

Characterized Diffusion and Spatial-Temporal Interaction Network for Trajectory Prediction in Autonomous Driving

Haicheng Liao^{1*}, Xuelin Li^{2*}, Yongkang Li², Hanlin Kong², Chengyue Wang¹,
Bonan Wang¹, Yanchen Guan¹, KaHou Tam¹, Zhenning Li^{1*†}, Chengzhong Xu¹

¹University of Macau

²University of Electronic Science and Technology of China

{yc27979, chengyuewang, mc3500, yc37976, yc374361, zhenningli, czxu}@um.edu.com,
lxl.cooper@outlook.com, franklin1234560@163.com, hanlinkong@foxmail.com

Abstract

Trajectory prediction is a cornerstone in autonomous driving (AD), playing a critical role in enabling vehicles to navigate safely and efficiently in dynamic environments. To address this task, this paper presents a novel trajectory prediction model tailored for accuracy in the face of heterogeneous and uncertain traffic scenarios. At the heart of this model lies the Characterized Diffusion Module, an innovative module designed to simulate traffic scenarios with inherent uncertainty. This module enriches the predictive process by infusing it with detailed semantic information, thereby enhancing trajectory prediction accuracy. Complementing this, our Spatio-Temporal (ST) Interaction Module captures the nuanced effects of traffic scenarios on vehicle dynamics across both spatial and temporal dimensions with remarkable effectiveness. Demonstrated through exhaustive evaluations, our model sets a new standard in trajectory prediction, achieving state-of-the-art (SOTA) results on the Next Generation Simulation (NGSIM), Highway Drone (HighD), and Macao Connected Autonomous Driving (MoCAD) datasets across both short and extended temporal spans. This performance underscores the model’s unparalleled adaptability and efficacy in navigating complex traffic scenarios, including highways, urban streets, and intersections.

1 Introduction

In the domain of autonomous driving (AD), trajectory prediction plays a pivotal role by providing invaluable insights for the subsequent trajectory planning module, thereby enhancing the safety of navigation in complex and dynamic traffic scenarios [Huang *et al.*, 2022]. The continuous presence of mixed traffic flow necessitates a trajectory prediction model that deeply understands the heterogeneity and uncertainty of traffic scenarios [Liao *et al.*, 2024e]. Despite the proliferation of trajectory prediction models, significant gaps remain in the thorough investigation of the impact of heterogeneous and uncertain traffic scenarios on future motion.

* Authors contributed equally; †Corresponding author.

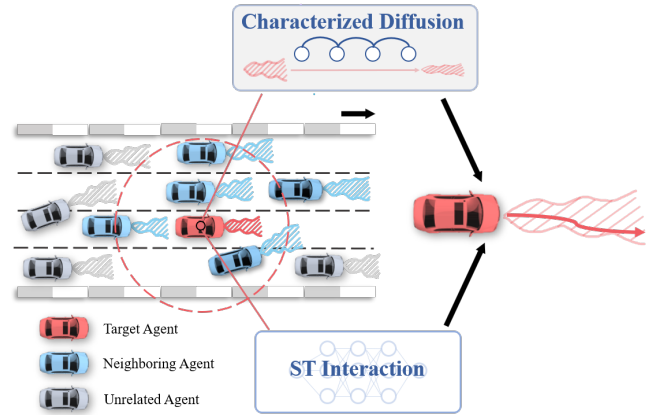


Figure 1: Overview of our model for processing past states. The framework utilizes two specialized modules to accomplish trajectory prediction for the target agent: Characterized Diffusion and Spatial-Temporal Interaction Network. In situations of high uncertainty, characterized diffusion employs a noisy Gaussian function to define a confidence region for the trajectory distribution. Continuous denoising isolates confidence features for future predictions. Meanwhile, the spatial-temporal interaction network extracts features to understand spatial relations and temporal dependency.

The initial gap we pinpoint hinges on the accurate simulation of future traffic scenarios—a cornerstone for enhancing trajectory prediction precision [Wang *et al.*, 2023; Liao *et al.*, 2024d]. The challenge is amplified by the intrinsic uncertainties characterizing traffic dynamics, making the accurate forecast of future scenarios a complex endeavor. Prevailing models have primarily concentrated on uncertainties inherent to the target agent [Zhao *et al.*, 2019; Alahi *et al.*, 2016; Gupta *et al.*, 2018], thereby neglecting the comprehensive uncertainty pervasive in the overall traffic scenarios. This oversight highlights an imperative need for trajectory prediction frameworks to adeptly navigate and mitigate uncertainties, facilitating a more accurate and holistic simulation of future traffic scenarios.

The second identified gap related to the sophisticated mechanisms by which traffic scenarios exert influence on human driving behaviors. The decision-making processes of human drivers are profoundly shaped by their interactions with

other traffic agents, with such interactions predicated on a nuanced interplay between spatial and temporal dimensions. Nevertheless, prevailing models have primarily focused on capturing spatial interactions, starkly overlooking the critical temporal dynamics [Wang *et al.*, 2023]. This neglect highlights the critical necessity for models that adeptly incorporate both spatial and temporal interaction.

