LEVERAGING FUTURE RELATIONSHIP REASONING FOR VEHICLE TRAJECTORY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the interaction between multiple agents is crucial for realistic and plausible vehicle trajectory prediction. Accordingly, existing methods tried to model and predict the interaction using observed past trajectories of agents with pooling, attention, or graph-based methods. However, we observed that they easily fail under complex road structures. It is because that they do not explicitly utilize the map information for predicting the relationship, and they only model the relationship between vehicles in a deterministic manner, not in a stochastic manner. In this paper, we propose a new method to model a stochastic future relationship among agents using the lane information contained in the map. Our method first predicts a probability of lane-level waypoint occupancy of vehicles. Then the proposed method utilizes the temporal probability of passing the same lanes to learn the interaction between agents. In addition, we model the interaction using probabilistic distribution. This distribution is learned by the posterior distribution of interaction from GT future trajectory. We validate our method on popular trajectory prediction datasets: nuScenes and Argoverse. The code will be available in public upon acceptance.

1 INTRODUCTION

For safe autonomous driving, predicting the future trajectory of a vehicle is very important. Early heuristic prediction models utilized only the past trajectory of the target vehicle (Zeng et al. (2020); Rhinehart et al. (2018)). Recently, thanks to deep learning, more accurate prediction can be made by considering not only its own state but also the relationship with the High-Definition (HD) map (Liang et al. (2020); Zeng et al. (2021)) or surrounding agents (Lee et al. (2017); Chandra et al. (2020)). Since surrounding vehicles are not stationary in current position, predicting relationships with them is much more complicated and has become essential for realistic trajectory prediction. In addition, since each vehicle is controlled by an individual driver, its interaction has stochastic nature.

Previous works utilized past trajectories of the surrounding vehicles to model interaction by pooling global features, operating multi-head attention, or adapting spatio-temporal graph methods. However, we observed that they easily fail under complex road structures. On the left of Fig. 1, the past trajectories of agents are shown, and on the right is a plot showing the attention weight among agents obtained by the previous method (Mercat et al. (2020)) that learned the interaction among agents using multi-head attention (MHA). Since agent 0 and agent 4 are expected to join in the future, the attention weight between the two is expected to be high. However, it can be seen that the model predicts low attention weight between them. This shows the difficulty of reasoning future relationships between agents based only on past trajectories. The reasoning should have been much easier if it leveraged the road structure.

Looking at the decision-making process of human drivers, they first set their goal and roughly infer how the surrounding vehicles will behave *in the future* while they head to their own goal. After that, they probabilistically infer the interaction with other agents by inferring how much the future path of other vehicles will overlap the path set by themselves. We define interaction from this process as a "Future Relationship" to mimic the decision-making process of human drivers. To model and train the Future Relationship, as shown in Fig. 3, we use the following approaches.

First, because vehicles on adjacent lanes will have much influence on each other, a future location of each vehicle should be obtained to leverage Future Relationship. From the fact that vehicles mainly

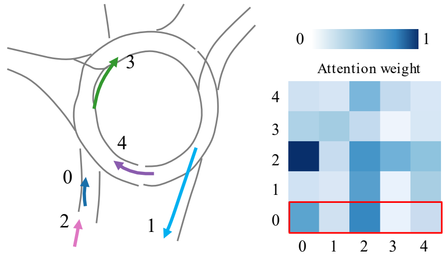

Figure 1: Past motions and their attention map from previous work (Mercat et al. (2020)). Weak relationship is inferred between agents that will highly interact in the future: agent 0 and 4.

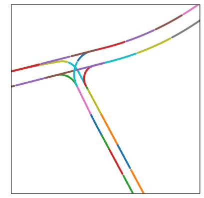

Figure 2: Lane segments represented in different colors.

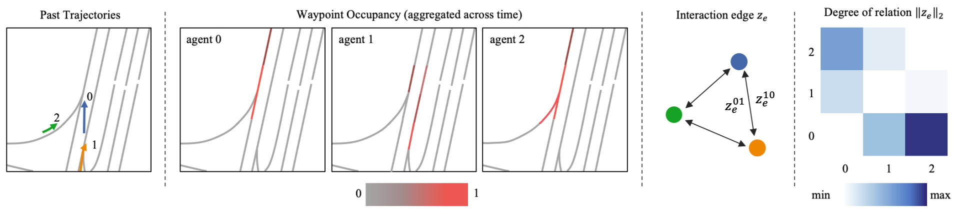

Figure 3: Key concept of the proposed method. Interaction between agents is defined as an edge. The more probability two agents will pass an adjacent lane in the future, the more interaction is expected.

move along lanes, we utilize the lane structure to predict the future location of each vehicle in a stochastic manner. As the future is uncertain, the future location is expressed in waypoint occupancy, which is the probability of a vehicle passing a specific lane segment at every intermediate timestamp. We represent the waypoint occupancy using categorical distribution and sample a future location feature from waypoint occupancy to infer Future Relationship. Fig. 3 shows the waypoint occupancy of each agent aggregated across time. The probability that a vehicle passes that lane at each timestamp is represented using the tone of red color. Each agent chooses multiple lane segments as their future location with probability. In conclusion, waypoint occupancy enables the model to learn how to obtain Future Relationship among vehicles on adjacent lanes in a stochastic manner.

Second, based on sampled future location feature, we infer the Future Relationship in probabilistic distribution. In most vehicle trajectory prediction tasks, interaction between agents is still made in a deterministic manner. Because interaction between agents is highly stochastic and there can be multiple possible interactions, the deterministic relation inference actually averages out interactions that interrupts socially-aware trajectory prediction. Therefore, we model Future Relationship in Gaussian Mixture (GM) distribution. Motivated by Neural Relational Inference (NRI) (Kipf et al. (2018)), we propose a method to train the distribution with the posterior distribution explicitly.

