the prior distribution $P(z|c)$, and the output heatmaps $\mathcal{H}$ are generated from the distribution $P(\mathcal{H}|z, c)$. As a result, multiple z drawn from the conditional distribution allows the system to model multiple outputs using the same observation c, where $c = q \boxtimes g$ is the concatenation of past motion context q encoded from $\mathcal{X}$ and estimated intention g. In general, the true posterior $P(z|\mathcal{H}, c)$ in maximum likelihood inference is intractable. Therefore, we consider an approximate posterior $Q(z|\mathcal{H}, c)$ with variational parameters predicted by a neural network. The objective of the model is thus written as follows:

$$\mathcal{L}_C = -KL\left(Q(z|\mathcal{H}, c) \| P(z|c)\right) + \mathbb{E}_{Q(z|\mathcal{H},c)}[\log P(\mathcal{H}|z, c)], \tag{1}$$

where $z_l \sim Q(z_l|\mathcal{H}, c) = \mathcal{N}(0, \mathrm{I})$ is modeled as Gaussian distribution.

We respectively build $Q(z|\mathcal{H}, c)$ and $P(\mathcal{H}|z, c)$ as a CVAE encoder and trajectory predictor, on top of convolutional neural networks. At training time, the observed condition c is first concatenated with heatmaps $\mathcal{H}$, and we train the CVAE encoder to learn to approximate the prior distribution $P(z|c)$ by minimizing the Kullback-Leibler divergence. Once the model parameters are learned, the latent variable z can be drawn from the same Gaussian distribution. At test time, the random sample $z \sim \mathcal{N}(0, \mathrm{I})$ is generated and masked with the relational features $\mathcal{F}$ using the element-wise multiplication operator. The resulting variable is passed through the trajectory predictor and concatenated with the observation c to generate δ number of heatmaps $\widehat{\mathcal{H}}$. Details of the network architecture are described in the supplementary material.

4.2.2 Intentional Goal Estimation

We also train the intention estimator for goal-oriented future prediction which employs prior knowledge about the intention of vehicles (at time $t = t_0 + \delta$). Given the relational features $\mathcal{F}$ extracted from vehicle interactions, we estimate the softmax probability S_g for each intention category

$g \in \{1, ..., G\}$ ($G = 5$ in Fig. 1) through a set of fully connected layers with a following ReLU activation function. We compute the cross-entropy from the softmax probability:

$$\mathcal{L}_S = -\sum_{m=1}^{G} \mathbb{1}(m = g) \log S_g, \tag{2}$$

where g is an estimated intention category and $\mathbb{1}$ is the indicator function, which equals 1 if m equals g or 0 otherwise. We use the estimated intention g to condition the process of model prediction. The computed softmax probability S_g is later used at test time to sample z with respect to its distribution.

4.3 Explicit Penalty Modeling

We introduce additional penalty terms specifically designed to constrain the model toward reliance on perceptual scene context and spatio-temporal priors.

4.3.1 Penetration penalty

We encourage the model to forecast all future locations within a boundary of the drivable road in a given environment. To ensure that the predictions do not penetrate outside the road (*i.e.*, sidewalks or buildings), we penalize the predicted points outside the drivable road using the following term:

$$\mathcal{L}_P = \frac{1}{\delta} \sum_{t=t_0+1}^{t_0+\delta} \sum_{j=1}^{J} \left(\mathcal{D}_j \times B(\widehat{\mathcal{H}}_{t,j}) \right), \tag{3}$$

where the function B is the binary transformation with a threshold ϵ_B, $\mathcal{D}$ is the binary mask annotated as zero inside the drivable road, and $J \in \mathbb{R}^{H \times W}$ is total pixels in each likelihood heatmap.

4.3.2 Inconsistency penalty

In order to restrict our model from taking unrealistic velocity changes between adjacent frames, we encourage temporal consistency between frames as a way to smooth the predicted trajectories. We hypothesize that the current velocity at $t = t_0$ should be near to the velocity of both the previous frame ($t = t_0$-1) and next frame ($t = t_0$+1). The inconsistency penalty is defined as

$$\mathcal{L}_I = \frac{1}{\delta - 1} \sum_{t=t_0+1}^{t_0+\delta-1} E(\mathrm{v}_{t-1}, \mathrm{v}_t, \mathrm{v}_{t+1}), \tag{4}$$

where v_t denotes velocity at time t and

$$E(a, x, b) = \max(0, \min(a, b) - x) + \max(x - \max(a, b), 0) \tag{5}$$

is the term to softly penalize the predictions outside of the velocity range.

4.3.3 Dispersion penalty

We further constrain the model to output more natural future trajectories, penalizing the cases where large prediction error is observed. In order to discourage the dispersion of an actual distance error distribution of the model, we use the following penalty:

$$\mathcal{L}_D = \mathrm{Var}\left(\left\{ \|Y_t - \widehat{Y}_t\|_2^2 \right\}_{t=t_0+1}^{t_0+\delta} \right) = \frac{1}{\delta} \sum_{t=t_0+1}^{t_0+\delta} (d_t - \bar{d})^2, \tag{6}$$

where d_t is an Euclidean distance between the predicted location and ground truth at time t and $\bar{d}$ denotes a mean of $\boldsymbol{d} = \{d_{t_0+1}, ..., d_{t_0+\delta}\}$. We observe that the $\mathcal{L}_D$ penalty is particularly helpful to obtain accurate future locations with the concurrent use of the $\mathcal{L}_P$ term.

