Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

Beyond Patterns: Harnessing Causal Logic for Autonomous Driving Trajectory Prediction

Bonan Wang[1], **Haicheng Liao**[1], **Chengyue Wang**[1], **Bin Rao**[1], **Yanchen Guan**[1], **Guyang Yu**[1], **Jiaxun Zhang**[1], **Songning Lai**[2], **Chengzhong Xu**[1], **Zhenning Li**[1†]
[1]University of Macau
[2]The Hong Kong University of Science and Technology (Guangzhou)
{mc35002, yc27979, yc47938, binrao, yc37976, yc47415, czxu, zhenningli}@um.edu.mo, yuguyangsam@gmail.com, songninglai@hkust-gz.edu.cn

Abstract

Accurate trajectory prediction has long been a major challenge for autonomous driving (AD). Traditional data-driven models predominantly rely on statistical correlations, often overlooking the causal relationships that govern traffic behavior. In this paper, we introduce a novel trajectory prediction framework that leverages causal inference to enhance predictive robustness, generalization, and accuracy. By decomposing the environment into spatial and temporal components, our approach identifies and mitigates spurious correlations, uncovering genuine causal relationships. We also employ a progressive fusion strategy to integrate multimodal information, simulating human-like reasoning processes and enabling real-time inference. Evaluations on five real-world datasets—ApolloScape, nuScenes, NGSIM, HighD, and MoCAD—demonstrate our model's superiority over existing state-of-the-art (SOTA) methods, with improvements in key metrics such as RMSE and FDE. Our findings highlight the potential of causal reasoning to transform trajectory prediction, paving the way for robust AD systems.

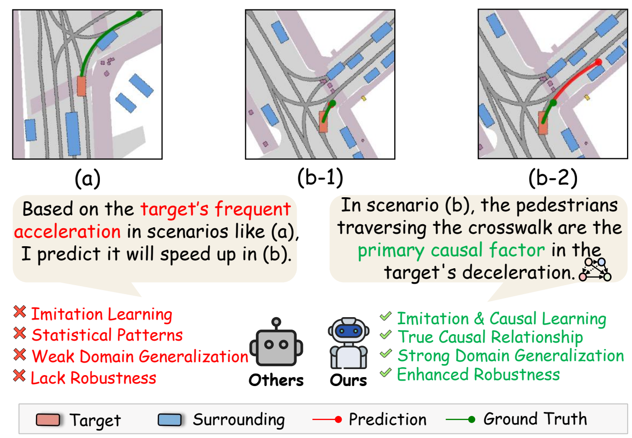

Figure 1: Illustration of causal relationships in traffic scenarios. Panel (a) represents the training stage, where the target agent is frequently observed accelerating through crosswalks. In the test stage, a traditional data-driven model predicts that the target agent will similarly accelerate, as shown in (b-1). In contrast, our model employs a causal inference to discern the true causal relationship, enabling the accurate prediction of the target agent's behavior of stopping at the crosswalk, as demonstrated in (b-2).

1 Introduction

Autonomous driving (AD) technology holds promise for revolutionizing transportation by enhancing traffic efficiency and safety [Liao *et al.*, 2025c; Guan *et al.*, 2024]. A critical component of AD systems is trajectory prediction, which involves forecasting the future positions of vehicles and other traffic participants. Accurate trajectory prediction enables these systems to anticipate and respond to dynamic changes in the environment, providing essential inputs for decision-making, route planning, and collision avoidance [Liao *et al.*, 2024b]. Despite significant advancements in trajectory prediction for autonomous vehicles (AVs), modern data-driven models primarily focus on identifying **statistical patterns** within large datasets [Chen *et al.*, 2021]. While this approach has been effective in capturing correlations, it often overlooks the deeper causal relationships that underpin driving behaviors. These models [Liao *et al.*, 2024c] tend to treat driving environments as mere data-generating processes without recognizing the causative factors that influence vehicle dynamics.

As illustrated in Figure 1, the model may learn that vehicles tend to accelerate in particular scenarios, such as scenario (a). This insight allows the model to anticipate acceleration maneuvers at analogous points in the future, as demonstrated in scenario (b). However, such predictions are based solely on observed patterns and lack an understanding of the underlying causes, such as the presence of crosswalks or sharp curves in the road. This superficial reliance on data correlations results in models that are tethered to the training data, without the capability to infer causal links, such as the impact of pedestrian movement (temporal agent data) or changes in road layout (spatial map data). These limitations become particularly problematic when the model encounters scenarios that are not well-represented in the training data. Without an understanding of causal mechanisms, the model struggles to generalize to new situations, potentially leading to inaccurate predictions and safety risks [Chen *et al.*, 2021;

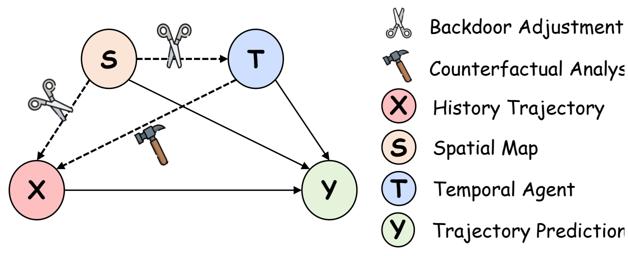

Figure 2: Illustration of the proposed new causal paradigm utilizes backdoor inference and counterfactual analysis to mitigate the confounding effects of spatial map S and temporal agent T data.

Liao *et al.*, 2025b; Ge *et al.*, 2023]. This highlights the need for trajectory prediction methods that incorporate causal reasoning to better interpret and predict complex traffic behaviors. This observation highlights the need for a paradigm shift toward **causal inference** in trajectory prediction, emphasizing the understanding of cause-and-effect relationships within data. Such an approach is crucial for AD systems for several reasons. First, causal inference enhances the robustness of the model [Bagi *et al.*, 2023]. AVs operate in dynamic scenes and may encounter sensor noise, missing data, or abnormal driving behaviors. By identifying underlying causal mechanisms, causal inference can effectively filter out these noises or random perturbations, maintaining the stability and accuracy of predictions. Second, causal inference improves generalization to unseen scenarios, which is essential for reliable performance in diverse environments [Ge *et al.*, 2023; Chen *et al.*, 2021]. Finally, causal models offer increased interpretability by providing insights into why specific predictions are made, thereby enhancing transparency and trust.

