Cross Time Domain Intention Interaction for Conditional Trajectory Prediction

Yuxiang Zhao
Sun Yat-sen University
Shenzhen, China
Alibaba Group
Beijing, China
zhaoyao.zyx@alibaba-inc.com

Wei Huang*
Sun Yat-sen University
Shenzhen, China
huangwei5@mail.sysu.edu.cn

Haipeng Zeng
Sun Yat-sen University
Shenzhen, China
zenghp5@mail.sysu.edu.cn

Huan Zhao
Sun Yat-sen University
Shenzhen, China
zhaoh77@mail2.sysu.edu.cn

Yujie Song
Sun Yat-sen University
Shenzhen, China
songyj28@mail2.sysu.edu.cn

Abstract

Human behavior has the nature of mutual dependencies, which requires human-robot interactive systems to predict surrounding agents' trajectories by modeling complex social interactions, avoiding collisions and executing safe path planning. While there exist many trajectory prediction methods, most of them do not incorporate the own motion of the ego agent and only model interactions based on static information. We are inspired by the humans' theory of mind during trajectory selection and propose a **C**ross time domain **i**ntention-interactive method for conditional **T**rajectory prediction(CiT). Our proposed CiT conducts joint analysis of behavior intentions over time, and achieves information complementarity and integration across different time domains. The intention in its own time domain can be corrected by the social interaction information from the other time domain to obtain a more precise intention representation. In addition, CiT is designed to closely integrate with robotic motion planning and control modules, capable of generating a set of optional trajectory prediction results for all surrounding agents based on potential motions of the ego agent. Extensive experiments demonstrate that the proposed CiT significantly outperforms the existing methods, achieving state-of-the-art performance in the benchmarks.

CCS Concepts

• **Human-centered computing** → **Interaction techniques**.

Keywords

Human-robot Interaction, Social Interaction, Conditional Prediction

*Corresponding Author.

MM '25, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754709

ACM Reference Format:
Yuxiang Zhao, Wei Huang, Haipeng Zeng, Huan Zhao, and Yujie Song. 2025. Cross Time Domain Intention Interaction for Conditional Trajectory Prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3754709

1 Introduction

Trajectory prediction of surrounding agents is a crucial component for ensuring safety in autonomous driving systems, as it enables avoiding collisions with human-driven cars. Moreover, predicting trajectories of surrounding agents has extensive applications in human-robot interaction, social robots, drones, and other domains[3, 6, 20, 32]. Humans can navigate through various social scenarios because they have an intrinsic theory of mind, which is the capacity to reason about other human's actions based on their mental states. Imbuing autonomous systems with such capability could enable more informed decision making and motion planning[10, 29, 37, 39]. However, predicting trajectory of agents in real world is challenging since an agent's trajectory is not determined by itself but involves complex social interactions with surrounding agents. Therefore, previous works[7, 11, 12, 38, 41] have proposed a series of methods to model such interactions.

Despite the remarkable progress have been made, these methods still face three critical problems. First, far less attention is paid to the ego agent's own motion, which hinders direct application of these models to real-world robotic systems. Being able to predict surrounding agents' corresponding reactions based on different potential motions of the ego agent is a very crucial capability for downstream tasks such as decision making, motion planning in robotic systems. For example, when the ego agent is faced with multiple trajectories to choose from, it can generate the predicted future trajectories of the surrounding target agents for each candidate trajectory respectively. Then, the trajectory that optimizes the overall system time/efficiency is selected as the final execution trajectory. Second, they do not conduct dynamic modeling of social interactions over time. During the movement of the agent, its intention dynamically changes as it interacts with surrounding agents. Therefore, trajectory prediction models should jointly analyze intentions over time to achieve dynamic modeling of social interactions. Third,

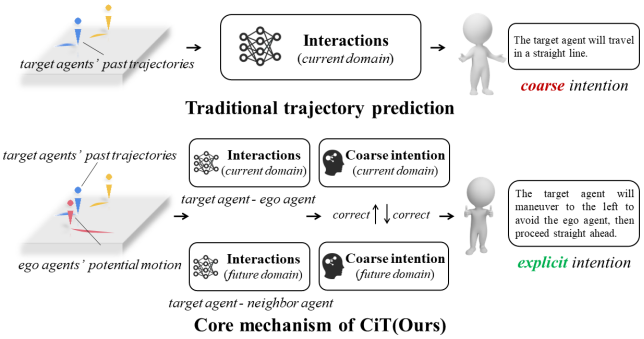

Figure 1: CiT incorporates potential motion of the ego agent to model social interactions, capturing the evolving intentions of the target agent over time. Furthermore, by jointly analyzing the intentions, it achieves feature refinement and information complementarity, leading to clearer and more accurate intention information for trajectory prediction.

different surrounding agents have varying degrees of influence on the target agent whose trajectory we want to predict. Many convolution and social pooling methods extract features of surrounding agents and directly concatenate them without letting the network learn the degree of influence in a prioritized manner. In order to model complex social interactive behaviors more delicately and tightly integrate with robotic system downstream planning and control tasks, we propose the CiT to produce behavior trajectory prediction of all surrounding agents based on ego-agent motion plans. The core of the proposed CiT is to mutually complement and refine intention representations over time through semantic supplementation and feature correction. Figure 1 illustrates the main working mechanism of the proposed model.