4.4 Training

At training time, we minimize the total loss drawn in Eqn. 7. The first two terms are primarily used to optimize the CVAE modules which aims to approximate the prior and generate actual likelihood predictions. The third term mainly leads the model's output to be in the drivable road, and the last two terms are involved in generation of more realistic future locations. We set the loss weights as $\zeta = 1$, $\eta = 0.1$, and $\mu = 0.01$ which properly optimized the entire network structures.

$$\mathcal{L}_{Optimize} = -\mathcal{L}_C + \mathcal{L}_S + \zeta\mathcal{L}_P + \eta\mathcal{L}_I + \mu\mathcal{L}_D. \tag{7}$$

Table 1: Quantitative comparison (ADE / FDE in *meters*) for **single-modal** prediction.

Single-modal	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art				
S-LSTM[1]	1.66 / 2.18	2.57 / 4.03	3.59 / 6.19	4.61 / 8.45
S-GAN[2]	1.61 / 3.01	2.06 / 3.83	2.32 / 4.35	4.28 / 7.92
S-ATTN[4]	1.17 / 1.45	1.69 / 2.61	2.41 / 4.45	3.29 / 6.67
Const-Vel[37]	0.52 / 0.85	1.27 / 2.63	2.34 / 5.38	3.70 / 8.88
Gated-RN[29]	0.74 / 0.98	1.14 / 1.79	1.60 / 2.89	2.13 / 4.20
Ours				
DROGON	**0.52 / 0.71**	**0.86 / 1.46**	**1.31 / 2.60**	**1.86 / 4.02**
Baseline				
w/o Intention	0.79 / 1.04	1.20 / 1.85	1.65 / 2.90	2.18 / 4.25
w/o Map	0.65 / 0.86	1.01 / 1.62	1.46 / 2.77	2.02 / 4.23
w/o Penalty	0.60 / 0.81	0.97 / 1.58	1.41 / 2.71	1.98 / 4.20

Table 2: Quantitative comparison (ADE / FDE in *meters*) for **multi-modal** prediction.

Multi-modal	1.0 *sec*	2.0 *sec*	3.0 *sec*	4.0 *sec*
State-of-the-art				
S-LSTM [1]	1.06 / 1.37	1.68 / 2.79	2.46 / 4.55	3.36 / 6.73
S-GAN [2]	1.50 / 2.84	1.94 / 3.52	1.99 / 3.75	3.43 / 6.47
S-ATTN [4]	1.35 / 1.69	1.73 / 2.10	2.09 / 3.11	2.66 / 5.10
Gated-RN [29]	0.60 / 0.80	0.93 / 1.49	1.33 / 2.48	1.82 / 3.74
Ours				
DROGON-Best	0.39 / 0.53	0.65 / 1.14	1.03 / 2.11	1.48 / 3.29
DROGON-Prob	**0.38 / 0.49**	**0.55 / 0.84**	**0.77 / 1.40**	**1.05 / 2.25**
Baseline				
DROGON-E	0.41 / 0.77	0.84 / 1.53	1.33 / 2.57	1.87 / 3.38

4.5 Extension of DROGON

At road intersections, we can define each potential destination in the scene as one of the zones based on its structural topology. In this way, each zone corresponds to the intentional destination of the driver as shown in Fig. 1. However, such a strategy cannot be directly applicable to pedestrians as their interactive environment is hypothesized as a nearly open space. The structural layout of the scene is not as informative as that of vehicles'. We thus relax the assumption for intentional destinations, so they can be any regions in the given environment for pedestrian trajectory prediction. By assuming every grid region in an image as zones, we can generalize the proposed behavior reasoning framework for pedestrian trajectory forecast in the open space. In the rest of this paper, we use a different abbreviation, DROGON-E, for the extended framework. Note that applying an extended method to driving scenes may cause a prediction failure since the future vehicle motion can be generated throughout non-drivable areas like sidewalks or buildings, as validated in Table 2.

5 Experiments

We comprehensively evaluate the proposed approach using Honda Intersection Dataset (HID). The detailed specifications of HID can be found in the supplementary material.

Although the authors are aware of other vehicle trajectory datasets such as [14, 15, 16], we do not use them for one or more of the following reasons: (i) None of the datasets provides the structure-specific intentional destinations of agents. Such zone labels should be acquired by hand-labeling as explained in Sec. 1. It thus makes the demonstration of DROGON infeasible to reason about behaviors conditioned on *internal* intentions; and (ii) Perceptual information such as RGB images or LiDAR point clouds is not provided, which is critical to visually infer relational behavior between agents from our framework. We did not find a straightforward way to evaluate our behavior reasoning framework on these datasets.

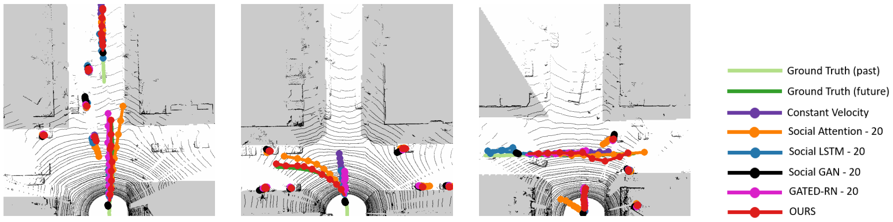
Additionally, we evaluate the extended framework DROGON-E for pedestrian trajectory prediction. Its generalization is validated using three public datasets (SDD [17], ETH [18], and UCY [19]) that contain pedestrian trajectories in diverse interaction scenarios.

5.1 Comparison to Baselines

We conduct ablative tests using HID dataset to demonstrate the efficacy of the proposed DROGON framework by measuring average distance error (ADE) during a given time interval and final distance error (FDE) at a specific time frame in *meters*.