To address these needs, we propose a novel causal inference framework for trajectory prediction in AD. Causal inference is a method that goes beyond merely identifying correlations in data to uncover the actual cause-and-effect relationships that drive the behaviors being observed [Pearl, 2009; Kuang *et al.*, 2020]. This is particularly important in dynamic environments like traffic scenarios, where understanding the underlying causes of actions, rather than just patterns, leads to more accurate trajectory predictions for AVs. Our approach constructs a causal graph to explicitly represent the relationships between key variables, such as spatial map data and temporal agent data. As illustrated in Figure 2, we employ a causal inference paradigm, including backdoor adjustment and counterfactual analysis, to isolate the confounding variables in the traffic environment. These paradigms eliminate spurious correlations, uncovering the true causal relationships in the data. Furthermore, we introduce a cross-modal progressive fusion strategy that mimics the gradual reasoning process of human drivers. This strategy incorporates a multi-stage attention mechanism to generate progressive causal queries to improve prediction accuracy.

In summary, our contributions are as follows:

- We propose a novel **causal inference** paradigm for trajectory prediction in AD, utilizing a causal graph to model cause-and-effect relationships between spatial map data and temporal agent data. This approach enables the identification of genuine causal links, improving prediction accuracy and robustness.
- We introduce a **cross-modal progressive fusion** strategy, using a multi-stage attention mechanism to integrate multimodal information. This enhances real-time inference, allowing the model to adapt to dynamic scenes.
- Our extensive experiments across five real-world datasets consistently demonstrate that both our causal inference model and its plug-in module outperform existing methods in diverse traffic scenarios, highlighting their generalization capabilities and reliability.

2 Related Work

2.1 Trajectory Prediction for Autonomous Vehicles

Trajectory prediction has undergone a remarkable evolution, transitioning from traditional physics-based models and classical machine learning approaches to sophisticated deep learning architectures. Early methods, such as the Kinematic bicycle model [Wang *et al.*, 2025a] and Kalman filtering [Wang *et al.*, 2023] are often limited in their adaptability to complex scenarios. As a result, these methods are typically restricted to simpler environments and short-term prediction tasks. The field has seen significant progress with the rapid advancement of deep learning technologies. Recent studies have introduced sophisticated models like gated recurrent unit (GRU) [Li *et al.*, 2024; Liao *et al.*, 2024f], long short-term memory (LSTM) [Deo and Trivedi, 2018; Xie *et al.*, 2021], graph attention network (GAT) [Mo *et al.*, 2022; Liao *et al.*, 2024b], generative models [Liao *et al.*, 2024a], and the Transformers [Liu *et al.*, 2024; Chen *et al.*, 2024; Liao *et al.*, 2025a; Liao *et al.*, 2024e], expanding the potential for accurate long-term prediction in complex scenarios. While contemporary trajectory prediction models have demonstrated remarkable performance, their reliance on purely data-driven approaches presents fundamental limitations. These methods excel at identifying statistical regularities in training data but fail to capture the essential causal relationships that dictate agent behaviors in dynamic scenes.

2.2 Causal Inference in Autonomous Driving

Causal inference is a statistical framework designed to estimate the impact of one variable on another by identifying causal effects and mitigating the influence of confounding factors. In this context, two fundamental methodologies have emerged as particularly powerful for causal identification: backdoor adjustment, which enables unbiased effect estimation through proper conditioning on observed covariates, and counterfactual analysis, which examines hypothetical scenarios under different intervention conditions. The former [Pearl, 2009] involves identifying and controlling for confounding variables to reduce bias, while the latter [Kuang *et al.*, 2020] evaluates causal effects by constructing hypothetical scenarios to explore "what if" questions. In the field of trajectory prediction, previous studies employing causal inference have respectively utilized counterfactual analysis [Chen *et al.*, 2021] and backdoor adjustment [Ge

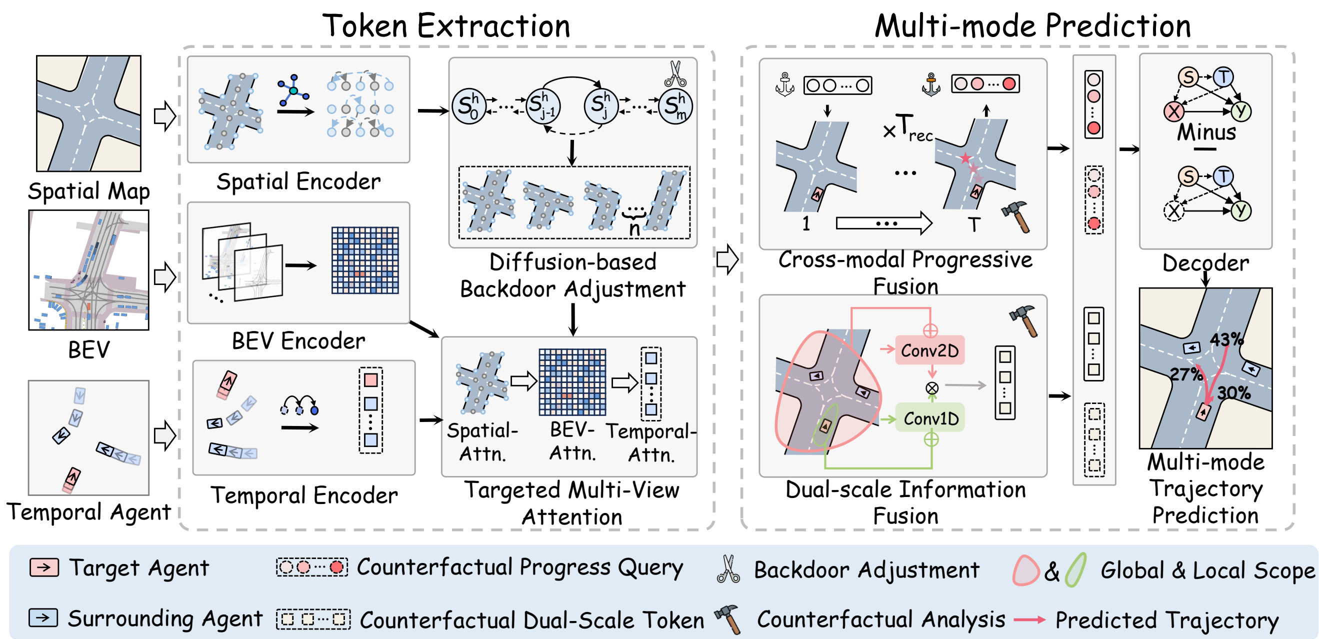

Figure 3: Overall framework of our two-stage model. The first stage involves token extraction with spatial, BEV, and temporal encoders producing tokens $\{S^h, B^h, T^h, X^h\}$. Spatial token S^h undergoes diffusion-based backdoor adjustment, generating $S^{h,i}$. Combining $S^{h,i}$ with $\{B^h, T^h, X^h\}$ via multi-view attention yields X^i_{attn}. In the second stage, X^i_{attn}, initial query Q_0, and counterfactual token X_c undergo cross-modal progressive fusion, producing final query Q^i and counterfactual query Q^i_c. In parallel, X^i_{attn} and X_c undergo dual-scale fusion to form G^i and G^i_c. The Causal Decoder then synthesizes the multi-modal predictions.