In summary, our contributions are:

1) We propose to inference Future Relationship using lane information, assumming that vehicles generally follow lane.

2) We propose the model to infer the distribution of interaction as Gaussian Mixture, and propose the new learning method to make the model learn the distribution from the posterior.

3) The proposed method is validated on popular real-world vehicle trajectory datasets: nuScenes and Argoverse. In both datasets, there was a remarkable improvement in prediction performance and achieved state-of-the-art performance in nuScenes dataset.

2 Related Work

2.1 Goal-conditioned trajectory prediction

Predicting future trajectories at once is a challenging task. Instead, the first goal-conditional prediction, which oversamples goal candidates and selects a final set of goals, is known to be helpful and has shown state-of-the-art performance (Zhao et al. (2021); Gu et al. (2021); Phan-Minh et al. (2020); Chai et al. (2019); Zhang et al. (2021)), especially in long-time prediction tasks. CoverNet (Phan-Minh et al. (2020)) and MultiPath (Chai et al. (2019)), which quantize the trajectory space to a set of anchors, often generate map-agnostic trajectories that cross non-drivable areas because the surrounding map is neglected. Recent studies have exploited the map information to obtain more performant goal candidates based on the assumption that vehicles follow lanes. TNT (Zhao et al. (2021)) uses goal points sampled from a lane centerline; GoalNet (Zhang et al. (2021)) uses lane segments as anchors and predicts multiple trajectories based on each anchor. Similarly, we set lane segments with their connectivity with nearby segments as goal candidates to model the inherent uncertainty of trajectory prediction. The relationship between lane segments, such as predecessor and successor, provides global connectivity, enabling the model to predict realistic and plausible trajectories.

2.2 Interaction modeling

Trajectory prediction considering the interaction between agents helps to predict socially aware future trajectory. In the very early stage, modeling interaction using a global pooling layer (Deo & Trivedi (2018); Gupta et al. (2018)) were proposed. In another way, researchers attempt to obtain interaction through an attention-based method (Ngiam et al. (2022); Mercat et al. (2020); Vemula et al. (2018)) and GNN-based method (Carrasco et al. (2021); Cao et al. (2021); Zeng et al. (2021); Casas et al. (2020); Liang et al. (2020); Gao et al. (2020)). In the previous method, interactions between agents are learned only with regression loss, which is insufficient to represent dynamic and rapidly changing situations. Besides, a line of works using Neural Relational Inference (NRI) that explicitly predicts and learn interaction using a latent interaction graph was proposed (Kipf et al. (2018)). EvolveGraph (Li et al. (2020)) utilizes two interaction graphs, static and dynamic, (Chen et al. (2021)) uses relation interaction mechanism and spatio-temporal message passing mechanism. Similarly, we apply NRI-based method to predict future trajectory explicitly in complex road structures.

2.3 Multi-modal trajectory prediction

The trajectory prediction task is the problem of predicting the future that has not occurred. Due to the nature of the problem; stochasticity, there is not a unique answer, but several futures are possible. Therefore, recent works use deep generative models such as GAN (Goodfellow et al. (2014)) and VAE (Kingma & Welling (2013)). (Gupta et al. (2018); Kosaraju et al. (2019); Li et al. (2021b)) use the GAN and (Ivanovic & Pavone (2019); Salzmann et al. (2020); Tang & Salakhutdinov (2019)) use the Conditional VAE (Sohn et al. (2015)) to predict multiple futures by sampling multiple latent vectors. As it is important to sample a meaningful latent vector for predicting diverse and plausible future trajectories, trying to define a well-organized latent space became a natural choice (Ma et al. (2021); Bae et al. (2022)). The most closely related work to ours is GRIN (Li et al. (2021a)), which also argues that multi-modality in trajectory prediction comes from two sources: i) personal intention and ii) social relations with other agents. GRIN only uses the past trajectories of agents to model the latent space of personal intention and social relations. However, our model first predicts the lanes to be occupied by agents and then models the latent space of social relations based on the occupancy. Leveraging the future occupancy information to model the interaction makes the model interpretable and easy to be trained since the pseudo labels of occupancy could be generated from ground-truth trajectories. Also, we sample the personal intention of a driver from a set of lane segments feature by following the standard goal-conditioned method (Zhao et al. (2020)).

3 Formulation

In each scene, setting the reference time step to 0, the trajectory during the past $t_p + 1$ time steps of N agents is denoted by $\mathbf{x}^- = \{\mathrm{x}^i_{-t_p:0}\}^{1:N}$ and $\mathbf{x}^+ = \{\mathrm{x}^i_{1:t_f}\}^{1:N}$ for ground-truth (GT) future trajectory where the location of the i-th agent at the t-th time step is $\mathrm{x}^i_t = (x^i_t, y^i_t)$. We use lane information from the HD map, expressed as M segmented lane polylines (Zhao et al. (2021); Varadarajan et al. (2022)). The segmented lane polylines are shown in Fig. 2, where different colors denote different individual segments. The lane information is represented as a graph $\boldsymbol{\mathcal{G}} = (\boldsymbol{\ell}, \mathbf{e})$ where the node $\boldsymbol{\ell} = \{x^m_{1:s}, y^m_{1:s}\}^{1:M}$ is the location of M lane segments of sequence length s. We define edge **e** as five types: predecessor e^m_{pre}, successor e^m_{suc}, left neighbor e^m_{left}, right neighbor e^m_{right}, and