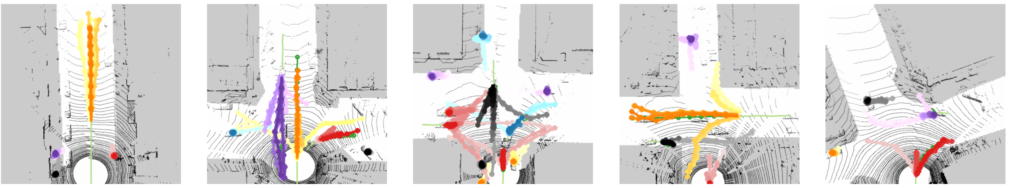
Prior knowledge of intention In order to investigate the efficacy of behavior reasoning, we design a baseline (w/o Intention) by dropping the intention estimator and CVAE encoder from DROGON. As a result, this baseline is not generative, outputting a single set of deterministic locations. In Table 1, the reported error rates indicate that behavior reasoning is essential to predict accurate trajectories under conditions of prior knowledge of intention. It is due to the fact that goal-oriented reasoning is practically helpful to condition the search space and guide the course of future motion. The mean average precision of intention estimation is 71.1% (from DROGON) and 70.2% (from w/o map).

Global scene context We define another baseline model (w/o Map) which does not use global scene context for trajectory forecast. For implementation, we did not add features extracted from the map M into the relational inference stage. In this way, the model is not guided to learn global road lay-



(a) (b) (c)

Figure 3: Qualitative comparison of DROGON with the state-of-the-art algorithms. We visualize the top-1 prediction. Gray mask is shown for non-drivable region.



(a) (b) (c) (d) (e)

Figure 4: (a-e) All 20 trajectories of DROGON-Prob-20 are plotted in interactive scenarios at interactions. We change the intensity of colors for those 20 samples and use different colors for different vehicles. Gray mask is shown for non-drivable region.

outs, similar to relational inference in [29]. As shown in Table 1, the prediction error of this baseline definitely increases against DROGON. The comparison indicates that discovering additional global context encourages the model to better understand about the spatial environment.

Explicit penalty We now remove the penalty terms in the total loss from the proposed DROGON framework at training time. The performance of this baseline model (w/o Penalty) is compared in Table 1. Although its performance is higher than other baseline models, it achieves higher error rate in comparison to DROGON. This is apparent in the sense that the model is not explicitly guided by physical constraints of the real world. Thus, we conclude that these penalty terms are dominant in forecasting accurate future trajectories.

5.2 Comparison with the State of the Arts

We compare the performance of DROGON to several state-of-the-art trajectory prediction approaches [1, 2, 4, 29] that have shown outstanding performance for vehicle trajectory forecast [15, 38, 39, 40]. Extensive evaluations are conducted on tasks for both single-modal and multi-modal prediction. As shown in Table 1 for single trajectory prediction, the performance of S-GAN [2] is consistently improved against S-LSTM [1] all over the time steps. S-ATTN [4] shows further improvement of both ADE and FDE by employing relative importance of individual vehicles. Interestingly, however, their performance is worse than or comparable to the simple constant velocity (Const-Vel) model in [37]. With additional perceptual priors, the network model (Gated-RN in [29]) then performs better than the heuristic approach. DROGON also employs visual information of the physical environment. Additionally, we generate intentional goals and predict a trajectory by reasoning about goal-oriented behavior. As a result, we achieve the best performance against the state-of-the-art counterparts.

For evaluation on multi-modal prediction in Table 2, we generate $S = 20$ samples and report an error of the s-th prediction with minimum ADE (*i.e.* , $\min_{s \in S} \|\mathcal{Y}^k - \widehat{\mathcal{Y}}_s^k\|_2^2$) as proposed in [24, 2]. We design two variants of DROGON with a different sampling strategy: (i) DROGON-Best-20 generates trajectories only conditioned on the best intention estimate; and (ii) DROGON-Prob-20 conditions the model proportional to the softmax probability S_g of each intention category. Similar to single-modal prediction, our models show a lower error rate than that of other approaches. It validates the effectiveness of our behavior reasoning framework for goal-oriented future forecast. In Fig. 3, we display their qualitative comparison in general driving scenarios 3a, by considering the influence of environments (parked cars and road layouts) while making turns 3b, and with an ability to socially avoid potential collisions 3c. DROGON properly forecasts trajectories considering interactions with other vehicles and the environment. Moreover, we achieve the best performance with DROGON-Prob-20. By taking adaptive condition on potential goals, we can eventually ease

Table 3: Quantitative comparison of DROGON-E with state-of-the-art methods using the SDD [17] dataset. In a range of 1 - 4 sec, FDE is reported in *pixels* at 1/5 resolution following [24, 29]. For 4.8 sec in the future, both ADE and FDE are reported using the original resolution as in [41, 31].

Time	Metric	CVAE	S-LSTM	DESIRE	CAR-Net	Gated-RN	MultiPath	Ours
1.0 sec	FDE	1.84	3.38	1.29	-	2.11	-	**1.24**
2.0 sec		3.93	5.33	2.35	-	3.83	-	**2.19**
3.0 sec		6.47	9.58	3.47	-	5.98	-	**3.36**
4.0 sec		9.65	14.57	5.33	-	8.65	-	**4.94**
4.8 sec	ADE	30.91	31.19	19.25	25.72	26.67	17.51	**17.06**
	FDE	61.40	56.97	34.05	51.80	53.93	58.38	**30.90**

Table 4: Quantitative comparison (ADE / FDE in *meters*) of the proposed approach (DROGON-E) with the state-of-the-art methods – S-GAN [2], SoPhie [42], S-BiGAT [43], PMP-NMMP [44], S-STGCNN [45] – using the ETH [18] and UCY [19] dataset.