To address these challenges, we introduce a novel generative model, CDSTraj, which is built on a dual architecture. As shown in Fig. 1. Our model employs an encoder designed to generate spatial-temporal features from past states, while significantly fusing confidence features to ensure the stability and reliability of the representation. This enriched feature set serves as the foundational input for a decoder that generates future trajectory predictions. A key innovation in our approach is the introduction of a characterized diffusion, which is seamlessly integrated with a spatial-temporal interaction network. This synergistic combination allows our model to be aware of the indeterminacy associated with both scene-to-agent and agent-to-agent contexts. Consequently, this leads to trajectory predictions that are both more accurate and reliable, even in dynamically changing environments. Overall, our main contributions can be summarized as follows:

- We introduce the Characterized Diffusion Module, a novel approach that enhances trajectory prediction by dynamically simulating future traffic scenarios through iterative uncertainty mitigation. This module significantly augments the predictive accuracy by integrating complex, contextual scenario features, allowing for a more nuanced understanding of potential motion.
- We unveil the Spatial-Temporal Interaction Module, which leverages a spatio-temporal attention mechanism to meticulously model and analyze the intricate interactions characteristic of traffic scenarios. Unique to this module is its three-stage architecture, meticulously designed to efficiently capture and process information across spatial and temporal dimensions.
- Our rigorous empirical investigations underscore the superiority of our model over the existing trajectory prediction model. Through extensive experiments, our model achieves top performance on public datasets such as NGSIM, HighD, and MoCAD. The exceptional performance on the MoCAD dataset is especially significant, offering a fresh perspective for evaluation through its unique right-hand drive configuration and obligatory left-hand traffic regime, thus underscoring our model’s adaptability and accuracy in varying driving scenarios.

2 RELATED WORKS

Trajectory Prediction for Autonomous Driving. Early trajectory prediction methods primarily relied on manual feature engineering and rule-based techniques, including linear regression and Kalman filters [Prevost *et al.*, 2007]. These methods were limited in capturing complex interactions in dynamic environments. The field evolved significantly with the introduction of deep learning, specifically Recurrent Neural Networks (RNNs) [Kim *et al.*, 2017] and Long Short-Term Memory (LSTM) networks [Altché and de La Fortelle,

2017; Alahi *et al.*, 2016; Liao *et al.*, 2024b]. These advancements enabled the capture of temporal dependencies in trajectories. Further innovation came with Graph Neural Networks (GNNs) [Zhou *et al.*, 2021; Liao *et al.*, 2024a; Liao *et al.*, 2024c], which provided a more nuanced approach to modeling interactions among agents in crowded scenes.

Generative Models for Trajectory Prediction. Generative models like Generative Adversarial Networks (GANs) [Gupta *et al.*, 2018] and Variational Auto-Encoders (VAEs) [Lee *et al.*, 2017] have gained prominence in trajectory prediction. GANs involve a generator and a discriminator engaged in mutual learning, while VAEs use a generative model and a variational posterior, the optimization of which can be complex. Diffusion models, on the other hand, offer a simplified training process by focusing on matching forward and inverse diffusion processes. To the best of our knowledge, this work is the first to leverage diffusion models for capturing the confidence features.

Denoising Diffusion Probabilistic Models. Denoising Diffusion Probabilistic Models (DDPM) [Ho *et al.*, 2020], known as diffusion models, have gained prominence as powerful generative models for various applications, including image [Ramesh *et al.*, 2022; Rombach *et al.*, 2022; Liao *et al.*, 2024f], video [Ho *et al.*, 2022], and 3D shape [Poole *et al.*, 2022] generation. Inspired by the diffusion models’ enormous representation capacities in numerous generation tasks, our work introduces the novel application of diffusion models to trajectory prediction in autonomous driving, addressing the challenges of modeling uncertainties and complex agent interactions in dynamic environments.

3 Problem Formulation

The paramount objective of this study is the precise prediction of trajectories for all entities within the proximity of an autonomous vehicle (AV) situated in an environment characterized by mixed autonomy. For this purpose, every entity proximal to the AV is designated as a *target agent*. At a specific time t_c , our model endeavors to utilize the historical states of both the target agent and its neighboring agents to predict the future trajectory of the target agent, represented as \mathbf{Y}_0 , extending to a future time $t_c + t_f$. The historical state since the time $t_c - t_h$ is denoted by \mathbf{X}_0 for the target agent and \mathbf{X}_i for the neighboring agents.