et al., 2023] to alleviate the impact of environmental confounders in traffic scenarios. However, due to an incomplete decomposition of these environmental factors, these methods often fall short in uncovering the underlying causal relationships. To address this limitation, our work advances beyond this limitation through a novel decomposition that separately models spatial and temporal environmental factors, enabling more precise confounding control. By unifying backdoor adjustment with counterfactual reasoning, we establish a more complete causal representation that preserves genuine relationships while eliminating spurious correlations, ultimately achieving superior robustness compared to existing methods.

3 Methodology

3.1 Problem Formulation

At each time step t, our model utilizes historical observations from $t - t_h$ to t, denoted as $X_0^{t-t_h:t}$, along with comprehensive traffic information–spatial map data S, temporal agent data T and a BEV of the traffic scene to forecast the vehicle's trajectory over the next t_f time steps, represented as $Y = \{X_0^{t+1:t+t_f}\}$. Specifically, S encompasses lane identifiers and precise coordinates of lane markings, while T includes the historical and predicted states of n surrounding agents, represented as $T = \{X_{1:n}^{t-t_h:t+t_f}, p_{1:n}^{t:t+t_f}\}$.

3.2 A Causal View on Trajectory Prediction

Causal Graph. This study constructs a causal graph to effectively map the relationships among the historical observations of the target agent X, its future trajectory Y, the temporal agent data T, and the spatial map data S. As shown in Figure 1 (c), X and T influence Y, represented by the edges $X \to Y$ and $T \to Y$. Additionally, S impacts both the historical and future trajectories, creating backdoor paths $Y \leftarrow S \to X$ and $Y \leftarrow S \to T$. In causal inference, variables like S and T that influence multiple other variables are termed confounders. The presence of confounders, particularly S, can bias the model's ability to learn the distributional properties of X and T, skewing it towards more common scenarios. For example, if straight-line trajectories dominate training data, the model may neglect specific traffic features, leading to incorrect predictions for turns. The confounding effect of T further hampers the model's ability to assess the impacts of T and X on Y, causing overfitting in interactions.

Casual Inference. Given that spatial map data S is relatively fixed and observable, backdoor adjustment mitigates bias by categorizing confounding factors into distinct groups and making predictions for each group. Therefore, it is necessary to systematically enumerate S and eliminate interference in the pathways $X \leftarrow S \to Y$ and $T \leftarrow S \to Y$. This enables the model to capture genuine causal relationships.

$$\tilde{Y} = \sum_{i=1}^{n} g_\theta\left(X, S = s_i, T\right) P\left(s_i\right) \tag{1}$$

where g_θ is the trajectory model, s_i denotes the enumerated spatial map data, and $P(s_i)$ is set to $1/n$ based on the principle of maximum entropy, with n being the assumed number of environmental categories [Ge *et al.*, 2023]. Although backdoor adjustment eliminates the confounding effects of spatial map data S, it does not address the complexities introduced by temporal agent data T. Causal theory suggests that counterfactual analysis can mitigate these effects by intervening in the historical trajectory data X to isolate the impact of T. This involves using the $do(\cdot)$ operator to replace the factual trajectory X with a hypothetical one, fixing X to a specific value and isolating it from its antecedents. Formally,

$$\tilde{Y}_c = \sum_{i=1}^{n} g_\theta \left(do\left(X = X_c\right), S = s_i, T \right) P(s_i) \tag{2}$$

where X_c represents counterfactual values. We combine backdoor adjustment and counterfactual analysis together in a comprehensive manner to derive: $Y = \tilde{Y} - \tilde{Y}_c$.

3.3 Proposed Model

Figure 3 showcases the overall pipeline of our model—token extraction, and multi-mode prediction—into a cohesive system. This subsection outlines the structure of our model.

Token Extraction

From a causal inference perspective, predicting future trajectories requires isolating the factors influencing the target object's decisions and movements. Due to the heterogeneity of traffic information, we employ specialized encoders–Spatial Encoder, Temporal Encoder, and BEV Encoder–to extract comprehensive features from diverse data sources.

Spatial Encoder. In this encoder, the spatial map data $S \in \mathbb{R}^{N_m \times n \times W_m}$ is tokenized into a sequence using the GRU layers. Then, GAT is applied to extract the spatial tokens $S^h \in \mathbb{R}^{N_m \times D}$ for the following backdoor adjustment process. Here, N_m represents the number of map polylines, n signifies the number of points within each polyline. w_m indicates the number of attributes for each point, such as location and road type, D denotes the dimension after encoding.

Temporal Encoder. The GRU layers are used in this encoder to produce target $X^h \in \mathbb{R}^D$ and surrounding $T^h \in \mathbb{R}^{N_a \times D}$ tokens for the raw observations $X \in \mathbb{R}^{T \times W_a}$ and temporal agent data $T \in \mathbb{R}^{N_a \times T \times W_a}$. Here, T represents the number of history frames, W_a is the number of state information, N_a signifies the number of surrounding agents.

BEV Encoder. Recent studies have shown that explicitly modelling the heterogeneity of driving scenes by incorporating diverse modalities can substantially improve a model's ability to interpret complex interactions, thereby enhancing its accuracy. Unlike the polyline representation, the BEV format uses rasterized images to represent spatial data. Our proposed encoder hierarchically extracts frame-wise pyramid features and fuses them into BEV tokens to facilitate the interaction between these two modalities. BEV is reorganized into a three-dimensional tensor with a hierarchical structure, comprising three distinct semantic layers: Agent, Map and Raster. To mitigate the risk of occlusion that arises when directly overlaying semantic layers onto images, we employ a series of convolutional kernels with varying sizes to extract key information from these semantic layers. This allows for a comprehensive capture of the interactions among the heterogeneous elements of the scene. Finally, We apply an MLP to refine the feature, resulting in the generation of B^h.