	ETH_hotel	ETH_eth	UCY_univ	UCY_zara01	UCY_zara02	Average
S-GAN [2]	0.87 / 1.62	0.67 / 1.37	0.76 / 1.52	0.35 / 0.68	0.42 / 0.84	0.61 / 1.21
SoPhie [42]	0.70 / 1.43	0.76 / 1.67	0.54 / 1.24	**0.30** / 0.63	0.38 / 0.78	0.54 / 1.15
S-BiGAT [43]	0.69 / 1.29	0.49 / 1.01	0.55 / 1.32	**0.30** / 0.62	0.36 / 0.75	0.48 / 1.00
PMP-NMMP [44]	**0.61** / 1.08	0.33 / 0.63	**0.52** / 1.11	0.32 / 0.66	0.29 / 0.61	0.41 / 0.82
S-STGCNN [45]	0.64 / 1.11	0.49 / 0.85	0.44 / **0.79**	0.34 / **0.53**	0.30 / 0.48	0.44 / 0.75
Ours	0.68 / **0.95**	**0.10** / **0.16**	0.53 / 0.89	0.45 / 0.81	**0.27** / **0.46**	**0.41** / **0.65**

the impact of misclassification in intention estimation. In Fig. 4, we visualize goal-oriented trajectories reasoned from DROGON-Prob-20. While approaching 4a and passing the intersection 4b-4e, DROGON accordingly predicts goal-oriented trajectories using the intentional destination (zone) of vehicles. Note that our framework is able to predict future dynamic motion of the static vehicles (red and purple in 4d), which can eventually help to avoid potential collisions that might be caused by their unexpected motion.

5.3 Generalization of DROGON

As detailed in Sec. 4.5, we assume every grid region in an image as zones to generalize the proposed behavior reasoning framework. In this way, we further conduct cross-domain validation by evaluating DROGON-E on the widely used benchmark datasets for pedestrian trajectory forecast. We first use the SDD dataset to evaluate the proposed framework comparing with the current state-of-the-art methods [1, 24, 41, 29, 31] on two standard benchmark measures, (i) FDE at 1-4 sec as used in [24, 29] and (ii) ADE / FDE at 4.8 sec as reported in [41, 31]. For evaluation, we divide the original image space into 5×5 regions ($G = 25$), assuming the intentional destination of the target agent belongs to one of regions. As shown in Table 3, DROGON-E achieves the best performance over all time steps compared to the state-of-the-art methods. It validates the efficacy of the proposed behavior reasoning framework. We next evaluate DROGON-E using the ETH and UCY datasets. The same number of intention categories ($G = 25$) are assumed as goals. We report ADE / FDE at 4.8 sec in $meters$ in Table 4. DROGON-E outperforms the state-of-the-art methods [2, 42, 43, 44, 45] from hotel, eth, and zara02 subset, improving average errors. These results further validate the generalization capability of our approach toward pedestrian trajectory prediction.

6 Conclusion

We presented a Deep RObust Goal-Oriented trajectory prediction Network, DROGON, which aims to predict a behavior of human drivers conditioned on their intentions. Motivated by the real world scenarios, the proposed framework estimates the intention of drivers based on their relational behavior. Given prior knowledge of intention, DROGON reasons about the behavior of vehicles as intermediate paths. To this end, multiple possible trajectories of each vehicle are generated considering physical constraints of the real world. For comprehensive evaluation, we collected a large-scale dataset with highly interactive scenarios at intersections and tested DROGON comparing with the current state-of-the-art methods. We further provided a way to generalize the proposed framework for pedestrian trajectory prediction, which also validates the efficacy of behavior reasoning.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

[2] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018.

[3] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani. Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In *CVPR*, 2018.

[4] A. Vemula, K. Muelling, and J. Oh. Social attention: Modeling attention in human crowds. In *ICRA*, 2018.

[5] N. Deo and M. M. Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *IV*, 2018.

[6] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi. Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *IV*, 2018.

[7] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *ICRA*, 2019.

[8] S. Malla, I. Dwivedi, B. Dariush, and C. Choi. Nemo: Future object localization using noisy ego priors. *arXiv preprint arXiv:1909.08150*, 2020.

[9] S. Malla, B. Dariush, and C. Choi. Titan: Future forecast using action priors. In *CVPR*, 2020.

[10] X. Huang, S. McGill, B. C. Williams, L. Fletcher, and G. Rosman. Uncertainty-aware driver trajectory prediction at urban intersections. *ICRA*, 2019.

[11] K. A. Feinfield, P. P. Lee, E. R. Flavell, F. L. Green, and J. H. Flavell. Young children's understanding of intention. *Cognitive Development*, 1999.

[12] J. Colyar and J. Halkias. Us highway 101 dataset. *Federal Highway Administration (FHWA), Tech. Rep. FHWA-HRT-07-030*, 2007.

[13] J. Colyar and J. Halkias. Us highway i-80 dataset. *Federal Highway Administration (FHWA), Tech. Rep. FHWA-HRT-07-030*, 2007.

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[15] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, 2019.

[16] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019.

[17] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.

[18] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.

[19] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer graphics forum*, 2007.

[20] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 1995.

[21] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *CVPR*, 2011.

[22] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *ECCV*, 2020.

[23] I. Dwivedi, S. Malla, B. Dariush, and C. Choi. Ssp: Single shot future trajectory prediction. In *IROS*, 2020.

[24] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017.