The novelty of our model lies in its exploitation of anticipated future traffic scenarios, specifically the future trajectories of neighboring agents, to enhance the accuracy of trajectory prediction for the target agent. To this end, we develop an Characterized Diffusion Module, designed to systematically mitigate the uncertainty inherent in the trajectories of neighboring agents, thereby enabling the accurate prediction of their future trajectories \mathbf{Y}_i . Formally, our prediction model Φ is represented as:

$$\mathbf{Y}_0 = \Phi(\mathbf{X}_0, \mathbf{X}_i, \mathbf{Y}_i) \quad \forall i \in [1, n] \quad (1)$$

4 Methodology

4.1 Overview of the Model Framework

As illustrated in Figure 2, our model comprises three primary components: the Characterized Diffusion Module, the

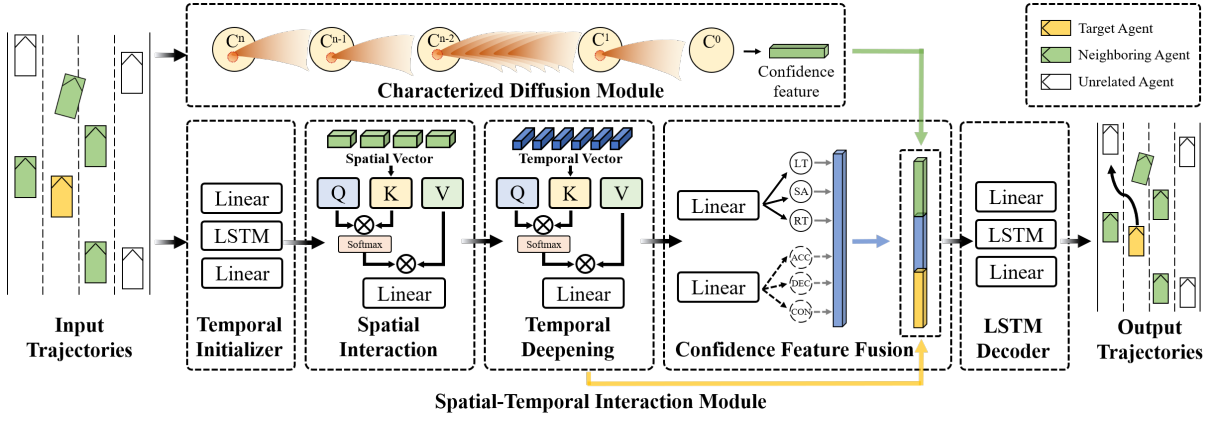


Figure 2: Framework of the proposed model.

ST Interaction Module, and a Multi-modal Decoder. The Characterized Diffusion Module employs an inverse diffusion process to generate the future trajectories of neighboring agents. Concurrently, the ST Interaction Module extracts Spatial-Temporal Interaction features through a methodical alternation between spatial and temporal dimensions. Ultimately, the predicted trajectories are generated using a multi-modal decoder, which synthesizes the processed information to produce accurate trajectory prediction.

4.2 Characterized Diffusion

In order to predict the future trajectories of neighboring agents, the Characterized Diffusion Module treats trajectory prediction as the reverse process of motion Characterized Diffusion and gradually eliminates the uncertainty of future trajectories by learning a parameterized Markov chain with the observed historical states. More specifically, during the diffusion process, the uncertainty inherent in future trajectories is simulated by iteratively introducing Gaussian noise. Conversely, in the inverse diffusion process, this uncertainty is iteratively mitigated to derive the anticipated future trajectories accurately. The detailed procedure is shown in Fig. 3. Mathematically, let \mathbf{C} be the future trajectory of the neighboring agents. Firstly, We initialize the diffused unit \mathbf{C}^0 :

$$\mathbf{C}^0 = \mathbf{C} \quad (2)$$

We use a forward diffusion operation $f_{diffuse}(\cdot)$ to add uncertainty to $\mathbf{C}^{\delta-1}$ and transition to diffused unit \mathbf{C}^δ :

$$\mathbf{C}^\delta = f_{diffuse}(\mathbf{C}^{\delta-1}), \delta = 1, \dots, \Gamma \quad (3)$$

where \mathbf{C}^δ is the diffused unit at the δ^{th} diffusion step, Γ is the total number of steps.

After n iterations, our model is able to capture a comprehensive spectrum of uncertain traffic scenarios with maximum coverage. Then, the inverse diffusion process is applied to accurately derive the future trajectories of neighboring vehicles. This step-by-step refinement ensures high fidelity in predicting vehicle movements, effectively addressing the inherent complexity of dynamic traffic scenarios. Due to the indeterminacy of future trajectories, in the denoising procedure,

it is usually more reliable to capture more than one inverse unit to extract sufficient trajectory information. Therefore, we draw K independent and identically distributed samples to initialize the denoising unit $\hat{\mathbf{C}}_k^\Gamma$ from a normal distribution.

$$\hat{\mathbf{C}}_k^\Gamma \stackrel{i.i.d}{\sim} \mathcal{P}(\hat{\mathbf{C}}^\Gamma) = \mathcal{N}(\hat{\mathbf{C}}^\Gamma; \mathbf{0}, \mathbf{I}), \text{ sample } K \text{ times}, \quad (4)$$