Diffusion-based Backdoor Adjustment. This component aims to mitigate the impact of confounding factors in spatial maps by generating a diverse set of potential traffic scenarios, a process known as backdoor adjustment. Since backdoor adjustment requires the stratification of S, we propose leveraging spatial maps generated by a diffusion model to automatically approximate the stratified structure of S. Through this approach, we effectively sever the directed edges $S \rightarrow X$ and $S \rightarrow T$, thereby eliminating spurious correlations and ensuring more accurate causal relationships [Ge *et al.*, 2023]. Specifically, we utilize the output S^h from the spatial encoder as the input S_0^h for backdoor adjustment and define a diffusion sequence $(S_0^h, S_1^h, \ldots, S_k^h)$, with the forward process gradually introduces Gaussian noise to systematically disrupt the deterministic structure of the road network, eventually rendering it entirely stochastic. Formally,

$$q(S_j^h) = f_{noised}\left(S_{j-1}^h\right) \text{ for } j = 1, \ldots, m \tag{3}$$

Conversely, the backward process employs a transformer-based architecture to iteratively reconstruct the typical road structure from this stochastic state, ultimately generating multiple instances. Mathematically,

$$p_\theta\left(S_{j-1}^h\right) = f_{denoised}\left(S_j^h\right) \textit{ for } j = 1, \ldots, m \tag{4}$$

By repeating this generative process n times, we obtain a backdoor spatial token set $\bar{S} = \{S^{h,1}, S^{h,2}, \ldots, S^{h,n}\}$.

Targeted Multi-View Attention Module. To enhance the understanding of environmental influences on the target future trajectory and facilitate causal integration, we employ a multi-view attention mechanism across spatial, BEV, and temporal contexts. Formally,

$$X_{s_i} = \textit{Spatial-Attn.}\left(q = X^h, k = S^{h,i}, v = S^{h,i}\right) \tag{5}$$

$$X_b = \textit{BEV-Attn.}\left(q = X^h, k = B^h, v = B^h\right) \tag{6}$$

$$X_t = \textit{Temporal-Attn.}\left(q = X^h, k = T^h, v = T^h\right) \tag{7}$$

The extracted tokens are processed through an aggregated MLP, producing the contextual target token X_{attn}^i.

Multi-mode Predictions

Cross-modal Progressive Fusion. This component is designed to integrate cross-modal tokens using progressive fusion. We initialize with an anchor-free query Q_0 to forecast the target agent's future trajectories. In the following progressive stages, this initial query operates as an adaptive anchor point. Specifically, an attention mechanism is employed to progressively refine our adaptive anchor, identifying the agent's intended position at the current moment, which then becomes the initial query for the subsequent update. After T_{rec} iterations, the definitive progressive inference anchor Q^i is developed. Notably, counterfactual analysis entails envisioning an alternate scenario where actual data is replaced with hypothetical data. To reduce the impact of confounding variables T in the trajectory prediction system, we substitute the historical trajectories with zero vectors and rerun the progressive fusion module to generate counterfactual anchor Q_c^i.

Dual-scale Information Fusion. This component employs a CNN-based framework to extract and integrate the target

Model	WSADE	ADEv	ADEp	ADEb	WSFDE	FDEv	FDEp	FDEb
TPNet [Fang *et al.*, 2021]	1.2800	2.2100	0.7400	1.8500	2.3400	3.8600	1.4100	3.4000
TP-EGT [Yang *et al.*, 2024]	1.1900	2.0500	0.7000	1.7200	2.1400	3.5300	1.2800	3.1600
S2TNet [Chen *et al.*, 2022a]	1.1679	1.9874	0.6834	1.7000	2.1798	3.5783	1.3048	3.2151
MSTG [Tang *et al.*, 2023]	1.1546	1.9850	0.6710	1.6745	2.1281	3.5842	1.2652	3.0792
SafeCast [Liao *et al.*, 2025a]	1.1253	1.9372	0.6561	1.6247	2.1024	3.5061	1.2524	3.0657
CoT-Drive [Liao *et al.*, 2025b]	1.0958	1.8933	0.6179	1.6305	2.0260	3.3541	1.1893	3.0244
Ours	**1.0681**	**1.8275**	**0.6138**	**1.5423**	**1.9174**	**3.1992**	**1.2164**	**2.6004**

Table 1: Evaluation results on the ApolloScape dataset. $\text{ADE}_{v/p/b}$ and $\text{FDE}_{v/p/b}$ are the ADE and FDE for the vehicles, pedestrians, and bicycles, respectively. **Bold** and underlined values represent the best and second-best performance in each category.

Model	$\text{minADE}_{10} \downarrow$	$\text{minADE}_{5} \downarrow$	FDE $\downarrow$
Trajectron++ [Salzmann *et al.*, 2020]	1.51	1.88	9.52
MultiPath [Chai *et al.*, 2019]	1.14	1.44	7.69
LaPred [Kim *et al.*, 2021]	1.12	1.53	8.12
PGP [Deo *et al.*, 2022]	0.94	1.27	7.17
EMSIN [Ren *et al.*, 2024]	-	1.77	-
Traj-LLM [Lan *et al.*, 2024]	0.99	1.24	-
DEMO [Wang *et al.*, 2025c]	1.04	1.20	6.90
LAformer [Liu *et al.*, 2024]	**0.93**	1.19	6.95
NEST [Wang *et al.*, 2025b]	-	1.18	6.87
Ours	**0.93**	**1.17**	**6.71**

Table 2: Performance comparison of various models on nuScenes dataset. **Bold** values represent the best performance in each category. "-" denotes the missing value.