[25] H. Xue, D. Q. Huynh, and M. Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *WACV*, 2018.

[26] C. Tang and R. R. Salakhutdinov. Multiple futures prediction. In *NeurIPS*, 2019.

[27] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. Trajectron++: Multi-agent generative trajectory forecasting with heterogeneous data for control. In *ECCV*, 2020.

[28] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, et al. Tnt: Target-driven trajectory prediction. *arXiv preprint arXiv:2008.08294*, 2020.

[29] C. Choi and B. Dariush. Looking to relations for future trajectory forecast. In *ICCV*, 2019.

[30] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *ICCV*, 2019.

[31] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *CoRL*, 2020.

[32] J. Li, H. Ma, W. Zhan, and M. Tomizuka. Coordination and trajectory prediction for vehicle interactions via bayesian generative modeling. In *IV*, 2019.

[33] X. Li, X. Ying, and M. C. Chuah. Grip: Graph-based interaction-aware trajectory prediction. In *ITSC*, 2019.

[34] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. In *CVPRW*, 2018.

[35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

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

[37] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *RAL*, 2020.

[38] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi. Non-local social pooling for vehicle trajectory prediction. In *IV*, 2019.

[39] D. Roy, T. Ishizaka, C. K. Mohan, and A. Fukuda. Vehicle trajectory prediction at intersections using interaction based generative adversarial networks. In *ITSC*, 2019.

[40] H. Bi, Z. Fang, T. Mao, Z. Wang, and Z. Deng. Joint prediction for kinematic trajectories in vehicle-pedestrian-mixed scenes. In *CVPR*, 2019.

[41] A. Sadeghian, F. Legros, M. Voisin, R. Vesel, A. Alahi, and S. Savarese. Car-net: Clairvoyant attentive recurrent network. In *ECCV*, 2018.

[42] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *CVPR*, 2019.

[43] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *NeurIPS*, 2019.

[44] Y. Hu, S. Chen, Y. Zhang, and X. Gu. Collaborative motion prediction via neural motion message passing. In *CVPR*, 2020.

[45] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *CVPR*, 2020.

Supplementary Material

A Contribution

The main contributions of the proposed paper are summarized as follows:

- Propose a trajectory forecast framework to estimate the intention of vehicles by analyzing their relational behavior.
- Reason about the behavior of agents as trajectories conditioned on their intentional destination and intermediate configuration for more accurate prediction.
- Create a new vehicle trajectory dataset with highly interactive scenarios at road intersections in urban areas and residential areas.
- Generalize the proposed framework to the pedestrian trajectory forecast tasks.

B Honda Intersection Dataset

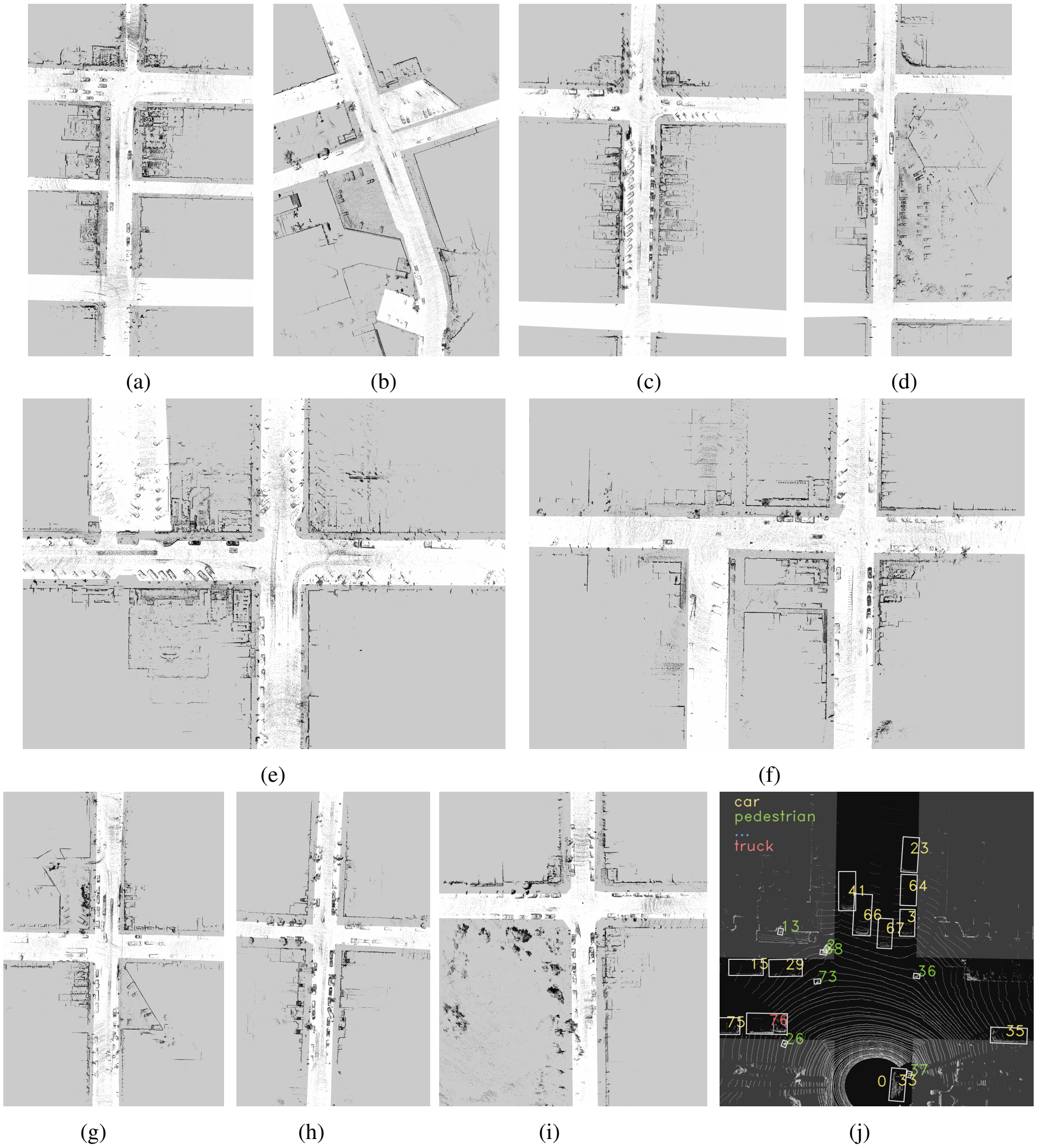
A large-scale dataset is collected in the San Francisco Bay Area (San Fransisco, Mountain View, San Mateo, and Santa Cruz), focusing on highly interactive scenarios at four-way intersections. We chose 213 scenarios in both urban and residential areas, which contain interactions between road users toward an environment. Our intersection dataset consists of LiDAR-based point clouds (full 360° coverage), track-IDs of traffic participants, their 3D bounding boxes, object classes (8 categories including cars and pedestrians), odometry of the ego-car, heading angle (in rad), drivable area mask, and potential destination as zone (intentional goal).