We formulate the trajectories generation process as a reverse diffusion by iteratively applying a denoising operation $f_{denoise}(\cdot)$ to obtain the denoised unit $\hat{\mathbf{C}}_k^\delta$ conditioned on historical states $\mathbf{X}_0, \mathbf{X}_i$ and, unit $\hat{\mathbf{C}}_k^{\delta+1}$:

$$\hat{\mathbf{C}}_k^\delta = f_{denoise}(\hat{\mathbf{C}}_k^{\delta+1}, \mathbf{X}_0, \mathbf{X}_i), \delta = \Gamma - 1, \dots, 0, \quad (5)$$

In a denoising module, two parts are trainable: a transformer-based context encoder $f_{context}(\cdot)$ to learn a social-temporal embedding and an uncertainty estimation module $f_\epsilon(\cdot)$ to estimate the uncertainty to reduce. Mathematically, the δ^{th} denoising step works as follows:

$$\mathbf{C}_{encoder} = f_{context}(\mathbf{X}_0, \mathbb{X}_i) \quad (6)$$

$$\epsilon_\theta^\delta = f_\epsilon(\hat{\mathbf{C}}_k^{\delta+1}, \mathbf{C}_{encoder}, \delta + 1) \quad (7)$$

$$\hat{\mathbf{C}}_k^\delta = \frac{1}{\sqrt{\alpha_\delta}} \left(\hat{\mathbf{C}}_k^{\delta+1} - \frac{1 - \alpha_\delta}{\sqrt{1 - \bar{\alpha}_\delta}} \epsilon_\theta^\delta \right) + \sqrt{1 - \alpha_\delta} \mathbf{z} \quad (8)$$

where α_δ and $\bar{\alpha}_\delta = \prod_{i=1}^\delta \alpha_i$ are parameters in the diffusion process and $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ is the uncertainty. We leverage a context encoder $f_{context}(\cdot)$ on historical states $(\mathbf{X}, \mathbb{X}_N)$ to obtain the context condition $\mathbf{C}_{encoder}$, and estimates the uncertainty ϵ_θ^δ in the uncertain trajectory $\hat{\mathbf{Y}}_k^{\delta+1}$ through uncertainty estimation $f_\epsilon(\cdot)$ implemented by multi-layer perceptions with the context \mathbf{C} ; Eq. 7 provides a standard denoising step.

The final K extracted units are $\hat{\mathbf{C}} = \{\hat{\mathbf{C}}_1^0, \hat{\mathbf{C}}_2^0, \dots, \hat{\mathbf{C}}_K^0\}$. Finally, we obtain the future trajectories with:

$$\mathbf{Y}_i = \Omega(\hat{\mathbf{C}}, W_{cf}) \quad (9)$$

where Ω denotes a multi-layer perception with learnable parameter matrix W_{cf} .

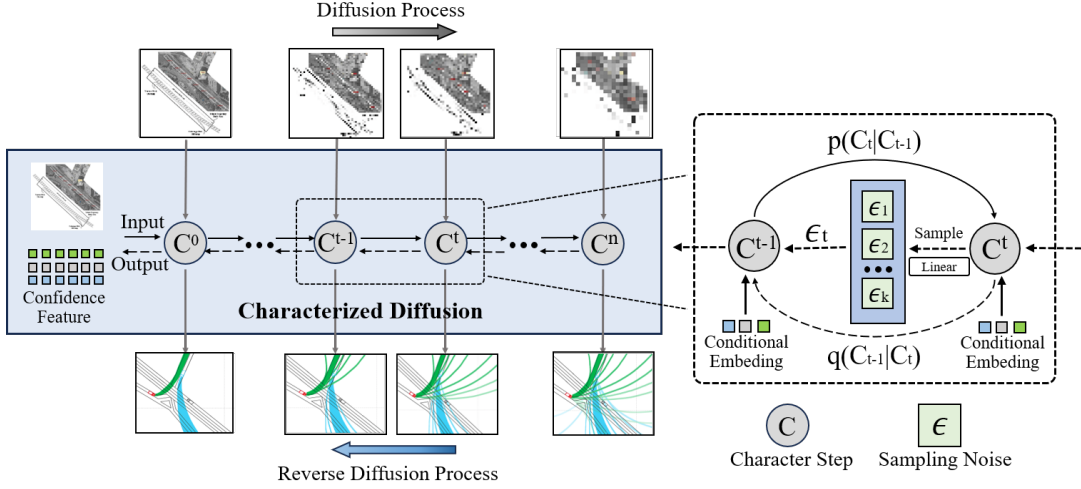


Figure 3: Overview of the Characterized Diffusion Module. The observed historical states are used as input for forward diffusion, noise is added for every steps by the Gaussian distributions with controllable means and variances. When $t = n$, the model performs reverse diffusion guided by the conditional embedding between each step to estimate the noise from the sampled noises with a Linear network, and finally output future trajectories.