CiT contains four key designs: First, we introduce the future trajectory of the ego agent. Note that CiT does not require the exact future trajectory, which is actual undetermined during prediction. CiT only conditions a rough trajectory which can be easily obtained by trajectory generator. This allows our proposed model to generate optional predictions based on candidate trajectories proposed by downstream planning and control modules. Second, in the intention graph construction module, by analyzing the past trajectories of the target agent and neighbor agents, we can infer the current intention of the target agent. To preserve spatial information, we map this intention onto a social tensor according to the target agent's location and refer to it as the "Intention Graph in the Current Time Domain." Furthermore, by incorporating the future trajectory of the ego agent, we can model the potential social interactions between the target agent and the ego agent in the future and predict the future intention of the target agent. Similarly, we map this future intention onto a social tensor based on its location and refer to it as the "Intention Graph in the Future Time Domain." Third, since the intention information from both time domains during the construction of the intention graph is partial and coarse, in the interaction cross domain module, intention information from different time domains interacts with each other. The intention in one time domain proposes a Query to the other time domain and corrects its own intention through the Key and Value from the other time domain. Through joint analysis of intention information over time, features across different agents, spaces, and time domains are fully extracted and fused to obtain a more precise intention representation. Fourth, in the intention influence evaluation module, the network estimates the degree of influence of different intentions on the future trajectory of the target agent, further refining the interaction process. The main contributions are concluded as follows:

- We propose CiT, which comprises four novel designs, including 1) future motion incorporation, captures the interactive aspect in human-robot interaction, 2) intention graphs construction, constructs two types of intention graphs, 3) interaction cross domain, achieves information complementarity between the two intentions, integrating information across different time domains and agents, 4) intention influence evaluation, enables the network to consider the degree of influence from different agents in a prioritized manner.
- In robotic systems, multiple candidate trajectories can be generated to evaluate their corresponding performance in the prediction module, the CiT will provide a highly valuable interface for integrating this trajectory prediction model into robotic system.
- We conduct experiments on two real-world datasets to evaluate our method. Experimental results show that CiT achieves state-of-the-art performance.

2 Related Work

2.1 Trajectory Prediction

Given an observed past trajectory, trajectory prediction aims to forecast each agent's future trajectory over a period of time. Many scene context-aware methods[6, 19, 31] attempt to incorporate bird's eye view road images or lane information as inputs to prune unlikely or low probability trajectories. VectorNet[8] encodes both image information and each agent as vectors and uses a novel graph convolution network. RedMotion[31] proposes a transformer-based model, which learns environment representations through two types of redundancy reduction. LAformer[19]uses an attention-based temporally dense lane-aware estimation module to assess the alignment probability between motion and HD map scene information, enhancing environmental comprehension. Some generative models[16, 35, 40] including GANs and VAEs have also been applied to trajectory prediction tasks to better capture the multimodal nature of future trajectories. Multi-Path[4] uses a set of dense trajectory anchors to capture the agent intentions. CoverNet[25] formulates trajectory prediction as classifying among a set of dense trajectories. These generative methods require a very large number of samples to cover all possible trajectories, especially the ones with low probability that are not necessarily unimportant.

2.2 Interaction-Aware Trajectory Prediction

The movements of an agent is not only influenced by its past trajectory but also related to other neighboring agents. Numerous works[5, 5, 28, 32] have been proposed on how to model the social interaction between agents. A classic study proposed early is the social force model (SFM)[13] which reflects the interaction between

Method	Interaction(tar-sur)	Interaction(sur-sur)	Integration capability	Considers dynamics	Ego agent's motion
S-LSTM (CVPR) [1]	✓				
C-LSTM (CVPR) [7]	✓				
PiP (ECCV) [28]	✓		✓		✓
CF-LSTM (ICRA) [37]	✓				
WSiP (AAAI) [32]	✓	✓			
BAT (AAAI) [17]	✓	✓			
C2F-TP (AAAI) [34]	✓	✓		✓	
Ours	✓	✓	✓	✓	✓

Table 1: A summary of recent state-of-the-art trajectory forecasting methods. Interaction(tar-sur) and Interaction(sur-sur) respectively denote that the proposed method takes into account the interactions between the target agent and surrounding agents, as well as the interactions within the group of surrounding agents. Integration capability denotes the model can be easily integrated into the robotic system. Considers dynamics implies that the method factors in the intention changes during the movement of agents, thereby dynamically modeling social interaction. Ego agent's motion signifies that the method accounts for the potential motion of the ego agent.

the agents by attractive and repulsive forces. Social LSTM[1] extracts the historical trajectory features of each agent with LSTMs separately and discovers the interactions between neighboring agents using social pooling structure. Convolution Social Pooling[7] improves upon this by incorporating convolution layers, which better captures the spatial connections between different agents. INT2[38] presents a large-scale interactive trajectory dataset for interactive trajectory prediction. SocialCircle[36] builds a new angle-based social interaction representation. In this work, we propose a novel cross time domain intention-interaction model for trajectory prediction. Compared to previous methods, our method fuses information across different time domains and spatial scales, empowering the network with stronger feature representation capabilities.

2.3 Conditional Trajectory Prediction

Conditional prediction methods[14, 18, 24, 30] explore the relationship between candidate trajectories of the ego agent and future trajectories of other surrounding agents. Trajectron++[27] incorporates heterogeneous data and future motions in predictions. Precog[26] and PiP[28] attempts to incorporate the future trajectory of the ego agent to explore how surrounding agents would be affected when the ego agent take different candidate trajectories. M2I[30] focuses on agent pairs, where one of the agents is predicted as influencer and the other as reactor. Then it utilizes a conditional model to generate consistent future trajectories for the reactor based on the influencer's marginal future predictions. Different from existing conditional prediction methods, our proposed model only requires rough instead of precise future trajectory, which can be easily obtained even in non-autonomous system scenarios. For instance, a rough left turn trajectory can be inferred based on a vehicle turning on left turn signals, and a roughly straight trajectory can be inferred for a vehicle driving in a straight lane.

3 Problem Formulation

Considering human-robot interactive autonomous system scenarios, the ego agent is commanded by planning and control modules, while the perception module detects surrounding agents around the ego agent. We formulate the multi-agent trajectory prediction problem as predicting future trajectories of all surrounding target agents given historical trajectories of surrounding agents within a certain range and future motion of the ego agent. The objective is to learn the posterior distribution $P(\mathcal{Y}|\mathcal{X}, \mathcal{L})$ of multiple targets' future trajectories $\mathcal{Y} = \{Y_i | v_i \in V_{tar}\}$, where V_{tar} is the set of predicted targets selected within an ego-vehicle-centered area $\mathcal{A}_{tar}$. The $\mathcal{X} = \{X_i | v_i \in V_{tar}\}$ denotes the past trajectories of the target agent. At any time t, the history trajectory and future trajectory of an agent i are denoted as $X_i = \{x_i^{t-T_{obs}+1}, x_i^{t-T_{obs}+2} ..., x_i^t\}$ and $Y_i = \{y_i^{t+1}, y_i^{t+2} ..., y_i^{t+T_{pred}}\}$, where the elements of $x_i, y_i \in \mathbb{R}^2$ represent coordinates in the past and future, respectively, while T_{obs} and T_{pred} refer to the number of frames for observation and prediction. We also consider the setting where we condition on the ego-agent's future motion plan, for example when evaluating responses to a set of motion primitives. In this setting, we additionally assume that we know the ego-agent's future motion plan for the next T timesteps, $\mathcal{L} = Y_{ego} = \{y_{ego}^{t+1}, y_{ego}^{t+2} ..., y_{ego}^{t+T_{pred}}\}$.