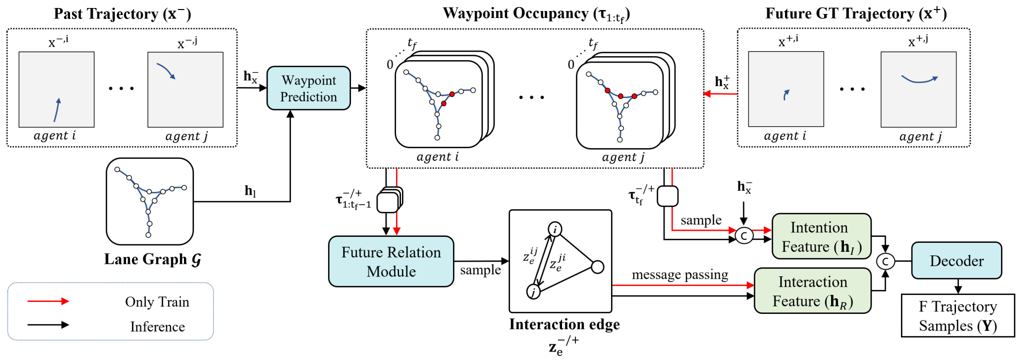

Figure 4: Overall structure. By prediction from past motions or directly getting from future motions, waypoint occupancy $\boldsymbol{\tau}_{1:t_f}$ is obtained. Goal lane features sampled from $\boldsymbol{\tau}_{t_f}$ and agent feature is aggregated to obtain intention feature. $\boldsymbol{\tau}_{1:t_f-1}$ is fed to FRM then interaction edge among agents are sampled. With the edge, message passing is performed to obtain interaction feature. Decoder predicts F future trajectory from intention feature and interaction feature.

in-the-same-intersection e^m_{inter}. The overall input observation is denoted by $\mathbf{X} = \{\mathbf{x}^-, \boldsymbol{\mathcal{G}}\}$. We also define the future lane occupancy (waypoint occupancy) denoted as $\boldsymbol{\tau}_{1:t_f} = \prod_{i,m}^{N,M} p^i_{\ell,1:t_f}(\ell^m)$, where $p^i_{\ell,1:t_f}(\ell^m)$ is the probability of the m-th lane segment being occupied by the i-th vehicle during future time step $1 : t_f$. Predicted waypoint occupancy from $\mathbf{X}$ is denoted by $\boldsymbol{\tau}^-_{1:t_f}$ and (binary) GT waypoint occupancy obtained from GT future trajectory is denoted by $\boldsymbol{\tau}^+_{1:t_f}$.

4 Method

As mentioned above, we focus on modeling "Future Relationship" between agents. The naive method to infer the future relationship is to predict all future trajectories of the vehicles and then use the similarity among the trajectories. However, this method needs to perform prediction twice, which is inefficient and redundant. Also, the criteria for calculating the similarity between predicted paths are not clear. Inspired by the idea that the vehicles mainly follow lanes, we predict the interaction between vehicles using the probability of two agents pass the same lane segment. The key idea is that *if two agents are expected to pass the same lane, they will have a high chance to interact in the future*.

The overall structure is shown in Fig. 4. Our model first predicts the waypoint occupancy, that the probability of lane segment where the vehicle is expected to pass during the future time step 1:t_f (Sec. 4.1). The waypoint occupancy of vehicles is fed to a Future Relationship Module (FRM) (Sec. 4.2). From each vehicle's waypoint occupancy, the FRM infers interaction feature between agents as Gaussian Mixture. Message passing through inferred interaction features is performed to update the agent feature. In the decoding stage (Sec. 4.4), updated agent features and goal lane feature are concatenated and fed to the decoder. In addition, the decoder predicts future trajectories from the aggregation of the interaction feature and intention feature. The interaction is learned in two manners: waypoint occupancy is trained by minimizing the cross entropy, and distribution of interaction feature is trained by CVAE (Sec. 4.3). Details of the proposed methods are described in the following sections.

4.1 Waypoint prediction

Past trajectories $\mathbf{x}^-$ and lane graph $\boldsymbol{\mathcal{G}}$ are encoded into agent and lane feature: $\boldsymbol{h}^-_{\mathrm{x}}$ and $\boldsymbol{h}_\ell$. Then following TNT (Zhao et al. (2021)), we first predict goal lane probability ($\boldsymbol{\tau}^-_{t_f}$) as categorical distribution, then sample goal lane feature ($\hat{\boldsymbol{h}}^-_\ell$) for all agents. Here, we additionally predict waypoint occupancy during 1:$t_f - 1$ timesteps ($\boldsymbol{\tau}^-_{1:t_f-1}$). Because waypoint occupancy varies according to the sampled goal, it is predicted through Cross Attention (CA) as:

$$\boldsymbol{\tau}^-_{1:t_f-1} = \mathrm{softmax}(\mathrm{CA}([\boldsymbol{h}^-_{\mathrm{x}}, \boldsymbol{h}_\ell], \hat{\boldsymbol{h}}^-_\ell)) \quad \in \mathbb{R}^{N\times M\times(t_f-1)} \tag{1}$$