4.3 Spatial and Temporal Interaction

To enhance the precision of modeling the temporal and spatial dynamics of vehicle interactions within the environment, the Space-Time (ST) Interaction Module is designed with a novel structure that alternates between temporal and spatial dimensions. This module is composed of three key components, as illustrated in Fig. 2: 1. **Temporal Encoder**: where the temporal dependencies of all the agents are extracted from the historical states by a temporal encoder. 2. **Spatial Encoder**: which plays an essential role in extracting the spatial relation between the target agent and neighboring agents. 3. **ST Fusion**: which aims to deeply capture the spatial-temporal interaction.

1) **Temporal Encoder**: To begin with, a temporal embedding vector F^t is obtained from the historical states x^t at the t^{th} timestamp using a fully connected layer with a learnable parameter matrix W_{emb} as follows:

$$F^t = \delta(\phi(x^t, W_{emb})) \quad (10)$$

where δ is the fully connected layer and Φ is the LeakyReLU activation function. It can be defined as follows:

$$h^t = f_{tem}(F^t, h^{t-1}, W_{init}) \quad (11)$$

where W_{init} denotes the learnable parameter matrix of encoder f_{tem} and h^t denotes the temporal feature at the t^{th} timestamp, which is updated at each timestamp based on the hidden state at the previous timestamp and the embedding vector at the current timestamp. Here, we apply f_{tem} to every single agent shared with the parameters of encoding f_{tem} to reduce variation across numerous agents at different timestamps. Finally, for the target agent, we obtain $H_0 = [h_0^{-T_p+1}, h_0^{-T_p+2}, \dots, h_0^0] \in R^{T_p \times D}$ representing the temporal feature over T_p timestamps, and $\bar{H}_i = [\bar{h}_i^{-T_p+1}, \bar{h}_i^{-T_p+2}, \dots, \bar{h}_i^0] \in R^{T_p \times D}$ denoting the temporal

feature of i^{th} neighboring agents and D denotes the number of hidden dimensions.

2) **Spatial Encoder**: Apparently, naive use of one temporal feature per agent does not capture agent-to-agent spatial relations. Therefore, capturing the spatial relations between agents in the same scene is highly necessary. Given the success of attention mechanisms in sequence-based prediction [Hu, 2020], we adopt a multi-head attention mechanism to obtain spatial relations between agents, which can be represented as follows:

$$Q, K, V = f_{sp}(H, \bar{H}, W_q, W_k, W_v) \quad (12)$$

where f_{sp} is the spatial attention.

In a detailed elaboration, $Q = [q^{-T_p+1}, q^{-T_p+2}, \dots, q^0]$, $K = [k^{-T_p+1}, k^{-T_p+2}, \dots, k^0]$, $V = [v^{-T_p+1}, v^{-T_p+2}, \dots, v^0]$ respectively denote linear projected vectors query, key, value and W_q, W_k, W_v indicate three learnable parameter matrices. We apply normalization Π to query and key to represent the importance of agents influencing each other. Formally,

$$\omega = \Pi\left(\frac{(Q \cdot K)}{\sqrt{D_{init}}}\right) \quad (13)$$

where ω is the attention score representing the similarity between Q and K , (\cdot) denotes the operation of matrix multiplication. Subsequently, we leverage attention scores to discern the significant connections among agents. Mathematically,

$$\Upsilon = \omega \cdot V \quad (14)$$

where v is the output of the single-head attention. Compared to single-head, multi-head attention can comprehensively capture local and global spatial relations. Therefore, we apply the multi-head attention mechanism to obtain $\Upsilon = [v_1, v_2, \dots, v_n]$, carrying spatial relations. We introduce an innovative gating mechanism H_g to control the importance of

different heads and selectively amplify or suppress specific heads. This mechanism acts as a gatekeeper for the Υ by adjusting the activation level consisting of two linear layers as:

$$H_a = \kappa(\Upsilon), H_g = \sigma(\kappa(\Upsilon)), S = H_a \odot H_g \quad (15)$$

where σ denotes the activation function sigmoid and \odot denotes the multiplication of the corresponding elements of the matrix, κ denotes the linear layer. Note that the output of the spatial encoder we simplify as follows:

$$S = [s^{-T_p+1}, s^{-T_p+2}, \dots, s^0] \in R^{T_p \times D} \quad (16)$$

3) **ST Fusion:** Following the spatial encoder, we introduce a ST Fusion f_{ST} to deeply capture spatial-interaction from S produced by the preceding module. Formally,

$$\bar{Q}, \bar{K}, \bar{V} = f_{ST}(S, \bar{W}_q, \bar{W}_k, \bar{W}_v) \quad (17)$$