Dataset	Model	Prediction Horizon (s)				
		1	2	3	4	5
NGSIM	WSiP [Wang *et al.*, 2023]	0.56	1.23	2.05	3.08	4.34
	MHA-LSTM [Messaoud *et al.*, 2020]	0.41	1.01	1.74	2.67	3.83
	STDAN [Chen *et al.*, 2022b]	0.39	0.96	1.61	2.56	3.67
	DACR-AMTP [Cong *et al.*, 2023]	0.57	1.07	1.68	2.53	3.40
	GaVa [Liao *et al.*, 2024f]	0.40	0.94	1.52	2.24	3.13
	HLTP++ [Liao *et al.*, 2024b]	0.46	0.98	1.52	2.17	3.02
	Ours	**0.32**	**0.83**	**1.47**	**2.09**	**2.87**
HighD	WSiP [Wang *et al.*, 2023]	0.20	0.60	1.21	2.07	3.14
	MHA-LSTM [Messaoud *et al.*, 2020]	0.19	0.55	1.10	1.84	2.78
	GaVa [Liao *et al.*, 2024f]	0.17	0.24	0.42	0.86	1.31
	DACR-AMTP [Cong *et al.*, 2023]	0.10	0.17	0.31	0.54	1.01
	HLTP++ [Liao *et al.*, 2024b]	0.11	0.17	0.30	0.47	0.75
	BAT [Liao *et al.*, 2024c]	0.08	0.14	0.20	0.44	0.62
	Ours	**0.06**	**0.12**	**0.19**	**0.39**	**0.58**
MoCAD	CS-LSTM [Deo and Trivedi, 2018]	1.45	1.98	2.94	3.56	4.49
	MHA-LSTM [Messaoud *et al.*, 2020]	1.25	1.48	2.57	3.22	4.20
	WSiP [Wang *et al.*, 2023]	0.70	0.87	1.70	2.56	3.47
	HLTP++ [Liao *et al.*, 2024b]	0.60	0.81	1.56	2.40	3.19
	BAT [Liao *et al.*, 2024d]	0.34	0.70	1.32	2.01	2.57
	NEST [Wang *et al.*, 2025b]	0.32	0.75	**1.27**	2.01	2.42
	Ours	**0.30**	**0.68**	**1.27**	**1.97**	**2.39**

Table 3: Evaluation results for our model and SOTA baselines in the NGSIM, HighD, and MoCAD datasets. RMSE (m) is the metric.

agent and its surrounding agent information, resulting in the feature representation G^i. For counterfactual analysis, the original target values are substituted with counterfactual values, and the module is applied synchronously to obtain the corresponding counterfactual representation G^i_c.

Causal Decoder. In this decoder, the factual tokens Q^i and G^i, along with the corresponding counterfactual tokens Q^i_c and G^i_c are integrated using MLPs to produce composite token $\tilde{Y}$ and $\tilde{Y}_c$. This composite token is then processed by a GRU-based decoder, which generates the predicted probability of maneuver y^k_{man} with the corresponding maneuver-based trajectory probability mixture model Y^k.

3.4 Learning Strategy

The training process of our model is divided into two steps. The first step involves training the backdoor adjustment module. In the second step, with the parameters of the backdoor adjustment module fixed, we train the entire model.

Step-1: Diffusion-based Backdoor Adjustment Training. In the initial step, we train the backdoor adjustment module using the following loss:

$$L_{back}(\theta) = \mathbb{E}_{\epsilon, S^h_0, j} \left\| \epsilon - \epsilon_\theta(S^h_j, j) \right\| \tag{8}$$

where ϵ is sampled from a normal distribution $\mathcal{N}(0, \mathbf{I})$, and the variable S^h_j is defined as: $S^h_j = \sqrt{\bar{\alpha}_j} S^h_0 + \sqrt{1 - \bar{\alpha}_j}\epsilon$.

Step-2: Full Model Training. In the second step, we train the entire model by combining two loss functions: the intention loss (L_{int}) and the trajectory loss (L_{traj}). Intention loss captures the driver's operational intent, while trajectory loss ensures the accuracy of the trajectory prediction. Intention loss is calculated as follows:

$$L_{int} = -\sum_{k=1}^{K} \hat{y}^k_{man} \log\left(y^k_{man}\right) \tag{9}$$

Here, $\hat{y}^k_{man}$ and y^k_{man} are the true probability and the predicted probability of maneuver k, respectively. Moreover, the trajectory loss L_{traj} adapts to different datasets and metrics:

$$L_{traj} = \lambda_0 L_0 + \lambda_1 L_{NLL} \tag{10}$$

where L_0 varies by dataset: minADE_k for nuScenes, WSADE for ApolloScape, and RMSE for NGSIM, MoCAD, and HighD. In addition, L_{NLL} represents the negative log-likelihood loss, while the coefficients λ_0 and λ_1 are learnable weights for L_0 and L_{NLL}, respectively. Hence, the overall loss L for the model is the sum of intention loss L_{int} and trajectory loss L_{traj}, i.e. $L = \lambda_{\text{int}} L_{\text{int}} + \lambda_{\text{traj}} L_{\text{traj}}$, where λ_{int} and λ_{traj} are the weights for intention loss and trajectory loss.

4 Experiments

4.1 Experiment Setup

The robustness and accuracy of our model are evaluated on five prominent real-world datasets, including nuScenes [Caesar *et al.*, 2020], ApolloScape [Huang *et al.*, 2018], MoCAD

Models	Add Noise		Drop Frame	
	minADE$_5$/FDE $\alpha = 8$	**minADE$_5$/FDE** $\alpha = 16$	**minADE$_5$/FDE** 20%	**minADE$_5$/FDE** 40%
PGP [Deo *et al.*, 2022]	1.56/9.03	1.64/9.34	1.49/8.29	1.58/9.15
Q-EANet [Chen *et al.*, 2024]	1.44/9.52	1.50/9.15	1.36/8.03	1.40/8.24
NEST [Wang *et al.*, 2025b]	1.45/8.47	1.59/8.98	1.29/7.54	1.37/8.12
DEMO [Wang *et al.*, 2025c]	1.49/8.65	1.55/9.18	1.31/7.38	1.39/8.06
Ours	**1.32/7.64**	**1.40/8.03**	**1.23/6.89**	**1.26/7.08**

Table 4: Robustness of different models on the nuScenes dataset.

Component	Ablated variants				
	A	B	C	D	E
BEV Encoder	✗	✔	✔	✔	✔
Human-like Progressive Fusion	✔	✗	✔	✔	✔
Dual-scale Information Fusion	✔	✔	✗	✔	✔
Causal Inference	✔	✔	✔	✗	✔

Table 5: Different components of ablation study.

Models	nuScenes		ApolloScape	
	minADE$_5$	minADE$_{10}$	WSADE	WSFDE
A	1.27	1.12	-	-
B	1.45	1.21	1.2870	2.1900
C	1.24	1.07	1.2210	2.0913
D	1.59	1.35	1.3264	2.2934
E	**1.17**	**0.93**	**1.0681**	**1.9174**

Table 6: Ablation results on nuScenes and ApolloScape.