4 Motivation

In this work we are interested in developing a trajectory prediction method that (1) directly designed for real-world robotic systems and can be integrated with downstream modules such as decision making and motion planning; (2) reasons the target agent's intention changes over time to more accurately infer its future trajectory choices; and (3) models social interaction in greater detail, selectively distinguishing the varying degrees of influence that different agents have on the target agent's future trajectory.

By introducing the potential motion of the ego agent, this method can evaluate the future trajectories of surrounding target agents under different potential motions of the ego agent, thereby providing a basis for downstream modules of the robotic system. It is worth re-emphasizing that we merely require a coarse potential motion of the target agent, rather than precise and complete planned trajectory. Additionally, we find that incorporating the potential motion of the ego agent enables the modeling of social interactions between the

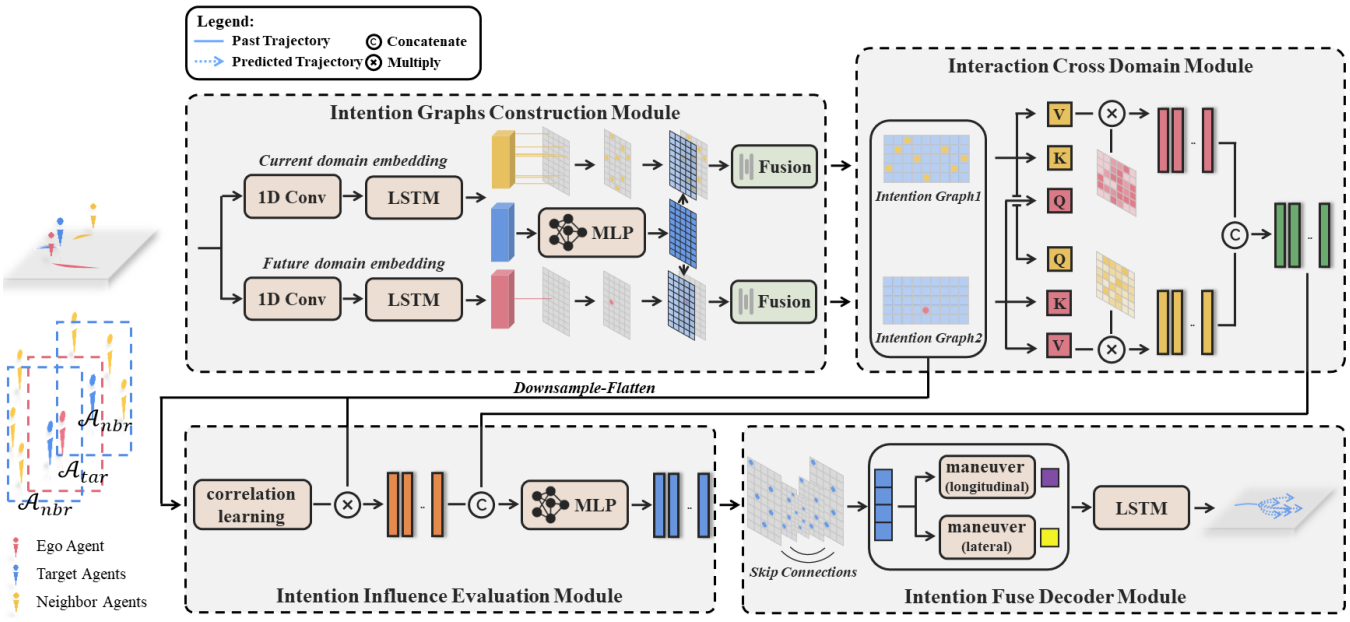

Figure 2: The overview of our proposed method: For each predicted target, two intention graphs are first constructed. Cross-domain interactions between these graphs are then modeled to refine behavior intention. Subsequent prioritized evaluation of other agents' influences further enhances intention estimation. Finally, multi-scale feature fusion and decoding generate diverse plausible trajectories.

ego agent and target agents in the future time domain. This enables us to jointly analyze intention information across both the current and future time domains, achieving fusion of feature information in spatial dimensions, temporal dimensions, and among different agents, thereby obtaining more accurate future intentions for the target agent. Finally, the model also learns the varying degrees of influence that different agents have on the target's future trajectory selection, enabling a hierarchical consideration of the interactions between the target agent and its surroundings, and achieving a more granular modeling process.

5 Proposed method

Our proposed method CiT is summarized in Figure 2, which processes each predicted agent in its own centric area $\mathcal{A}_{nbr}$. Thus, the area $\mathcal{A}_{nbr}$ contains three types of agents in total: the ego agent, the target agent, and the other neighboring agents. CiT contains four modules in total. Sec.5.1 proposes an Intention Graphs Construction Module to construct behavior intentions of the target agent in two time domains, Sec.5.2 proposes an Interaction Cross Domain Module to dynamically model social interactions across time domains, Sec.5.3 proposes an Intention Influence Evaluation Module to assess which agent's influence the target agent should focus more on, and Sec.5.4 proposes an Intention Fusion Decoder Module to fuse features and generate multiple possible future trajectories for each target agent.