where $\bar{Q} = [\bar{q}_1^t, \bar{q}_2^t, \dots, \bar{q}_M^t]$, $\bar{K} = [\bar{k}_1^t, \bar{k}_2^t, \dots, \bar{k}_M^t]$, $\bar{V} = [\bar{v}_1^t, \bar{v}_2^t, \dots, \bar{v}_M^t]$ respectively denote linear projected vectors query, key, value at the t^{th} timestamp, and $\bar{W}_q, \bar{W}_k, \bar{W}_v$ indicate three learnable parameter matrices. In analogy with the spatial interaction module, we also use normalization and gating mechanisms to obtain long-term temporal features simplified as follows:

$$U = [u^{-T_p+1}, u^{-T_p+2}, \dots, u^0] \in R^{T_p \times D} \quad (18)$$

4.4 Decoder

This study defines the trajectory prediction task as a conditional probabilistic prediction problem. Specifically, the decoder is designed to predict the future trajectory for the target agent based on different lateral and longitudinal maneuver classes:

$$P(\hat{Y}) = P(Y|P_{lat}, P_{lon}) \cdot P_{lat} \cdot P_{lon} \quad (19)$$

where $P(\hat{Y})$ is the conditional probabilistic distribution for the predicted trajectory. In detail, we use the LSTM decoder to implement the final multi-modal trajectory prediction.

$$\hat{y}^t = f_{LSTM}(F, \hat{y}^{t-1}, W_{decoder}) \quad (20)$$

where \hat{y}^t is the predicted 2D spatial coordinate at future t^{th} timestamp, $W_{decoder}$ denotes the parameter matrix to be learned in the LSTM. Based on the comprehensive feature vectors, we can only apply the simplest LSTM decoder to predict the accurate trajectory easily with fewer parameters.

5 Experiment

To evaluate the performance of our model, we perform expensive experiments on real-world datasets. This study uses a consistent segmentation framework for all three datasets. Each sample is divided into 8-second segments, with the first 16 timestamps (3 seconds) serving as historical data and the following 25 timestamps (5 seconds) for evaluation.

5.1 Datasets

Next Generation Simulation (NGSIM): This dataset [Deo and Trivedi, 2018b] consists of vehicle trajectory datasets from US-101 and I-80, containing approximately 45 minutes

of vehicle trajectory data at 10 Hz. It is critical for the analysis of vehicle behavior in a variety of traffic scenarios and assists in the development of reliable AD models.

Highway Drone (HighD): HighD [Krajewski *et al.*, 2018] is a dataset of vehicle trajectories collected from six locations on German highways. It includes 110,000 vehicles, including cars and trucks, and a total distance traveled of 45,000 km. This dataset provides detailed information about each vehicle, including type, size, and maneuvers, making it invaluable for advanced vehicle trajectory analysis and AD research.

Macau Connected Autonomous Driving (MoCAD): It [Liao *et al.*, 2024b] was collected from the first Level 5 autonomous bus in Macau, which has undergone extensive testing and data collection since its deployment in 2020. The data collection period spans over 300 hours and covers various scenarios, including a 5-kilometer campus road dataset, a 25-kilometer dataset covering city and urban roads, and complex open traffic environments captured under different weather conditions, time periods, and traffic densities.

Table 1: Evaluation of the proposed model and baselines on the NGSIM dataset over a 5-second prediction horizon. The accuracy metric is RMSE (m). Cases marked as ('-') indicate unspecified values. **Bold** and underlined values represent the best and second-best performance in each category.

Model	Prediction Horizon (s)				
	1	2	3	4	5
S-LSTM [Alahi <i>et al.</i> , 2016]	0.65	1.31	2.16	3.25	4.55
S-GAN [Gupta <i>et al.</i> , 2018]	0.57	1.32	2.22	3.26	4.40
CS-LSTM [Deo and Trivedi, 2018a]	0.61	1.27	2.09	3.10	4.37
MATF-GAN [Zhao <i>et al.</i> , 2019]	0.66	1.34	2.08	2.97	4.13
DRBP [Gao <i>et al.</i> , 2023]	1.18	2.83	4.22	5.82	-
M-LSTM [Deo and Trivedi, 2018b]	0.58	1.26	2.12	3.24	4.66
IMM-KF [Lefkopoulou <i>et al.</i> , 2020]	0.58	1.36	2.28	3.37	4.55
GAIL-GRU [Kuefler <i>et al.</i> , 2017]	0.69	1.51	2.55	3.65	4.71
MFP [Tang and Salakhutdinov, 2019]	0.54	1.16	1.89	2.75	3.78
NLS-LSTM [Messaoud <i>et al.</i> , 2019]	0.56	1.22	2.02	3.03	4.30
MHA-LSTM [Messaoud <i>et al.</i> , 2021]	0.41	1.01	1.74	2.67	3.83
WSIP [Wang <i>et al.</i> , 2023]	0.56	1.23	2.05	3.08	4.34
CF-LSTM [Xie <i>et al.</i> , 2021]	0.55	1.10	1.78	2.73	3.82
TS-GAN [Wang <i>et al.</i> , 2022]	0.60	1.24	1.95	2.78	3.72
STDAN [Chen <i>et al.</i> , 2022]	0.42	1.01	1.69	2.56	3.67
BAT [Liao <i>et al.</i> , 2024b]	0.23	0.81	<u>1.54</u>	<u>2.52</u>	3.62
FHIF [Zuo <i>et al.</i> , 2023]	0.40	0.98	1.66	<u>2.52</u>	3.63
DACR-AMTP [Cong <i>et al.</i> , 2023]	0.57	1.07	1.68	2.53	<u>3.40</u>
Our model	<u>0.36</u>	<u>0.86</u>	1.36	2.02	2.85