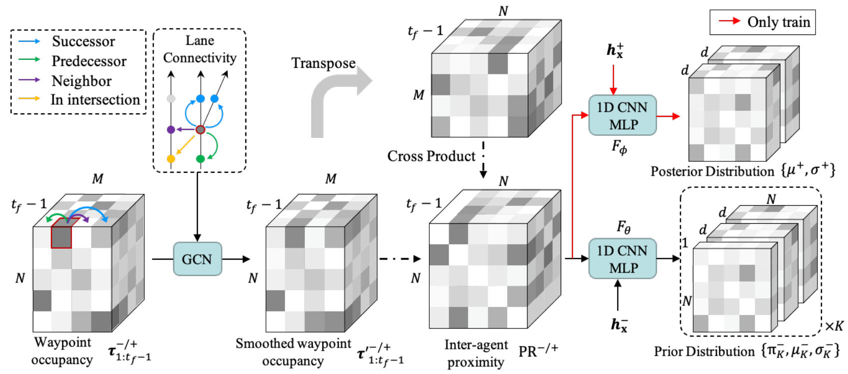

Figure 5: Future Relationship Module. During inference, inferred waypoint occupancy of agents ($\boldsymbol{\tau}^-_{1:t_f-1}$) is fed to GCN. Then, it is cross-producted by itself to get inter-agent proximity (PR$^-$). Prior of interaction ($\boldsymbol{\pi}^-_K, \boldsymbol{\mu}^-_K, \boldsymbol{\sigma}^-_K$) is then obtained. During training, GT waypoint occupancy ($\boldsymbol{\tau}^+_{1:t_f-1}$) is fed to get posterior of interaction ($\boldsymbol{\mu}^+, \boldsymbol{\sigma}^+$). Prior is modeled in Gaussian Mixture, and Posterior in Gaussian distribution.

where $[\cdot,\cdot]$ means concatenation. As waypoint occupancy is a probability, a softmax layer is used so that it becomes $\sum_m^M \tau^{-,i,m}_{1:t_f} = 1$. Sampled goal lane features ($\hat{\boldsymbol{h}}^-_\ell$) is concatenated to agent feature and this *is interpreted as the agents' intentions feature* ($\boldsymbol{h}^-_I$) to reach on the map, so it is utilized as agents' personal intention. Besides, $\boldsymbol{\tau}^-_{1:t_f-1}$ can be thought as a temporal feature containing probabilistic future of the vehicle so it is passed to FRM.

4.2 Future Relationship Module (FRM)

The procedure for the prior distribution of Future Relationship is depicted in Fig. 5. At first, waypoint occupancy is smoothed by Graph Convolutional Network (GCN) (Welling & Kipf (2016)). It is because when a vehicle passes a specific lane, the vehicle will affect other vehicles that pass the adjacent lane, not only exactly the same lane. Different smoothing is applied for each lane connectivity (predecessor, successor, neighbor, in-the-same-intersection) by employing different 2-hop GCN layers where one layer is expressed as:

$$\boldsymbol{\tau}^{-'}_{1:t_f-1} = \sum_{\substack{e \in \{succ, pred,\\ right, left, inter\}}} \sigma\left(D_e^{-1} A_e \boldsymbol{\tau}^-_{1:t_f-1} W_e\right) \tag{2}$$

Here, σ is softmax followed by ReLU activation, $D_e \in \mathbb{R}^{M\times M}$ is a degree matrix, $A_e \in \mathbb{R}^{M\times M}$ is an adjacency matrix, and $W_e \in \mathbb{R}^{t_f-1\times d}$ is a weight matrix for each edge type.

Next, with the obtained waypoint occupancy, we compute the inter-agent proximity (PR) between vehicles i and j (pr^{ij}) which is the temporal probability that two agents would occupy adjacent lanes. The PR can be obtained with a cross product of $\boldsymbol{\tau}^{-'}_{1:t_f-1}$ across lane axis.

$$\mathrm{PR}^- = \{pr^{ij}\}^{1:N,1:N} = \boldsymbol{\tau}^{-'}_{1:t_f-1} \cdot (\boldsymbol{\tau}^{-'}_{1:t_f-1})^\top \quad \in \mathbb{R}^{N\times N\times(t_f-1)} \tag{3}$$

The PR is then utilized as an initial bi-directional interaction edge between agents. Unlike the physical system where we can categorize interaction clearly, the interaction between chaotic systems cannot be clearly categorized. Therefore, its distribution is assumed as a Gaussian Mixture (GM) distribution. In addition, interactions among multiple agents are independent, so we predict interaction distribution for every pair of agents. Then, we define interaction edge e^{ij} between agent i and agent j as a d-dimentional feature which has GM distribution ($p_\theta(\boldsymbol{e}|\mathbf{X}) \sim \prod_{k=1}^K \boldsymbol{\pi}_k \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I}\boldsymbol{\sigma}_k^2)$) following GMVAE (Dilokthanakul et al. (2016)). After that, the distribution parameters ($\boldsymbol{\mu}^-_K, \boldsymbol{\sigma}^-_K, \boldsymbol{\pi}^-_K = \{\boldsymbol{\mu}^{ij}_K, \boldsymbol{\sigma}^{ij}_K, \boldsymbol{\pi}^{ij}_K\}^{1:N,1:N}$) are given by neural network F_θ:

$$\boldsymbol{\mu}^{ij}_K, \boldsymbol{\sigma}^{ij}_K, \boldsymbol{\pi}^{ij}_K = F_\theta([pr^{ij}, h^i_{\mathrm{x}}, h^j_{\mathrm{x}}]) \quad \in \mathbb{R}^{Kd}, \mathbb{R}^{Kd}, \mathbb{R}^K \tag{4}$$