5.1 Intention Graphs Construction

First, the trajectories X_{ego}, X_{tar}, X_{nbr} are encoded by the one-dimensional (1D)-CNN followed by the long short term memory (LSTM) model. Since the trajectories of target agent, neighbor agents and ego agent belong to different time domains, CNN-LSTMs with different weights are adopted. While 1D-CNN-LSTM can extract temporal properties, it can hardly capture the spatial interactions between target agent and its surrounding agents. Therefore, we build a grid centered at the location of the target agent and adopt the convolutional social pooling layer proposed by [7]. The ego agent encoding ξ_{ego} is placed to a social tensor in the target-centric grid with respect to its locations, then convolution and pooling layers are applied on the social tensor to capture social context for the spatial interaction between target agent and ego agent in the future time domain. A similar processing method is also adopted for the neighbor agents encoding ξ_{nbr} to capture the social context for the spatial interaction between the target agent and neighbor agents in the current time domain. The target encoding ξ_{tar} is concatenated with the two time domain social contexts to generate two Intention Graphs $\mathcal{V}^c, \mathcal{V}^f \in \mathbb{R}^{H\times W\times(D+1)}$, where H and W are the height and width of the target-centric grid $\mathcal{A}_{nbr}$ respectively, and D is the dimension after trajectory encoding.

$$\xi_{tar} = LSTM(1D-CNN(X_{tar})) \tag{1}$$

$$\xi_{ego} = LSTM(1D-CNN(X_{ego})) \tag{2}$$

$$\xi_{nbr} = LSTM(1D-CNN(X_{nbr})) \tag{3}$$

5.2 Interaction Cross Domain

Both Intention Graphs already contain the historical trajectory information of the target agent and social interaction information with surrounding agents. Therefore, each of the intention graph has a certain capability of estimating the future trajectory of the

target agent. However, taking the humans' theory of mind that they can navigate through various scenarios without collisions with surroundings as a reference, the future trajectory inferred unilaterally from intention in one time domain is inadequate. Because humans often comprehensively consider social behaviors in different time domains before selecting trajectories, which requires the model to jointly analyze behavior intentions in different time domains. Starting from the completion of Intention Graphs Construction Module, $\mathcal{V}^c$ is transformed to generate the current time domain intention matrix $\mathcal{M}_c = [\mathcal{V}_1^c, \mathcal{V}_2^c, ..., \mathcal{V}_{D+1}^c]$, with $V_d^c \in \mathbb{R}^{HW}$ and similarly for the future time domain intention matrix $\mathcal{M}_f$ is generated from $\mathcal{V}^f$, where $\mathcal{M}_c, \mathcal{M}_f \in \mathbb{R}^{HW\times(D+1)}$. The Intention Graph in the current time domain proposes Queries to the Intention Graph in the future time domain, and explores which interactions in the future time domain will affect its current Behavior Intention through the Key and Value of the future time domain Intention Graph, so as to revise its own Intention. The Intention Graph in the future time domain will also obtain an Intention that is more consistent with real human thoughts by proposing Queries to the Intention Graph in the current time domain. To conduct cross time domain intention analysis jointly, we propose the formulation of the cross time domain representation with:

$$\widetilde{\mathcal{M}}_c = \mathcal{F}^{trans}(\mathcal{M}_c, \mathcal{M}_f, \mathcal{M}_f) \tag{4}$$

$$\widetilde{\mathcal{M}}_f = \mathcal{F}^{trans}(\mathcal{M}_f, \mathcal{M}_c, \mathcal{M}_c) \tag{5}$$

$\widetilde{\mathcal{M}}_c, \widetilde{\mathcal{M}}_f \in \mathbb{R}^{HW\times D'}$, and the $\mathcal{F}^{trans}(Q, K, V)$ is defined as:

$$\mathcal{F}^{trans}(Q, K, V) = \mathcal{F}^{fc}(\mathcal{F}^{attn}(Q, K, V)) \tag{6}$$

Through the Interaction Cross Domain Module, the Intention Graphs in the two time domains are fully integrated with each other's information and adjust their own intentions accordingly. This allows the network to have the capability of dynamically modeling social interactions across time domains. We concatenate the $\widetilde{M}_c$ and $\widetilde{M}_f$ to obtain $\mathcal{I}$, a more precise representation of the target agent's intention which explores the correlation between social interaction in the two time domains.

5.3 Intention Influence Evaluation

The Intention Influence Evaluation Module learns the influence degree of social interactions in the two time domains on the target agent, which is a more delicate modeling of social interaction behaviors, allowing the network to consider social interactions in a prioritized manner. We flatten the downsampled Intention Graphs $\mathcal{V}^c$ and $\mathcal{V}^f$ and concatenate them with the encoding of the target agent to generate social contexts in the two time domains, which are then fed into fully connected and softmax layers to produce two weights β_1 and β_2 representing the importance of social information in each time domain respectively. The two weights are multiplied with the social contexts in the two time domains respectively to obtain vector $\mathcal{G}$. Finally, $\mathcal{G}$ is concatenated with the vector $\mathcal{I}$ obtained from the Interaction Cross Domain Module to generate the intention representation $\mathcal{Z}$ of the target agent.

Note that we multiply the weights of different time domains with the low-level Intention Graphs, rather than the deeper and more abstract Intention Graphs generated after Interaction Cross Domain. We want to make use of the shallow information contained in the low-level Intention Graphs through this method. At the same time, we want the model to remember to some extent the "original" intention generated from social interactions in its own time domain, preventing the model from over-focusing on social behaviors in the other time domain thus causing unnecessary intention correction. Before feeding it into the decoder, the Intention $\mathcal{Z}$ at this point already contains the social interaction information of the target agent with different agents across time domains and spaces. In addition, it incorporates both high-level and low-level intention features.