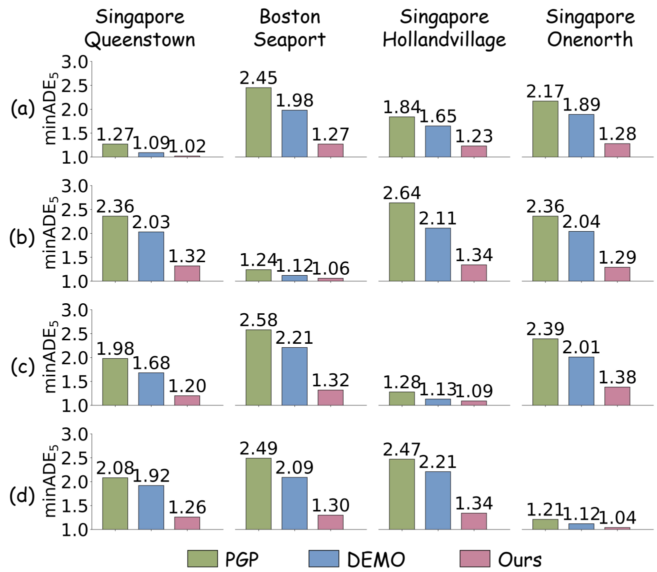

Figure 4: Exploration of the domain generalization ability for different models. (a) Singapore Queenstown, (b) Boston Seaport, (c) Singapore Hollandvillage, and (d) Singapore Onenorth. The results show the performance of models trained in these regions and tested in various other regions, which are evaluated using the minADE$_5$.

[Liao *et al.*, 2024c], NGSIM [Deo and Trivedi, 2018] and HighD [Krajewski *et al.*, 2018]. In accordance with the ApolloScape and nuScenes Forecasting Challenge, as well as the seminal work [Liao *et al.*, 2024c; Liao *et al.*, 2025b], five metrics are employed to assess the efficacy of our model: minADE, FDE, WSADE, WSFDE, and RMSE.

4.2 Evaluation Results

Comparison to SOTA Baselines. As shown in Tables 1, 2, and 3, our model outperforms the current SOTA baselines across five real-world datasets. On the ApolloScape dataset, our model surpasses MSTG [Tang *et al.*, 2023] in WSADE and WSFDE by 1.84% and 8.20%, respectively. In the nuScenes benchmark, it achieves improvements of at least 7.91% in minADE$_1$ and 4.00% in FDE, particularly in intersections. For NGSIM and HighD datasets, it improves the long-term prediction (3s-5s) by 13.82% and 6.45% in RMSE, respectively. For the MoCAD dataset, which represents urban roads with right-hand drive, we see an improvement of up to 6.60% in long-term predictions. These results validate the prediction accurancy of our model across challenging scenes, including highways, roundabouts, and congested urban roads.

Comparison of Robustness. We evaluate the robustness of our causal inference model using two methodologies: introducing controlled noise [Bagi *et al.*, 2023] and randomly removing frames. To simulate observation noise, we augment the nuScenes dataset by defining a noise variable σ_t as a function of curvature γ_t, formulated as follows:

$$\gamma_t := \left(\dot{x}_{t+\delta_t} - \dot{x}_t\right)^2 + \left(\dot{y}_{t+\delta_t} - \dot{y}_t\right)^2, \sigma_t := \alpha\left(\gamma_t + 1\right) \tag{11}$$

where $\dot{x}_t = x_{t+1} - x_t$ and $\dot{y}_t = y_{t+1} - y_t$ represent agent velocities, and α controls the noise magnitude. We set $\alpha \in \{1, 2\}$ during training and $\alpha \in \{8, 16\}$ during testing to simulate severe perturbations. To assess robustness against missing data, 20% or 40% of test frames are randomly removed and replaced with zeroes. Table 4 shows that under both noise and frame removal conditions, our model still outperforms SOTA baselines. These results demonstrate its ability to maintain high prediction accuracy by effectively learning and leveraging causal relationships.

Comparison of Domain Generalization. We partition the nuScenes dataset into four subsets and verify the statistical distinction in driving behaviors using a Kolmogorov-Smirnov test [Berger and Zhou, 2014], which yields p-values below the standard threshold. This confirm significant behavioral differences across regions. To assess the generalization capability of our causal inference model, we perform