F_θ is composed of MLP layer followed by 1d-convolution along the time dimension (Deo & Trivedi (2018)) to capture interdependency across time. Then, the interaction edge is sampled over $\boldsymbol{\pi}_K^-$ and $\boldsymbol{\epsilon}$. This allows to have K distinct interactions, each interaction following Gaussian distribution:

$$\boldsymbol{\mu}_k^-, \boldsymbol{\sigma}_k^- = \text{argmax}_k(\boldsymbol{\pi}_K^- + \boldsymbol{g}), \quad \boldsymbol{g} \sim \text{Gumbel}(0,1) \tag{5}$$

$$\boldsymbol{z_e}^- = \{z_e^{ij}\}^{1:N,1:N} = \boldsymbol{\mu}_k^- + \boldsymbol{\sigma}_k^- \boldsymbol{\epsilon} \quad \in \mathbb{R}^{N \times N \times d}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{6}$$

Message passing function between the agent feature is then performed based on the interaction edge. As we defined interaction as a d-dimensional feature that has the same dimension as an agent feature, the message function is composed of summation followed by Hadamard product ($\otimes$) of interaction edge and the other agent features. Interaction feature ($\boldsymbol{h}_R^i$) can be obtained by:

$$h_R^i = \sigma'\left(\frac{1}{N-1}\sum_{j \neq i}^{N} z_e^{ij} \otimes F_p(h_{\mathrm{x}}^j)\right) \quad \in \ \mathbb{R}^d \tag{7}$$

Stochasticity is obtained by giving flexibility within a limited range by predicted mean and standard deviation. It is because, unlike a symmetric and deterministic interaction between particles, such as pull or push (Kuramoto (1975)), each driver can asymmetrically consider stochastic interaction.

4.3 CVAE POSTERIOR

To learn the distribution of the interaction, the GT future trajectory ($\mathbf{x}^+$) and GT (binary) waypoint occupancy ($\boldsymbol{\tau}^+$) are utilized. The future trajectory is encoded with the same agent encoder to get the agent feature $\boldsymbol{h}_{\mathrm{x}}^+$. With GT waypoint occupancy ($\boldsymbol{\tau}^+$), the same procedure as described in Eqs. (2)-(3) are performed to obtain posterior spatial proximity (PR$^+$). The difference between the prior is that the posterior distribution of interaction is modeled in a single Gaussian distribution. Thus, F_θ is replaced with F_ϕ which outputs parameters of Gaussian distribution ($\boldsymbol{\mu}^+, \boldsymbol{\sigma}^+ = \{\boldsymbol{\mu}^{ij}, \boldsymbol{\sigma}^{ij}\}^{1:N,1:N}$):

$$\boldsymbol{\mu}^{ij}, \boldsymbol{\sigma}^{ij} = F_\phi([pr^{ij}, h_{\mathrm{x}}^i, h_{\mathrm{x}}^j]) \quad \in \mathbb{R}^d, \mathbb{R}^d \tag{8}$$

We only sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then interaction edge is obtained as $\boldsymbol{z_e}^+ = \boldsymbol{\mu}^+ + \boldsymbol{\sigma}^+ \boldsymbol{\epsilon}$. Message passing is performed the same as Eq. 7, then agent features are updated.

4.4 DECODER

The decoder predicts future trajectories from the aggregation of the interaction feature ($\boldsymbol{h}_R$) and intention feature ($\boldsymbol{h}_I$). During training, GT goal feature ($\hat{\boldsymbol{h}}_\ell^+$) is concatenated to agent feature ($\boldsymbol{h}_{\mathrm{x}}^-$) to get a unique GT intention feature. The GT intention feature is repeated F times, and we sample the interaction feature ($\boldsymbol{h}_R^+$) F times. During inference, the goal feature ($\hat{\boldsymbol{h}}_\ell^-$) is sampled F times from inferred waypoint occupancy. Intention feature is obtained by concatenation of sampled goal feature and agent feature ($\boldsymbol{h}_{\mathrm{x}}^-$). Interaction feature ($\boldsymbol{h}_R^-$) is also sampled F times from prior distribution and inferred goal feature ($\hat{\boldsymbol{h}}_\ell^-$) are used. The decoder is modeled by 2-layer MLP and predicts the sequence of x,y coordinates from the concatenation of the intention feature and interaction feature.

4.5 TRAINING

For intention, because the ground truth waypoint occupancy $\boldsymbol{\tau}^+$ is observable, we train $\boldsymbol{\tau}^-$ with the ground truth. Since the waypoint occupancy is modeled as a classification per agent, we learn it using negative log-likelihood (NLL). Therefore, the NLL loss is computed on lanes that the GT trajectory actually occupies: $\mathcal{L}_{nll} = -\boldsymbol{\tau}_{1:t_f}^+ log(\boldsymbol{\tau}_{1:t_f}^-)$.

In contrast, because interaction $\boldsymbol{e}$ is unobservable, we optimize the negative evidence lower bound (ELBO) to learn the distribution of interaction utilizing the CVAE framework.

$$ELBO = -\mathbb{E}[log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{e}, \boldsymbol{\tau})] + KL[q_\phi(\boldsymbol{e} \mid \mathbf{Y}, \boldsymbol{\tau}^+) \parallel p_\theta(\boldsymbol{e} \mid \mathbf{X}, \boldsymbol{\tau}^-)] \tag{9}$$

q is an approximate posterior distribution of interaction, and p is prior. Because our model only allows interaction posterior to be Gaussian distribution, the Kullback–Leibler (KL) divergence term

Table 1: Comparison on nuScenes test set. Best in **bold**, second best in <u>underline</u>.