5.4 Intention Fusion Decoder

In order to predict future trajectories for all target agents around the ego agent, each target agent intention represented as $\mathcal{Z}$ is placed into an ego agent-centric grid based on its location, resulting in a social intention tensor $\mathcal{S}$. To further extract interaction features at different spatial scales, we apply a fulley convolutional network (FCN) structure on the social intention tensor $\mathcal{S}$ and obtain the intention feature $\mathcal{Z}^+$. We refer to the method of [7] and predefine six maneuver categories $C = \{c_k | k = 1, 2, ..., 6\}$ in advance to address the inherent multi-modality of human behavior by predicting the distribution for each of the maneuver classes along with the probability for each maneuver class. The predefined maneuver categories including three lateral maneuvers (lane keeping, left lane change, right lane change) and two longitudinal maneuvers (normal driving and braking). We feed each target agent's intention feature $\mathcal{Z}^+$ into fully connected layers followed by softmax layers that output the lateral and longitudinal maneuver probabilities. These can be multiplied to give the value of each maneuver $P(c_k|\mathcal{L}, X)$. The lateral and longitudinal maneuver class are transformed into one-hot vectors and concatenated with the intention feature $\mathcal{Z}^+$. Then the resulted feature vector is fed into LSTM layers to generate the predicted location, which can be represented by a bivariate Gaussian distribution over T_{pred} frames:

$$\hat{y}_i^{t+T_{pred}} \sim \mathcal{N}(\mu_i^{t+T_{pred}}, \sigma_i^{t+T_{pred}}, \rho_i^{t+T_{pred}}) \tag{7}$$

where the mean vector is the sum of all displacements along the future frames T_{pred} with the location at the last frame t, the standard deviation vector $\sigma_i^{t+T_{pred}} \in \mathbb{R}^2$ and the correlation coefficient $\rho_i^{t+T_{pred}} \in \mathbb{R}$. Finally, the posterior probability of all surrounding target agents' future trajectories could be estimated from:

$$P(\mathcal{Y}|\mathcal{X}, \mathcal{L}) = \prod_{v_i \in V_{tar}} \sum_{k=1}^{|C|} P_{\theta_i}(Y_i|c_k, \mathcal{X}, \mathcal{L}) P(c_k|\mathcal{X}, \mathcal{L}) \tag{8}$$

where the Gaussian parameters for all future frames of target agent v_i is written as θ_i.

6 Experiments

Following existing methods[1, 7, 9, 12, 28, 32, 33, 40], we evaluate our proposed method on two public-available datasets NGSIM and HighD. For each dataset, 70% is split for training and 10%, 20% for validation and test. The objective is to predict the future 5s (25 frames) behavioral trajectories for all target agents surrounding

the ego agent using the 3s (15 frames) past trajectory and the ego agent's potential motion. In addition, to demonstrate our proposed method only requires the rough potential motion of the ego agent, ego agent's future trajectory is downsampled to 1Hz (5 frames) in the testing and evaluation.

6.1 Implementation Details

Each data instance contains an agent specified as the ego agent. The target agents whose trajectories we want to predict are all agents located in the ego-agent-centric area $\mathcal{A}_{tar}$. The area $\mathcal{A}_{tar}$ has a size of 200×35 feet, discretized into a 25×5 spatial grid. For each target agent, the operating area $\mathcal{A}_{nbr}$ has the same size as $\mathcal{A}_{tar}$. Ideally, we want to minimize the negative log-likelihood over all training instances.

$$-log(\sum_{v_i \in V_{tar}} P_\theta(Y_i|c_k, \mathcal{X}, \mathcal{L})P(c_k|\mathcal{X}, \mathcal{L})) \tag{9}$$

However, each training instance only corresponds to the one true maneuver class that was actually performed. Thus we minimize the negative log likelihood of the predictive distribution associated with the true maneuver.

$$-\sum_{v_i \in V_{tar}} log(P_\theta(Y_i|c_{true}, \mathcal{X}, \mathcal{L})P(c_{true}|\mathcal{X}, \mathcal{L})) \tag{10}$$

For the incorporation of the ego agent's potential motion $\mathcal{L}$, we use the ego agent's actual future trajectory over the prediction horizon 5s (25 frames) as input during training. In evaluation and testing, we downsample the ego agent's actual trajectory to 5 frames to verify our model only requires the ego agent's "rough intention". The maintained validity of our model when only "rough intention" is available illustrates the high robustness and convenience of real-world deployment of our proposed method. Meanwhile, the introduction of the ego agent's potential motion in our model can effectively combine with downstream trajectory planning and control tasks, having high application value.

All the experiments were conducted on NVIDIA RTX 4090(24GB). We employed 1D-conv to upsample the input of the model from dimension 2 to 32. The encoder LSTM has 64 dimensional state while the decoder has a 128 dimensional state. In these experiments, all parameter settings are aligned with the compared methods. We use the leaky-ReLU activation with $\alpha = 0.1$ for all layers.

6.2 Comparison with SOTA Methods

We measure the performance of different trajectory prediction methods using two metrics: RMSE and NLL. 1) RMSE calculates the deviation between predicted trajectories and ground truth future trajectories. For multiple predicted trajectories, we evaluate RMSE using the trajectory under the maneuver with highest probability. 2) The negative log-likelihood (NLL) of true trajectories under the predictive distribution fitted by the model is also adopted in evaluation, because RMSE tends to average all prediction results and has limitations for multi-modal trajectory prediction.

NGSIM dataset. We compare our method with the state-of-the-art prediction methods at different timestamps; see Table 2. The method we proposed has achieved either the first or the second place in the metrics at all timestamps, and it outperforms the current state-of-the-art (SOTA) methods in terms of the average metrics of RMSR and NLL. We have reduced the RMSE and NLL from 1.74/3.86 to 1.67/3.66. We have achieved improvements of **4.02%** and **5.18%** in the RMSE and NLL metrics respectively.

HighD dataset. We compare our method with the state-of-the-art prediction methods at different timestamps; see Table 3. The method we proposed has achieved either the first or the second place in the metrics at all timestamps, and it outperforms the current state-of-the-art (SOTA) methods in terms of the average metrics of RMSR and NLL. We have reduced the RMSE and NLL from 1.14/3.04 to 1.06/2.89. We have achieved improvements of **7.02%** and **4.93%** in the RMSE and NLL metrics respectively.

6.3 Qualitative Results

Visualization of predicted trajectory. Figure 3 compares the predicted trajectories of the previous conditional trajectory prediction state-of-the-art method, PiP(red), our CiT(yellow), historical trajectories(blue) and the ground-truth(green). In the first row, it is shown that although the previous state-of-the-art methods can correctly capture the motivation of the target agent, that is, the trend of the motion direction is correct, our prediction results are closer to the ground truth. In the second row, we can observe that the previous methods fail to correctly capture the motivation of the target agent. For example, in the first column, the historical trajectory of the target agent is changing lanes to the left and will continue to move leftward. However, PiP predicts that it will move straight after changing lanes to the left. In practical application scenarios, such misjudgment of the motion motivation is extremely dangerous. From this, we can see that by mutually correcting the intentions across the time domain, not only can more accurate trajectory information be obtained, but it is even possible to correct the originally mispredicted motivations.