The point cloud data is acquired using a Velodyne HDL-64E S3 sensor, and distortion correction is performed using the high-frequency GPS data. Odometry of the ego-vehicle is obtained via NDT-based point cloud registration. The labels are manually annotated at 2Hz and linearly interpolated to generate labels at 10Hz. We further use the registered point cloud data and divide the intersection by five regions (*i.e.*, in a clockwise direction as illustrated in Fig. 1. Individual road agents are then assigned $g \in \{1, ..., G\}$ to indicate which zone the agent belongs to, with respect to the ego-vehicle.

In Fig. 5, we visualize 10 example scenarios. (a-i) are the bird-eye view maps we created using LiDAR point clouds. Note that the gray mask displays non-drivable regions. Each scenario mainly focuses on interactions at the four-way intersection, and some cases additionally include other types of roads such as three-way intersection (b,f), four-way intersection (a,c,d), or parking lot (b,e). Fig. 5 (g-i) shows four-way intersections in urban areas and residential areas. The preprocessed example is visualized in (j). Also, we compare our intersection dataset with the existing trajectory datasets [14, 15, 16] in Table 5 and highlight unique features of the dataset.

Table 5: Comparison of our new dataset with the driving datasets for trajectory prediction - KITTI [14], Apolloscape [15], and Argoverse [16].

		KITTI [14]	Apolloscape [15]	Argoverse [16]	Ours
No. of scenarios		50	103	-	**213**
No. of frames ($\times 10^3$)		13.1	90	**192**	59.4
No. of object classes		8	5	**15**	8
Sampling frequency (fps)		10	2	10	**10**
Trajectory duration (sec)		flexible	flexible	fixed (5)	**flexible**
No. of intersections		-	-	-	**255**
Type of labels	3D bounding boxes	✓	no	no	**✓**
	Ego-car odometry	✓	no	no	**✓**
	LiDAR point cloud	✓	no	✓	**✓**
	360° coverage	✓	no	✓	**✓**
	Drivable area mask	no	no	✓	**✓**
	Intentional goal	no	no	no	**✓**



(a) (b) (c) (d)

(e) (f)

(g) (h) (i) (j)

Figure 5: Bird-eye view maps of interaction scenarios created using our intersection dataset. Gray mask is shown for non-drivable region.

C Additional Evaluation

C.1 Qualitative Results

We conduct additional qualitative evaluation using the presented intersection dataset.

In Fig. 6, we visualize all 20 trajectories generated from DROGON-Prob-20. By considering road layouts and interactions of each agent with others, the proposed framework appropriately generates goal-oriented trajectories conditioned on their estimated intentions. The output trajectories are inherently multi-modal. In Fig. 7, we also visualize different road layouts other than intersections. DROGON is capable of generating accurate trajectories with respect to road types.

If a car is stopped through a red light, most of existing approaches predict its future motion as a static point based on the observation. However, our DROGON framework is able to predict their future

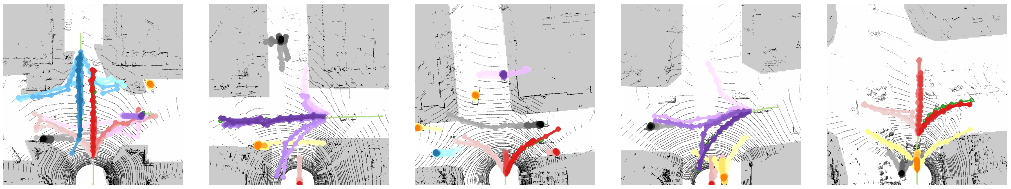

(a) (b) (c) (d) (e)

Figure 6: All 20 trajectories of DROGON-Prob-20 are plotted for multiple vehicles interactions. We change the brightness of a color for those 20 samples and use different color for different vehicles. DROGON accordingly predicts goal-oriented trajectories based on the intention of vehicles. Gray mask is shown for non-drivable region.