Paper	mADE_5	mADE_{10}	MR_5	MR_{10}	mFDE_1
Trajectron++ Salzmann et al. (2020)	1.88	1.51	0.70	0.57	9.52
AgentFormer Yuan et al. (2021)	1.86	1.45	-	-	-
LaPred Kim et al. (2021)	1.47	1.12	0.53	0.46	8.37
P2T Deo & Trivedi (2020)	1.45	1.16	0.64	0.46	10.5
MultiPath Chai et al. (2020)	1.44	1.14	-	-	7.69
GOHOME Gilles et al. (2022a)	1.42	1.15	0.57	0.47	6.99
Autobot Girgis et al. (2021)	1.37	1.03	0.62	0.44	8.19
THOMAS Gilles et al. (2022b)	1.33	1.04	0.55	0.42	<u>6.71</u>
PGP Deo et al. (2022)	<u>1.27</u>	<u>0.94</u>	<u>0.52</u>	<u>0.34</u>	7.17
Ours	**1.19**	**0.89**	**0.49**	**0.31**	**6.64**

Table 2: Impact of prediction time to the proposed modeling in terms of mADE_1/mADE_6.

	Baseline	Ours	Improvement
nuScenes (6sec)	3.23/1.17	2.89/1.10	**10.5**%/6.0%
nuScenes (3sec)	1.26/0.50	1.19/0.48	5.6%/4.0%
Argoverse (3sec)	1.41/0.71	1.33/0.68	5.7%/4.2%

can be simplified as below. The detailed derivation with the reparameterization trick is included in the supplementary material.

$$\mathcal{L}_{KL} = -KL[q_{\pi} \parallel p_{\theta}] \approx log \sum_{k} \boldsymbol{\pi}_k \text{exp}(-KL[q_{\phi} \parallel p_{\theta,k}]) \tag{10}$$

The common drawback of NRI-methods is the "degenerate" problem which leads the decoder to ignore the relation edge during training. So we train the network to that two random characteristics; intention and interaction to have different role. Because the GT trajectory is conditioned on the GT goal feature, we directly use the GT goal feature to compute the reconstruction term.Thus, training in this way restricts the role of interaction edge to be narrowed down to momentary motion: $\mathcal{L}_{recon} = \min_{\boldsymbol{e}}\{\mathbb{E}[log(p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{e}, \boldsymbol{\tau}^{+})]\}$.

The final loss is the sum of the three losses, and they are trained jointly: $\mathcal{L} = \mathcal{L}_{nll} + \mathcal{L}_{KL} + \mathcal{L}_{recon}$.

5 Experiment and Result

We train and evaluate on nuScenes (Caesar et al. (2020)) and Argoverse (Chang et al. (2019)) datasets. The nuScenes/Argoverse datasets provide the past 2/2 seconds of observation, and 6/3 seconds of future trajectory is predicted with 0.5/0.1 sampling time. Training/validation/test sets consist of real-world driving scenes of 32186/8560/9041 in nuScenes and 205942/39472/78143 in Argoverse. For a baseline model, we follow TNT for goal conditioned model and MHA is used to encode interaction from past trajectories. For implementation and training details, refer to the supplementary material.

5.1 Quantitative result

Our method outperforms the SoTA models in all metrics used in nuScenes benchmark, as shown in Tab. 1. The proposed model outperforms the runner-up method, PGP (Deo et al. (2022)), by a large margin. The superiority in validity implies that explicit interaction modeling via inferring waypoint occupancy helps scene understanding of our model compared to the implicit interaction modeling of PGP. When more samples were predicted, mADE_{10} and MR_{10} showed improvements of 5.3% and 8.8%. THOMAS (Gilles et al. (2022b)) had first placed in mFDE_1 by proposing a recombination module that post-processes the marginal predictions into the joint predictions aware of other agents. Our model shows higher performance than that of THOMAS, which indicates better interaction modeling ability without the post-processing. This is possible because inferring future relationship helps to understand the future interaction with other agents; details are provided in the ablation study.

We also validated our method on Argoverse dataset. The proposed Future Relationship Module improved prediction performance compared to baseline in mADE_6/mADE_1 by 4.2%/5.7%. Our method made competitive performance, but it did not achieve SoTA in Argoverse dataset. We

Table 3: Comparison on Argoverse validation set. Best in **bold**, second best in underline.

	mADE_6	mFDE_6	mADE_1	mFDE_1
TNT Zhao et al. (2021)	0.73	1.29	-	-
DenseTNT Gu et al. (2021)	0.73	1.05	-	-
TPCN Ye et al. (2021)	0.73	1.15	1.34	2.95
mmTransformer Liu et al. (2021)	0.71	1.15	-	-
LaneRCNN Zeng et al. (2021)	0.77	1.19	-	-
Multipath Chai et al. (2020)	0.80	1.68	-	-
HiVT Zhou et al. (2022)	**0.66**	**0.96**	-	-
Baseline	0.71	1.03	1.41	3.10
Ours	0.68	0.99	1.33	2.90

Table 4: Ablation studies on nuScenes validation set.

	F=1 mADE/mFDE	F=5 mADE/mFDE
Impact of model design		
Baseline	3.23/7.60	1.26/2.49
Ours w/o FR	3.21/7.59	1.26/2.50
Ours w/o GCN	3.04/6.94	1.22/2.41
Ours w/ Sym	2.99/6.78	1.22/2.35
Importance of multimodal stochastic interaction		
Ours w/ GP	2.98/6.78	1.20/2.33
Ours w/ Deterministic	2.96/6.80	1.28/2.52
Ours (Full)	**2.89/6.61**	**1.19/2.30**

conclude that the relatively small performance gain on Argoverse compared to nuScenes comes from the different configurations of datasets. A model should predict the 6-second future in nuScenes while it only needs to predict the 3-second future in Argoverse. Intuitively, the impact of interaction gets larger in long-time prediction task. We thus measure the performance gain on nuScenes when the prediction horizon is the same as that of Argoverse and report the results in Tab. 2. In 6-second prediction, the proposed interaction inference method improved mADE_1 by more than 10%, but its effect halves in a 3-second prediction which is a similar effect on Argoverse. This result shows that our interaction inference method is more effective in longer prediction tasks.