Visualization of multi-modal prediction. Figure 4 shows the multiple future trajectories predicted by our proposed method based on predefined maneuvers. The figure 4 contains multiple predicted future trajectories under the six pre-defined motivations in Sec.5.4. The yellow trajectory represents the future trajectory predicted under the motivation with the highest probability by the method, which is the prediction result output by the method. The purple trajectories are the future trajectories predicted under other motivations with relatively lower probabilities. We can see that the method can capture agent maneuvers like lane changing(first column), braking(third column) and normal driving(fourth column). In some challenging scenarios, for example, in the first column, the historical trajectory of the agent shows a lane change to the left, and the agent will continue to move leftward for a certain distance before driving straight. It is relatively difficult for the method to predict the straight trajectory after the lane change to the left. In the second column, the entire historical trajectory of the agent is changing lanes to the right. However, precisely within the prediction time window, the agent has completed the lane change and maintains a straight driving motion. In such a situation, the model often predicts that the agent will continue to change lanes to the right based on the information of the rightward lane change in the historical trajectory. Nevertheless, as can be seen from the figure

Time	S-LSTM [1]	C-LSTM [7]	MATF-GAN [40]	SAMMP [21]	PiP [28]	CF-LSTM [37]	Flash [2]	WSiP [32]	BAT [17]	C2F-TP [34]	Ours
1s	0.59/2.10	0.58/1.96	0.66	0.51	0.55/1.72	0.55	0.51	0.56/1.77	**0.23**	0.32	0.43/**1.61**
2s	1.29/3.66	1.27/3.46	1.34	1.13	1.18/3.30	1.10	1.15	1.23/3.30	**0.81**	0.92	0.88/**3.14**
3s	2.13/4.61	2.11/4.32	2.08	1.88	1.94/4.17	1.78	1.84	2.05/4.17	1.54	1.62	**1.47**/**3.95**
4s	3.21/5.37	3.19/4.95	2.97	2.81	2.88/4.80	2.73	2.64	3.08/4.80	2.52	2.44	**2.36**/**4.54**
5s	4.55/5.99	4.53/5.48	4.13	3.67	4.04/5.32	3.82	3.62	4.34/5.32	3.62	3.45	**3.23**/**5.04**
avg	2.35/4.35	2.34/4.03	2.24	2.00	2.12/3.86	2.00	1.95	2.25/3.87	1.74	1.75	**1.67**/**3.66**

Table 2: Comparison with baseline models on NGSIM dataset. RMSE/NLL are reported. Lower is better. Bold/underlined fonts represent the best/second-best result. Our method achieves the best performance in RMSE/NLL.

Time	S-LSTM [1]	C-LSTM [7]	S-GAN [12]	N-LSTM [22]	PiP [28]	M-LSTM [23]	DLM [15]	DRBP [9]	WSiP [32]	C2F-TP [34]	Ours
1s	0.21/0.46	0.24/0.43	0.30	0.20	0.17/**0.14**	0.19	0.22	0.41	0.20/0.31	**0.11**	0.12/0.15
2s	0.65/2.55	0.68/2.54	0.78	0.57	0.52/2.24	0.55	0.61	0.79	0.60/2.31	**0.41**	**0.41**/**2.14**
3s	1.31/3.81	1.26/3.72	1.46	1.14	1.05/3.48	1.10	1.16	1.11	1.21/3.51	0.92	**0.85**/**3.31**
4s	2.16/4.67	2.15/4.51	2.34	1.90	1.76/4.33	1.84	1.80	**1.40**	2.07/4.32	1.64	1.49/**4.11**
5s	3.29/5.35	3.31/5.13	3.41	2.91	2.63/4.99	2.78	2.80	-	3.14/4.95	2.60	**2.43**/**4.74**
avg	1.52/3.37	1.53/3.27	1.66	1.34	1.23/3.04	1.29	1.32	-	1.44/3.08	1.14	**1.06**/**2.89**

Table 3: Comparison with baseline models on HighD dataset. RMSE/NLL are reported. Lower is better. Bold/underlined fonts represent the best/second-best result. Our method achieves the best or second-best performance in RMSE/NLL.

Method	info (c)	info (f)	ICD	IIE	Fusion	1s	2s	3s	4s	5s	avg
Variant1						0.68/2.14	1.66/3.81	2.96/4.76	4.56/5.42	5.44/6.03	3.06/4.43
Variant2	✓				✓	0.51/**1.60**	1.09/3.19	1.79/3.98	2.69/4.62	3.83/5.15	1.98/3.71
Variant3		✓			✓	0.54/1.70	1.32/3.27	2.31/4.21	3.51/4.85	4.90/5.35	2.52/3.88
Variant4	✓	✓			✓	0.53/1.83	1.13/3.34	1.81/4.17	2.64/4.76	3.66/5.25	1.95/3.87
Variant5	✓	✓	✓		✓	0.46/1.64	**0.88**/3.16	1.56/3.98	2.50/4.56	3.41/5.06	1.76/3.68
Ours	✓	✓	✓	✓	✓	**0.43**/1.61	**0.88**/**3.14**	**1.47**/**3.95**	**2.36**/**4.54**	**3.23**/**5.04**	**1.67**/**3.66**

Table 4: Ablation study conducted on NGSIM dataset. info(c) refers to the trajectory information of the surrounding agents in the current time domain, info(f) refers to the potential motion of the ego agent in the future time domain. ICD stands for Interaction Cross Domain Module, IIE stands for Intention Influence Evaluation, and Fusion means performing feature fusion through the Fully Connected Network (FCN) structure.