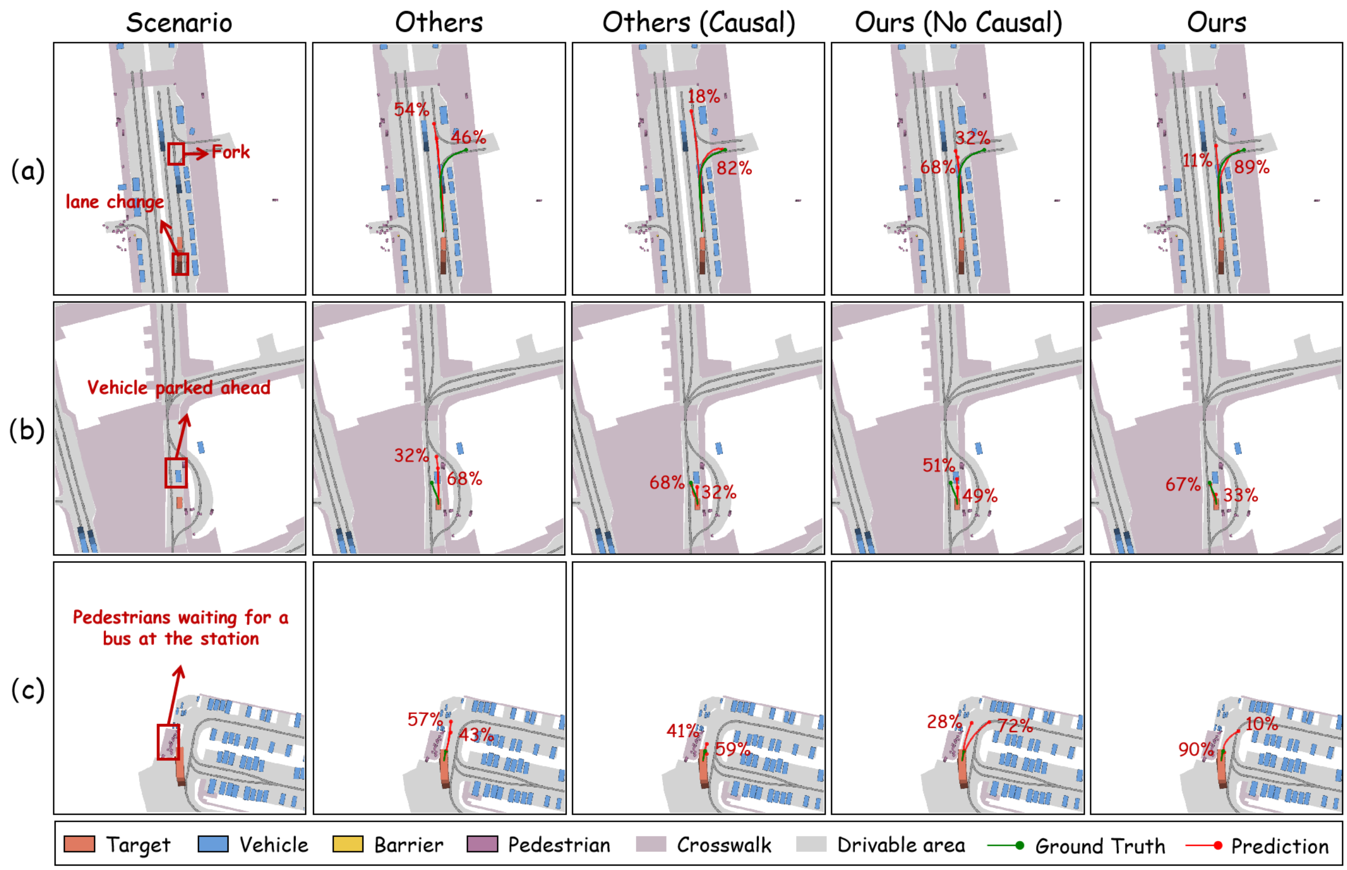

Figure 5: Comparative analysis of the impact of causal inference on our model and others across challenging scenes. For clarity, the number of predicted trajectories is set to $k = 2$. The probability of each predicted trajectory is shown in the subfigures.

cross-training and testing, where the model is trained on one subset and tested on others. As shown in Figure 4, our model achieves robust performance in unseen regions, demonstrating its effectiveness in enhancing domain generalization across diverse scenes.

4.3 Ablation Study

We present detailed ablation variants in Table 5 and Table 6. Method E represents the complete model, which achieves the SOTA performance across all metrics, demonstrating the synergistic effect of its components. Model D, which eliminates all causal analysis modules, including Backdoor Adjustment using diffusion, counterfactual analysis in progressive fusion, and dual-scale information fusion, and replaces the Differential Causal Decoder with a simple GRU, exhibits the poorest performance across most metrics, highlighting the significance of the causal inference paradigm in accurately learning casual relationship. Model B, which substitutes the progressive anchor fusion module with a single-stage anchor, also shows degraded performance, indicating the importance of progressive fusion in driving scenarios for precise trajectory prediction. The results of Method A and Method C show varying degrees of performance degradation, further confirming the necessity of each component within the model.

4.4 Validation for Plug-and-Play Causal Inference

We incorporate the causal paradigm into the PGP model by replacing its GRU-based encoder with a backdoor adaptation mechanism and by integrating counterfactual reasoning into the decoder. On the nuScenes dataset, as shown in Figure 5, our integrated model effectively captures causal relationships in complex lane-forking scenarios, where both the original and enhanced PGP versions fail to predict turning maneuvers. These results establish the role of causal inference in advancing behavior prediction performance.

5 Conclusion

This study advances trajectory prediction for AVs by integrating causal inference and progressive fusion strategies. Addressing the limitations of traditional data-driven methods that often neglect causal relationships, our model improves prediction accuracy and reliability through the use of causal graphs, backdoor adjustment, and counterfactual analysis, enabling it to distinguish true causal effects from correlations in complex traffic scenarios. Evaluation results on five real-world datasets show significant performance improvements across multiple metrics, marking a significant step forward in the field of trajectory prediction for fully AD systems.

Acknowledgements

This work was supported by the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Science and Technology Development Fund of Macau [0021/2022/ITP, 0122/2024/RIB2 and 001/2024/SKL], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055]. Bonan Wang and Haicheng Liao contributed equally to this work. Please ask Dr. Zhenning Li (zhenningli@um.edu.mo) for correspondence.

References

[Bagi *et al.*, 2023] Shayan Shirahmad Gale Bagi, Zahra Gharaee, Oliver Schulte, and Mark Crowley. Generative causal representation learning for out-of-distribution motion forecasting. In *ICML*, pages 31596–31612. PMLR, 2023.

[Berger and Zhou, 2014] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[Caesar *et al.*, 2020] 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 *CVPR*, pages 11621–11631, 2020.

[Chai *et al.*, 2019] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction, 2019.

[Chen *et al.*, 2021] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *ICCV*, pages 9824–9833, 2021.

[Chen *et al.*, 2022a] Weihuang Chen, Fangfang Wang, and Hongbin Sun. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving, 2022.

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

[Chen *et al.*, 2024] Jiuyu Chen, Zhongli Wang, Jian Wang, and Baigen Cai. Q-eanet: Implicit social modeling for trajectory prediction via experience-anchored queries. *IET Intelligent Transport Systems*, 18(6):1004–1015, 2024.

[Cong *et al.*, 2023] Peichao Cong, Yixuan Xiao, Xianquan Wan, Murong Deng, Jiaxing Li, and Xin Zhang. Dacr-amtp: Adaptive multi-modal vehicle trajectory prediction for dynamic drivable areas based on collision risk. *IEEE Transactions on Intelligent Vehicles*, 2023.

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

[Deo *et al.*, 2022] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *CoRL*, pages 203–212. PMLR, 2022.

[Fang *et al.*, 2021] Liangji Fang, Qinhong Jiang, Jianping Shi, and Bolei Zhou. Tpnet: Trajectory proposal network for motion prediction, 2021.

[Ge *et al.*, 2023] Chunjiang Ge, Shiji Song, and Gao Huang. Causal intervention for human trajectory prediction with cross attention mechanism. In *AAAI*, volume 37, pages 658–666, 2023.

[Guan *et al.*, 2024] Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Yunjian Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

[Huang *et al.*, 2018] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *CVPR workshops*, pages 954–960, 2018.