5.2 Qualitative result

In Fig. 6, prediction samples (F=6) of baseline and our method are depicted. To simply check the effect of our method, we took the samples where two agents are in and the plotted the prediction of a single agent per scene. Green, blue and red solid lines represent past trajectories, future trajectories of other agents and prediction samples of target agents. In two scenes, each target agent sets its intention to where the other agent is likely to pass in the future. The prediction samples from our model (right) generate samples that understand and leverage the interaction with other agent. Unlike the baseline method that ignores other agent and generates spatially uniform trajectories, our model considers other agents to surpass or wait them. Moreover, not only considering multi-modal interaction, we allow stochasticity within a single mode of interaction. As a result, it can be seen that our method generates diverse yet interaction aware samples.

In addition, our method can consider stochastic interaction when multiple agents are present. Similarly, we predicted trajectories of a single agent per scene, and we fixed the intention feature of the target vehicle. With fixed intention, two relation edges are sampled, and corresponding predicted trajectories and degree of relation ($\|\boldsymbol{z}_e\|$) are plotted. In the first row, target agent (0) infers large interaction with agent 2 in sample 1. Because agent 2 is moving in the same direction and predicted to go ahead, the model generates an accelerating trajectory to follow agent 2. In sample 2, in contrast, the interaction between agent 1 is sampled large because they are expected to be in the same intersection. In this case, the model generates decelerating trajectory considering future motion of agent 1. In addition, with fixed intention, it shows that all generated trajectories are properly on goal lane segments. It means that the proposed training method effectively restricts the role of interaction feature affecting momentary motion as we defined.

5.3 Ablation study

Ablation on the impact of model design is shown in the upper part in Tab. 4. **Ours w/o FR** does not consider Future Relationship in interaction modeling. It only utilizes past trajectories to infer the interaction and does not predict waypoint occupancy. Its performance is almost the same as the Baseline which uses MHA of past trajectories to model interaction. This result shows that leveraging Future Relationship indeed helps to infer interaction. **Ours w/o GCN** does not smooth predicted/GT waypoint occupancy, which deteriorates performance as it leads to insufficient lane features and

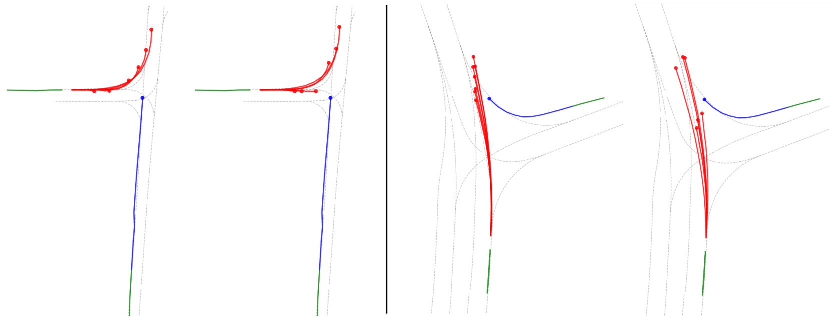

Figure 6: Qualitative results of the proposed method. Green solid line is past trajectories, red line is 6 predicted samples by Baseline (left) and our method (right), blue line is gt future trajectory of other vehicle. Lane centerlines are in gray dashed line. In complex road scene, baseline generate spatially uniform samples regardless of interaction with other vehicles. On the other hand, our method generate diverse yet interaction aware samples: wait or surpass other vehicles that would join in the future.

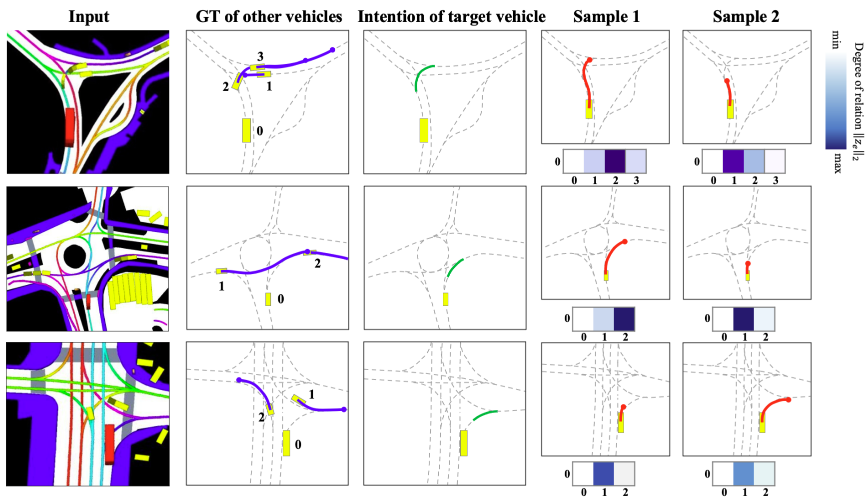

Figure 7: Qualitative results of the proposed method in multi-agent scene.

causes inaccurate inter-agent proximity (PR) estimation. In the proposed model, we allow asymmetric interaction between two agents. **Ours w/ Sym** applies hard symmetric interaction (Li et al. (2019)), and it shows that our asymmetric design is more suitable to model the driver interaction.