Method	info (c)	info (f)	ICD	IIE	Fusion	1s	2s	3s	4s	5s	avg
Variant1						0.22/0.55	0.65/2.65	1.32/3.94	2.22/4.87	3.43/5.59	1.57/3.52
Variant2	✓				✓	0.20/0.28	0.62/2.27	1.30/3.47	2.23/4.31	3.42/4.96	1.55/3.06
Variant3		✓			✓	0.21/0.34	0.65/2.34	1.33/3.52	2.23/4.33	3.35/4.95	1.55/3.10
Variant4	✓	✓			✓	0.22/0.20	0.61/2.31	1.19/3.44	1.96/4.28	2.95/4.88	1.39/3.02
Variant5	✓	✓	✓		✓	0.17/**0.14**	0.46/2.27	0.96/3.42	1.50/4.24	2.55/4.86	1.13/2.99
Ours	✓	✓	✓	✓	✓	**0.12**/0.15	**0.41**/**2.14**	**0.85**/**3.31**	**1.49**/**4.11**	**2.43**/**4.74**	**1.06**/**2.89**

Table 5: Ablation study conducted on HighD dataset. info(c) refers to the trajectory information of the surrounding agents in the current time domain, info(f) refers to the potential motion of the ego agent in the future time domain. ICD stands for Interaction Cross Domain Module, IIE stands for Intention Influence Evaluation, and Fusion means performing feature fusion through the Fully Connected Network (FCN) structure.

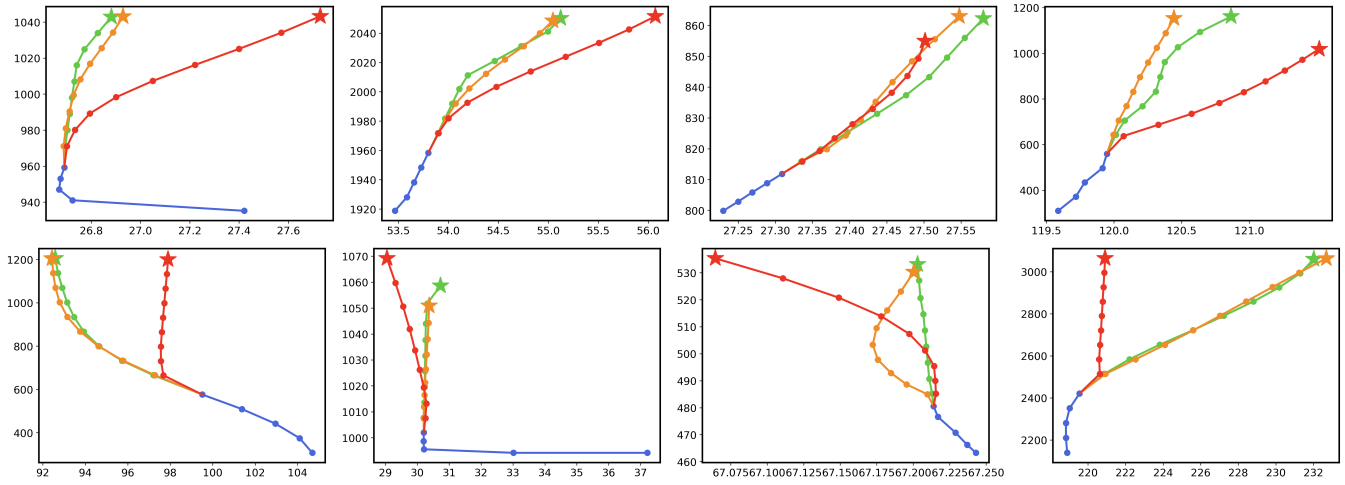

Figure 3: Visualization comparison on NGSIM and HighD. Given the past trajectories(blue), we illustrate the ground truth trajectories(green) and predicted trajectories by CiT(yellow) and PiP(red). Best viewed in color and zoom-in for more clarity.

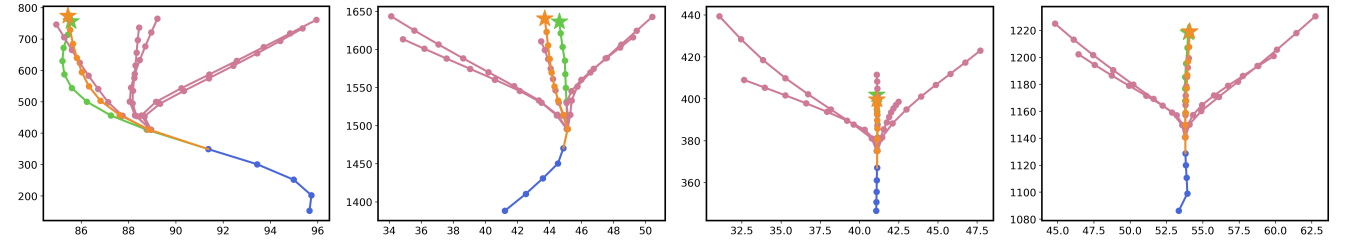

Figure 4: Visualization multi-modal prediction on NGSIM. Given the past trajectory(blue), we illustrate the ground truth trajectory(green), the predicted prediction(yellow), which corresponds to the maneuver with highest probability among all predicted trajectories, and the other predictions(purple). Best viewed in color and zoom-in for more clarity.

4, the method we proposed can correctly capture the information that the agent has completed the lane change and will maintain a straight driving motion.

6.4 Ablation Studies

Tables 4 and 5 show the contribution of each component added to model over the baseline. Variant1 is the baseline, which only uses LSTM to predict future trajectories based on past trajectories. Variant2 takes surrounding agents' state as conditional input, and uses fully convolutional networks(FCN). Variant3 takes ego agent's potential motion as conditional input and also utilizes the FCN structure. To demonstrate that the performance improvement achieved by the ICD module stems from the mutual complementarity of semantic intentions across time domains rather than simply increasing the parameters, we introduce Variant4, which employs self-attention mechanism for information fusion and shares almost the same number of parameters as the ICD module. Variant5 is based on Variant4, replacing the self-attention method to integrate information with the Interaction Cross Domain Module(ICD). Our method adds the Intention Influence Evaluation Module(IIE) on Variant5. The experimental results show that performance improves as we add the proposed modules. Compared to Variant2 and Variant3 which model social interactions unilaterally from either the current or future time domain, Variant4 achieves better results by integrating information from the two time domains. Moreover, Variant5 further improves over Variant4 by using social information from the other time domain to correct the intention in its own time domain, and conducts joint analysis of social interactions across time domains. It can be observed that prediction accuracy is improved with the IIE module. This indicates the model can differentiate the influence across time domains to refine social interaction modeling, and incorporating more global, shallow intention information can further enhance model performance.