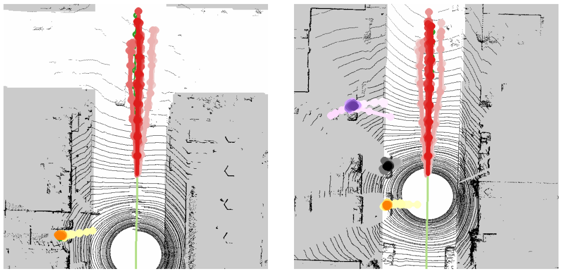

(a) (b)

Figure 7: All 20 trajectories of DROGON-Prob-20 are plotted for multiple vehicles interactions. DROGON accordingly predicts goal-oriented trajectories based on the intention of vehicles while approaching the intersection. Gray mask is shown for non-drivable region.

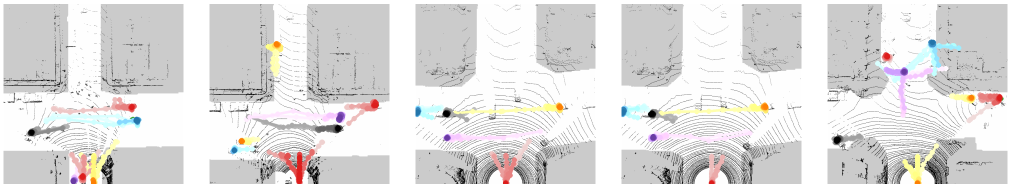

(a) (b) (c) (d) (e)

Figure 8: While the vehicles are stopped, our approach is still capable of predicting their potential movements. This is apparently helpful to avoid potential collisions caused by static road agents.

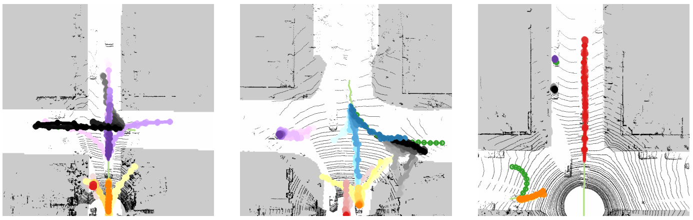

(a) (b) (c)

Figure 9: Failure cases are visualized. (a) Sometimes, samples are drawn through non-drivable region. (b) Some predictions of on-coming vehicle (blue color) are generated toward one-way road. (c) The model experiences difficulties to predict U-turn (Dark green is the ground-truth).

dynamic motion given the intentional destination. By conditioning the model on the intentions, we can encourage the system to generate potential trajectories as shown in Fig. 8. This is apparently helpful to avoid potential collisions that might be caused by unexpected motion of static road agents.

We visualize failure cases in Fig. 9 where (a) trajectories are generated through the non-drivable region (light yellow toward the bottom-right corner), (b) the model does not recognize one-way road (blue toward the bottom), and (c) predictions are not made for u-turn (green on the left). In this view, our future plans include (a) a design of more powerful penalty terms on top of non-drivable region masks, (b) a use of full semantics such as road signs and signals from supplemental RGB images, and (c) a collection of more diverse driving activities to cover such scenarios with more samples.

Table 6: Quantitative comparison (ADE / FDE in normalized *pixels*) of the proposed approach (DROGON-E) with the state-of-the-art methods – S-LSTM [1], SS-LSTM [25], S-GAN [2], Gated-RN [29], SSP [23] – using the ETH [18] and UCY [19] dataset.

	ETH_hotel	ETH_eth	UCY_univ	UCY_zara01	UCY_zara02	Average
S-LSTM	0.076 / 0.125	0.195 / 0.366	0.196 / 0.235	0.079 / 0.109	0.072 / 0.120	0.124 / 0.169
SS-LSTM	0.070 / 0.123	0.095 / 0.235	0.081 / 0.131	0.050 / 0.084	0.054 / 0.091	0.070 / 0.133
S-GAN	0.046 / 0.081	0.087 / 0.169	0.108 / 0.206	0.062 / 0.127	0.058 / 0.114	0.072 / 0.139
Gated-RN	0.018 / 0.033	0.052 / 0.100	0.064 / 0.127	0.044 / 0.086	0.030 / 0.059	0.044 / 0.086
SSP	0.018 / 0.031	0.036 / 0.064	0.059 / 0.120	0.038 / 0.078	0.046 / 0.094	0.039 / 0.077
DROGON-E	**0.011 / 0.018**	**0.022 / 0.031**	**0.035 / 0.059**	**0.030 / 0.054**	**0.018 / 0.031**	**0.020 / 0.033**

C.2 Quantitative Results

Following the evaluation metric in [25, 29], we additionally report ADE / FDE at 4.8 sec in normalized $pixels$ in Table 6. $G = 25$ is used as intentional goals. DROGON-E consistently outperforms the state-of-the-art methods [1, 25, 2, 29, 23].

D Implementation Detail

D.1 Preprocessing

Every $\tau + \delta$ (past and future) number of point clouds, we first transform this subset to the local coordinates at time $t = t_0 - \tau + 1$ using GPS/IMU position estimates in the world coordinate. Then, we project these transformed point clouds onto the top-down image space that is discretized with a resolution of $0.5m$. Each cell in projected top-down images $\mathcal{I}$ has a three-channel ($C_{\mathcal{I}} = 3$) representation of the height, intensity, and density. The height and intensity is obtained by a laser scanner, and we choose the maximum value of the points in the cell. The density simply shows how many points belong to the cell and is computed by $\log(N + 1)/\log(64)$, where N is the number of points in the cell. We further normalize each channel to be in the range of $[0, 1]$. From these projected top-down images $\mathcal{I} \in \mathbb{R}^{H \times W \times C_{\mathcal{I}}}$ where $H = W = 160$, we create the 2D coordinates of past $\mathcal{X}$ and future trajectories $\mathcal{Y}$ in the local coordinates at time $t = t_0 - \tau + 1$. We further convert 2D coordinates to heatmaps $\mathcal{H}$, following [29].