5.2 Training and Implement Details

We adopt a two-stage training approach to train our model. In the first stage, our model is trained to predict a future trajectory with the Mean Squared Error (*MSE*) [Pan, 2020] loss function as follows:

$$\mathcal{L}_{MSE}(\hat{y}, y) = \sum_{t=1}^{t_f} [(y_x^t - \hat{y}_x^t)^2 + (y_y^t - \hat{y}_y^t)^2] \quad (21)$$

where $(\hat{y}_x^t, \hat{y}_y^t)$ is the 2D spatial coordinate predicted and (y_x^t, y_y^t) denotes the corresponding ground-truth coordinate. It helps our model to learn the exact position information that approximates the true trajectory. When the model converges to a point using the *MSE* loss function, we use a Negative Log-Likelihood (*NLL*) [Kim *et al.*, 2020] loss function. This

Table 2: Evaluation of our model and SOTA baselines on HighD.

Model	Prediction Horizon (s)				
	1	2	3	4	5
S-LSTM [Alahi <i>et al.</i> , 2016]	0.22	0.62	1.27	2.15	3.41
S-GAN [Gupta <i>et al.</i> , 2018]	0.30	0.78	1.46	2.34	3.41
WSiP [Wang <i>et al.</i> , 2023]	0.20	0.60	1.21	2.07	3.14
CS-LSTM [Deo and Trivedi, 2018a]	0.22	0.61	1.24	2.10	3.27
MHA-LSTM [Messaoud <i>et al.</i> , 2021]	0.19	0.55	1.10	1.84	2.78
NLS-LSTM [Messaoud <i>et al.</i> , 2019]	0.20	0.57	1.14	1.90	2.91
DRBP [Gao <i>et al.</i> , 2023]	0.41	0.79	1.11	1.40	-
EA-Net [Cai <i>et al.</i> , 2021]	0.15	0.26	0.43	0.78	1.32
CF-LSTM [Xie <i>et al.</i> , 2021]	0.18	0.42	1.07	1.72	2.44
STDAN [Chen <i>et al.</i> , 2022]	0.19	0.27	0.48	0.91	1.66
GaVa [Liao <i>et al.</i> , 2024e]	<u>0.17</u>	<u>0.24</u>	<u>0.42</u>	<u>0.86</u>	<u>1.31</u>
Our model	0.13	0.21	0.32	0.38	1.05

Table 3: Evaluation of our model and SOTA baselines on MoCAD.

Model	Prediction Horizon (s)				
	1	2	3	4	5
S-LSTM [Alahi <i>et al.</i> , 2016]	1.73	2.46	3.39	4.01	4.93
S-GAN [Gupta <i>et al.</i> , 2018]	1.69	2.25	3.30	3.89	4.69
CS-LSTM [Deo and Trivedi, 2018a]	1.45	1.98	2.94	3.56	4.49
MHA-LSTM [Messaoud <i>et al.</i> , 2021]	1.25	1.48	2.57	3.22	4.20
NLS-LSTM [Messaoud <i>et al.</i> , 2019]	0.96	1.27	2.08	2.86	3.93
WSiP [Wang <i>et al.</i> , 2023]	0.70	0.87	1.70	2.56	3.47
CF-LSTM [Xie <i>et al.</i> , 2021]	0.72	0.91	1.73	2.59	3.44
STDAN [Chen <i>et al.</i> , 2022]	0.62	0.85	1.62	2.51	3.32
HLTP [Liao <i>et al.</i> , 2024a]	<u>0.55</u>	<u>0.76</u>	<u>1.44</u>	<u>2.39</u>	<u>3.21</u>
Our model	0.39	0.82	1.43	2.08	2.74

transition facilitates a more comprehensive exploration of uncertainty within the trajectory prediction.

$$Loss_{NLL}(\hat{y}, y) = \sum_{t=1}^{T_F} \alpha((\sigma_x^t)^2(\Delta_x^t)^2 + (\sigma_y^t)^2(\Delta_y^t)^2 - 2\rho_{xy}^t\sigma_x^t\sigma_y^t(\Delta_x^t)(\Delta_y^t)) - \log(P^t)$$

where Δ_x^t, Δ_y^t represents $(y_x^t - \hat{y}_x^t), (y_y^t - \hat{y}_y^t)$, and σ_x^t, σ_y^t denote the standard deviation of x-coordinate and y-coordinate at the t th timestamp. ρ_{xy}^t is the correlation coefficient between the x and y coordinates at t th timestamp. $P^t = \rho_x^T \rho_y^T \sqrt{1 - (\rho_{xy}^t)^2}$ is the standard deviation of the probability density function in x and y coordinate. α is an empirical constant value to simplify the calculation.