[Kim *et al.*, 2021] ByeoungDo Kim, Seong Hyeon Park, Seokhwan Lee, Elbek Khoshimjonov, Dongsuk Kum, Junsoo Kim, Jeong Soo Kim, and Jun Won Choi. Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents, 2021.

[Krajewski *et al.*, 2018] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018.

[Kuang *et al.*, 2020] Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, and Zhichao Jiang. Causal inference. *Engineering*, 6(3):253–263, 2020.

[Lan *et al.*, 2024] Zhengxing Lan, Lingshan Liu, Bo Fan, Yisheng Lv, Yilong Ren, and Zhiyong Cui. Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. *IEEE Transactions on Intelligent Vehicles*, 2024.

[Li *et al.*, 2024] Can Li, Wei Liu, and Hai Yang. Deep causal inference for understanding the impact of meteorological variations on traffic. *Transportation Research Part C: Emerging Technologies*, 165:104744, 2024.

[Liao *et al.*, 2024a] Haicheng Liao, Xuelin Li, Yongkang Li, Hanlin Kong, Chengyue Wang, Bonan Wang, Yanchen Guan, KaHou Tam, and Zhenning Li. Cdstraj: Characterized diffusion and spatial-temporal interaction network for trajectory prediction in autonomous driving. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7331–7339. International Joint Conferences on Artificial Intelligence Organization, 8 2024. AI for Good.

[Liao *et al.*, 2024b] Haicheng Liao, Yongkang Li, Zhenning Li, Chengyue Wang, Guofa Li, Chunlin Tian, Zilin Bian, Kaiqun Zhu, Zhiyong Cui, and Jia Hu. Less is more: Efficient brain-inspired learning for autonomous driving trajectory prediction. In *ECAI 2024*, pages 4361–4368. IOS Press, 2024.

[Liao *et al.*, 2024c] Haicheng Liao, Zhenning Li, Huanming Shen, Wenxuan Zeng, Dongping Liao, Guofa Li, and Chengzhong Xu. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. In *AAAI*, volume 38, pages 10332–10340, 2024.

[Liao *et al.*, 2024d] Haicheng Liao, Zhenning Li, Chengyue Wang, Huanming Shen, Dongping Liao, Bonan Wang, Guofa Li, and Chengzhong Xu. Mftraj: Map-free, behavior-driven trajectory prediction for autonomous driving. In Kate Larson, editor, *IJCAI*, pages 5945–5953, 8 2024.

[Liao *et al.*, 2024e] Haicheng Liao, Zhenning Li, Chengyue Wang, Huanming Shen, Bonan Wang, Dongping Liao, Guofa Li, and Chengzhong Xu. Mftraj: Map-free, behavior-driven trajectory prediction for autonomous driving. *arXiv preprint arXiv:2405.01266*, 2024.

[Liao *et al.*, 2024f] Haicheng Liao, Shangqian Liu, Yongkang Li, Zhenning Li, Chengyue Wang, Yunjian Li, Shengbo Eben Li, and Chengzhong Xu. Human observation-inspired trajectory prediction for autonomous driving in mixed-autonomy traffic environments. In *2024 ICRA*, pages 14212–14219. IEEE, 2024.

[Liao *et al.*, 2025a] Haicheng Liao, Hanlin Kong, Bin Rao, Bonan Wang, Chengyue Wang, Guyang Yu, Yuming Huang, Ruru Tang, Chengzhong Xu, and Zhenning Li. Safecast: Risk-responsive motion forecasting for autonomous vehicles. *arXiv preprint arXiv:2503.22541*, 2025.

[Liao *et al.*, 2025b] Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Wang Ye, Zhengbing He, Chengzhong Xu, and Zhenning Li. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*, 2025.

[Liao *et al.*, 2025c] Haicheng Liao, Zhenning Li, Guohui Zhang, Keqiang Li, and Chengzhong Xu. Toward human-like trajectory prediction for autonomous driving: A behavior-centric approach. *Transportation Science*, 05 2025.

[Liu *et al.*, 2024] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *CVPR*, pages 2039–2049, 2024.

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

[Mo *et al.*, 2022] Xiaoyu Mo, Zhiyu Huang, Yang Xing, and Chen Lv. Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9554–9567, 2022.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Ren *et al.*, 2024] Yilong Ren, Zhengxing Lan, Lingshan Liu, and Haiyang Yu. Emsin: Enhanced multi-stream interaction network for vehicle trajectory prediction. *IEEE Transactions on Fuzzy Systems*, 2024.

[Salzmann *et al.*, 2020] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *ECCV*, pages 683–700. Springer, 2020.

[Tang *et al.*, 2023] Luqi Tang, Fuwu Yan, Bin Zou, Wenbo Li, Chen Lv, and Kewei Wang. Trajectory prediction for autonomous driving based on multiscale spatial-temporal graph. *IET Intelligent Transport Systems*, 17(2):386–399, 2023.

[Wang *et al.*, 2023] Renzhi Wang, Senzhang Wang, Hao Yan, and Xiang Wang. Wsip: wave superposition inspired pooling for dynamic interactions-aware trajectory prediction. In *AAAI*, volume 37, pages 4685–4692, 2023.

[Wang *et al.*, 2025a] Chengyue Wang, Haicheng Liao, Zhenning Li, and Chengzhong Xu. Wake: Towards robust and physically feasible trajectory prediction for autonomous vehicles with wavelet and kinematics synergy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Wang *et al.*, 2025b] Chengyue Wang, Haicheng Liao, Bonan Wang, Yanchen Guan, Bin Rao, Ziyuan Pu, Zhiyong Cui, Chengzhong Xu, and Zhenning Li. Nest: A neuromodulated small-world hypergraph trajectory prediction model for autonomous driving. *AAAI*, 2025.

[Wang *et al.*, 2025c] Chengyue Wang, Haicheng Liao, Kaiqun Zhu, Guohui Zhang, and Zhenning Li. A dynamics-enhanced learning model for multi-horizon trajectory prediction in autonomous vehicles. *Information Fusion*, page 102924, 2025.

[Xie *et al.*, 2021] Xu Xie, Chi Zhang, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Congestion-aware multi-agent trajectory prediction for collision avoidance. In *2021 ICRA*, pages 13693–13700. IEEE, 2021.

[Yang *et al.*, 2024] Biao Yang, Fucheng Fan, Rongrong Ni, Hai Wang, Ammar Jafaripournimchahi, and Hongyu Hu. A multi-task learning network with a collision-aware graph transformer for traffic-agents trajectory prediction. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–14, 2024.