7 Conclusion

This paper presents a novel trajectory prediction method that models cross-time-domain social interactions to capture temporal intention dynamics. Our approach incorporates four key components: potential motion integration, intention graph construction, cross-domain interaction modeling, and intention influence assessment. These designs enable inter-temporal information exchange and intention refinement, while the integration of ego-agent potential motion facilitates seamless integration with downstream planning and control modules in robotic systems.

Acknowledgements

The study is supported by the National Key R&D Program of China (No. 2023YFB4301900).

References

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

[2] Morris Antonello, Mihai Dobre, Stefano V Albrecht, John Redford, and Subramanian Ramamoorthy. 2022. Flash: Fast and light motion prediction for autonomous driving with Bayesian inverse planning and learned motion profiles. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 9829–9836.

[3] Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. 2024. SingularTrajectory: Universal Trajectory Predictor Using Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17890–17901.

[4] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. 2019. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449* (2019).

[5] Xiaobo Chen, Huanjia Zhang, Feng Zhao, Yu Hu, Chenkai Tan, and Jian Yang. 2022. Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems* 23, 10 (2022), 19471–19483.

[6] Yuxiao Chen, Boris Ivanovic, and Marco Pavone. 2022. Scept: Scene-consistent, policy-based trajectory predictions for planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17103–17112.

[7] Nachiket Deo and Mohan M Trivedi. 2018. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1468–1476.

[8] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. 2020. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11525–11533.

[9] Kai Gao, Xunhao Li, Bin Chen, Lin Hu, Jian Liu, Ronghua Du, and Yongfu Li. 2023. Dual Transformer Based Prediction for Lane Change Intentions and Trajectories in Mixed Traffic Environment. *IEEE Transactions on Intelligent Transportation Systems* (2023).

[10] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. 2022. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17113–17122.

[11] Ke Guo, Wenxi Liu, and Jia Pan. 2022. End-to-end trajectory distribution prediction based on occupancy grid maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2242–2251.

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

[13] Dirk Helbing and Peter Molnar. 1995. Social force model for pedestrian dynamics. *Physical review E* 51, 5 (1995), 4282.

[14] Xin Huang, Guy Rosman, Ashkan Jasour, Stephen G McGill, John J Leonard, and Brian C Williams. 2022. TIP: Task-informed motion prediction for intelligent vehicles. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11432–11439.

[15] Mahrokh Khakzar, Andry Rakotonirainy, Andy Bond, and Sepehr G Dehkordi. 2020. A dual learning model for vehicle trajectory prediction. *IEEE Access* 8 (2020), 21897–21908.

[16] Mihee Lee, Samuel S Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. 2022. Muse-VAE: multi-scale VAE for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2221–2230.

[17] Haicheng Liao, Zhenning Li, Huanming Shen, Wenxuan Zeng, Dongping Liao, Guofa Li, and Chengzhong Xu. 2024. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10332–10340.

[18] Jerry Liu, Wenyuan Zeng, Raquel Urtasun, and Ersin Yumer. 2021. Deep structured reactive planning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4897–4904.

[19] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. 2024. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2039–2049.

[20] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. 2023. Leapfrog Diffusion Model for Stochastic Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5517–5526.

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

[22] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. 2019. Non-local social pooling for vehicle trajectory prediction. In *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 975–980.

[23] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. 2020. Attention based vehicle trajectory prediction. *IEEE Transactions on Intelligent Vehicles* 6, 1 (2020), 175–185.

[24] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. 2021. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations*.

[25] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. 2020. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14074–14083.

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

[27] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 683–700.

[28] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. 2020. Pip: Planning-informed trajectory prediction for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 598–614.

[29] Pinhao Song, Pengteng Li, Erwin Aertbeliën, and Renaud Detry. 2024. Robot Trajectron: Trajectory Prediction-based Shared Control for Robot Manipulation. *arXiv preprint arXiv:2402.02499* (2024).

[30] Qiao Sun, Xin Huang, Junru Gu, Brian C Williams, and Hang Zhao. 2022. M2i: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6543–6552.

[31] Royden Wagner, Omer Sahin Tas, Marvin Klemp, and Carlos Fernandez Lopez. 2023. Road Barlow Twins: Redundancy Reduction for Road Environment Descriptors and Motion Prediction. *arXiv preprint arXiv:2306.10840* (2023).

[32] Renzhi Wang, Senzhang Wang, Hao Yan, and Xiang Wang. 2023. Wsip: Wave superposition inspired pooling for dynamic interactions-aware trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4685–4692.

[33] Yu Wang, Shengjie Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. 2020. Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion. *IEEE Transactions on Intelligent Transportation Systems* 23, 1 (2020), 236–248.

[34] Zichen Wang, Hao Miao, Senzhang Wang, Renzhi Wang, Jianxin Wang, and Jian Zhang. 2024. C2F-TP: A Coarse-to-Fine Denoising Framework for Uncertainty-Aware Trajectory Prediction. *arXiv preprint arXiv:2412.13231* (2024).

[35] Theodor Westny, Björn Olofsson, and Erik Frisk. 2024. Diffusion-based environment-aware trajectory prediction. *arXiv preprint arXiv:2403.11643* (2024).

[36] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. 2024. SocialCircle: Learning the Angle-based Social Interaction Representation for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19005–19015.

[37] Xu Xie, Chi Zhang, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. 2021. Congestion-aware multi-agent trajectory prediction for collision avoidance. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 13693–13700.

[38] Zhijie Yan, Pengfei Li, Zheng Fu, Shaocong Xu, Yongliang Shi, Xiaoxue Chen, Yuhang Zheng, Yang Li, Tianyu Liu, Chuxuan Li, et al. 2023. Int2: Interactive trajectory prediction at intersections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8536–8547.

[39] Liang Zhang, Nathaniel Xu, Pengfei Yang, Gaojie Jin, Cheng-Chao Huang, and Lijun Zhang. 2023. TrajPAC: Towards Robustness Verification of Pedestrian Trajectory Prediction Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8327–8339.

[40] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. 2019. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12126–12134.

[41] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. 2023. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17863–17873.