In addition, we remove dynamically moving agents (vehicles and pedestrians) from raw point clouds to only leave the stationary elements such as roads, sidewalks, buildings, and lanes, similar to [24]. Resulting point clouds are registered in the world coordinate and accordingly cropped to build a map $M \in \mathbb{R}^{H \times W \times C_M}$ in the local coordinates at $t = t_0 - \tau + 1$ (same as $I_{t_0 - \tau + 1}$). We observed that the density is always high when the ego-vehicle stops moving, and the height of the hilly road is not consistent when registered. Therefore, only the intensity channel (*i.e.*, $C_M = 1$) is used.

D.2 Implementation of DROGON

Given input LiDAR sequence $\mathcal{I}$, we first concatenate τ images and pass through the interaction encoder that extracts spatial interactions within a receptive field. Details of the network structure are shown in Table 7. Another network with the same configuration is used as a feature extractor to extract rich perceptual information about global scene context using the top-down intensity image M. Then, we slice the previously extracted spatial features as τ maps and perform (i) element-wise addition to combine each map with scene context information and (ii) 1×1 convolution (with 24 kernels) for further feature embedding. For the rest of relational inference, we follow the temporal interaction encoder and relation gate module in [29].

For intentional goal estimation, we use the relational features $\mathcal{F}$ as input and go through four fully connected layers with a following ReLU activation function as in Table 8. To reason about the goal-oriented future behaviors as trajectories, we use a CVAE-based encoder and trajectory predictor. At training time, we train CVAE encoder to output the mean μ and the standard deviation σ, which learns to approximate the prior distribution $P(z|c)$ and thus samples z from $\mathcal{N}(0, \mathrm{I})$. Specifically,

Table 7: Overall architecture of the interaction encoder and feature extractor. Conv2D denotes 2D convolutional layer.

	1	2	3	4	5	6	7	8
Layers	Conv2D	Conv2D	Conv2D	Conv2D	Conv2D	Conv2D	Conv2D	Conv2D
No. of filters	24	24	24	24	24	24	24	24
Width	5	1	5	1	5	1	5	1
Height	5	1	5	1	5	1	5	1
Stride	3	1	3	1	2	1	2	1

we use the ground-truth heatmaps $\mathcal{H}$ as input and the past motion context q and the ground-truth intention g as condition. Note that we use the estimated intention g at test time. The concatenated information $c = q \boxtimes g$ is passed through four 2D convolutional layers (each with a leaky ReLU activation function) and two subsequent fully connected layers. The details of the network structures and output sizes are given in Table 9. Next, we generate z using the resulting μ and σ and perform element-wise multiplication with the relational features $\mathcal{F}$ similar to a guided drop out in [29]. Also note that we randomly sample z from the normal distribution $\mathcal{N}(0, \mathrm{I})$ at test time. The latent variable z is now concatenated with the past motion context q and the ground-truth intention g to condition our trajectory predictor. We use two fully connected layers and four deconvolutional layers with a ReLU function. The final output of trajectory preditor is a set of likelihood heatmaps of size $W \times H \times \delta$, where δ denotes future time steps.

Table 8: The structure of our intention estimator. FC: fully connected layer.

	1	2	3	4	5	6	7	8
Layers	FC	ReLU	FC	ReLU	FC	ReLU	FC	ReLU
Output size	1600		400		100		5	

The network models were trained with a GPU (NVIDIA's TITAN Xp) using the TensorFlow framework. We first trained the modules for relational inference (feature extractor, interaction encoder, relation encoder) with intention estimator. An adam optimizer was used with beta1= 0.9, beta2= 0.999, and batch size of 30. We started with a learning rate of 5e-4 and reduced by a factor of 2 after 15 epochs. These modules were trained for 20 epochs. Then, we trained the CVAE encoder and trajectory predictor together with optimizing the network models using the total loss $\mathcal{L}_{optimize}$ presented in Section 4.4 in the main manuscript. The initial learning rate of 5e-4 was reduced by a factor of 2 after 10 epochs, and the network converged after 15 epochs.

Table 9: The structure of the CVAE encoder. lrelu denotes a leaky ReLU activation function.

	1	2	3	4	5	6
Layers	Conv2D	Conv2D	Conv2D	Conv2D	FC	FC
Output size	$80 \times 80 \times 10$	$40 \times 40 \times 64$	$20 \times 20 \times 64$	$10 \times 10 \times 64$	1024	512
Width	3	3	3	3		
Height	3	3	3	3		
Stride	2	2	2	2		
Activation	lrelu	lrelu	lrelu	lrelu	lrelu	

Table 10: The structure of the trajectory predictor. Deconv denotes a deconvolutional layer.

	1	2	3	4	5	6
Layers	FC	FC	Deconv	Deconv	Deconv	Deconv
Output size	1024	6400	$20 \times 20 \times 64$	$40 \times 40 \times 64$	$80 \times 80 \times 64$	$160 \times 160 \times 10$
Width			3	3	3	3
Height			3	3	3	3
Stride			2	2	2	2
Activation	ReLU	ReLU	ReLU	ReLU	ReLU	softmax

D.3 Run Time

The inference time is 0.0082 sec (in avg.) for the first prediction using Nvidia TITAN X GPU. Note that the output features of Stage 1 in Fig. 1 can be shared by all other agents, which significantly reduces the run time to be 0.0063 sec (in avg.) from the next prediction in the same scene. We believe that our model is applicable to the autonomous driving system that runs in real-time.