The importance of multimodal stochastic interaction modeling is shown in the lower part of Tab. 4. **Ours w/ GP** models the prior distribution of interaction as Gaussian distribution instead of GM. As it only considers a single modality of agent interaction, the performance drops compared to when considering multiple modalities. **Ours w/ Deterministic** only predict mean of interaction feature in Eq. 4. Although it is able to model multi-modal interaction, the diversity is prone to be limited compared to the stochastic counterpart especially when the sample size F is large. The result also shows deterministic interaction modeling severely degrades prediction performance when F is large. On the other hand, **Ours w/ GP** shows relatively less drop in performance as it does not lose stochasticity even after removing the GM prior.

6 CONCLUSION

In this paper, we propose Future Relationship to learn the interaction between vehicles for trajectory prediction effectively. Beyond utilizing only past trajectory to infer interaction, Future Relationship has advantages in explicitly using lane information. We show that using lane information enables proper interaction inference even in complex road structures. In addition, we propose the model that obtains interaction in a probabilistic manner, and generates diverse yet socially plausible trajectory samples. The proposed method brings remarkable performance improvement in popular real-world trajectory prediction datasets: nuScenes and Argoverse.

REFERENCES

Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6477–6487, 2022.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628. IEEE Computer Society, 2020.

Defu Cao, Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Spectral temporal graph neural network for trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1839–1845. IEEE, 2021.

Sandra Carrasco, D Fernández Llorca, and MA Sotelo. Scout: Socially-consistent and understandable graph attention network for trajectory prediction of vehicles and vrus. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1501–1508. IEEE, 2021.

Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *2020 European Conference on Computer Vision (ECCV)*. Springer, 2020.

Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *CoRL*, 2019.

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

Rohan Chandra, Tianrui Guan, Srujan Panuganti, Trisha Mittal, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms. *IEEE Robotics and Automation Letters*, 5(3):4882–4890, 2020.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8740–8749. IEEE Computer Society, 2019.

Siyuan Chen, Jiahai Wang, and Guoqing Li. Neural relational inference with efficient message passing mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7055–7063, 2021.

Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1549–15498. IEEE, 2018.

Nachiket Deo and Mohan M Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020.

Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pp. 203–212. PMLR, 2022.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

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 *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 9107–9114. IEEE, 2022a.

Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Thomas: Trajectory heatmap output with learned multi-agent sampling. In *International Conference on Learning Representations*, 2022b.

Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *2021 Advances in neural information processing systems (NeurIPS)*, 2014.

Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15283–15292. IEEE, 2021.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2255–2264, 2018.

Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2375–2384, 2019.

ByeoungDo Kim, Seokhwan Lee, Seong Hyeon Park, Elbek Khoshimjonov, Dongsuk Kum, Junsoo Kim, and Jun Won Choi. Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pp. 14636–14645. IEEE Computer Vision and Pattern Recognition, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pp. 2688–2697. PMLR, 2018.

Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International symposium on mathematical problems in theoretical physics*, pp. 420–422. Springer, 1975.

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*, pp. 336–345, 2017.

Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020.

Longyuan Li, Jian Yao, Li Wenliang, Tong He, Tianjun Xiao, Junchi Yan, David Wipf, and Zheng Zhang. Grin: Generative relation and intention network for multi-agent trajectory prediction. *Advances in Neural Information Processing Systems*, 34:27107–27118, 2021a.

Xiao Li, Guy Rosman, Igor Gilitschenski, Cristian-Ioan Vasile, Jonathan A DeCastro, Sertac Karaman, and Daniela Rus. Vehicle trajectory prediction using generative adversarial network with temporal logic syntax tree features. *IEEE Robotics and Automation Letters*, 6(2):3459–3466, 2021b.

Yaguang Li, Chuizheng Meng, Cyrus Shahabi, and Yan Liu. Structure-informed graph auto-encoder for relational inference and simulation. In *ICML Workshop on Learning and Reasoning with Graph-Structured Data*, volume 8, pp. 2, 2019.

Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *2020 European Conference on Computer Vision (ECCV)*. Springer, 2020.

Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7573–7582. IEEE Computer Society, 2021.

Yecheng Jason Ma, Jeevana Priya Inala, Dinesh Jayaraman, and Osbert Bastani. Likelihood-based diverse sampling for trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13279–13288, 2021.

Jean Mercat, Thomas Gilles, Nicole El Zoghby, Guillaume Sandou, Dominique Beauvois, and Guillermo Pita Gil. Multi-head attention for multi-modal joint vehicle motion forecasting. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9638–9644. IEEE, 2020.

Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David J Weiss, Ben Sapp, Zhifeng Chen, and Jonathon Shlens. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *2022 International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=Wm3EA5OlHsG`.

Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14062–14071. IEEE, 2020.

Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. r2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *European Conference on Computer Vision*, pp. 794–811. Springer, 2018.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pp. 683–700. Springer, 2020.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 7814–7821. IEEE, 2022.

Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*, 2016.

Maosheng Ye, Tongyi Cao, and Qifeng Chen. Tpcn: Temporal point cloud networks for motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11318–11327, 2021.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9813–9823, 2021.

Wenyuan Zeng, Shenlong Wang, Renjie Liao, Yun Chen, Bin Yang, and Raquel Urtasun. Dsdnet: Deep structured self-driving network. In *European conference on computer vision*, pp. 156–172. Springer, 2020.

Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 532–539. IEEE, 2021.

Lingyao Zhang, Po-Hsun Su, Jerrick Hoang, Galen Clark Haynes, and Micol Marchetti-Bowick. Map-adaptive goal-based trajectory prediction. In *Conference on Robot Learning*, pp. 1371–1383. PMLR, 2021.

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.

Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pp. 895–904. PMLR, 2021.

Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8823–8833, 2022.