5.3 Comparison to State-of-the-arts

Our model’s performance is evaluated against more than 15 state-of-the-art (SOTA) methods on each referenced dataset. The experimental results, presented in Table 1, show that our model provides significant improvements in trajectory prediction over the prevailing SOTA baselines. Using Root Mean Square Error (RMSE) as the evaluation metric, our model consistently outperforms most baselines, achieving improvements of 29% and 22% over WSiP and STDAN, respectively, over a 5-second horizon. On the HighD dataset, our model consistently outperforms current SOTA baselines, with average improvements ranging from 43%-70% for short-term forecasts (1-3 seconds) and 62%-78% for long-term forecasts (4-5 seconds). These improvements highlight the importance of integrating spatio-temporal and confidence features. It is

noteworthy that the prediction error of the HighD dataset is significantly lower than that of the NGSIM dataset for all algorithms, probably due to the fact that the HighD dataset provides more accurate trajectory data, including detailed information on location and speed. In addition, the HighD dataset contains approximately twelve times more samples than the NGSIM dataset. In addition, on the MoCAD dataset, our model excels on busy urban roads, outperforming SOTA baselines by at least 37% for short-term predictions and reducing long-term prediction errors by at least 0.58 metres. These improvements highlight the importance of incorporating characterised diffusion and spatio-temporal interaction networks. In conclusion, our results confirm the effectiveness and efficiency of our model in predicting AV trajectories.

5.4 Ablation Study

Table 4 analyzes five critical components: characterized diffusion, temporal initializer, spatial interaction, temporal deepening, and confidence feature fusion modules. It shows that we test six models, labeled Model A through Model F. Evaluations against the NGSIM datasets reveal that the stripped-down versions (Models A-E) consistently underperform compared to Model F, which includes all components. Importantly, the integration of the characterized diffusion and confidence feature fusion modules significantly enhances performance, underscoring their vital role in improving prediction accuracy. The inclusion of the confidence feature fusion module, which merges spatial-temporal confidence features with maneuver states, enables more precise predictions of the target agent’s trajectory, particularly through the incorporation of characterized diffusion.

Table 4: Ablation studies for core components in NGSIM dataset.

Components	Ablation Models					
	A	B	C	D	E	F
Characterized Diffusion	✗	✓	✓	✓	✓	✓
Temporal Encoder	✓	✗	✓	✓	✓	✓
Spatial Encoder	✓	✓	✗	✓	✓	✓
ST Fusion	✓	✓	✓	✗	✓	✓
Decoder	✓	✓	✓	✓	✗	✓
RMSE	3.09	3.05	2.97	3.02	3.16	2.85

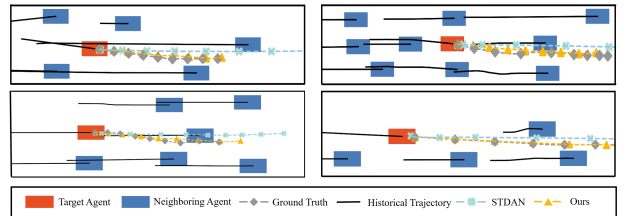


Figure 4: Visualization of our proposed model.

5.5 Qualitative Results

To qualitatively gain insight, we visualize the prediction results to assess the effectiveness of our proposed model. Fig.

4. displays the qualitative results of our model tested on the NGSIM dataset. As we can see, the prediction favorably covers the ground truth trajectory despite different numbers of neighboring agents in different scenes. These results demonstrate the impressive capabilities of the CDSTraj model in feature representation.

6 Conclusion

The intricate dynamics and uncertainties inherent in multi-agent interactions and scene contexts present a significant challenge in the development of fully AVs. To address this challenge, we introduce a novel generative model that employs a dual architecture integrating a characterized diffusion mechanism and a spatial-temporal interaction network. Empirical evaluations on the NGSIM, HighD, and MoCAD datasets demonstrate that our model, CDSTraj, consistently outperforms existing SOTA baselines in prediction accuracy over both short and long terms. Future research will explore the application of this model to pedestrian trajectory predictions and investigate the integration of spatial and temporal information. These endeavours may yield significant advancements in AD technologies.

Acknowledgements

This research is supported by the Science and Technology Development Fund of Macau SAR (File no. 0021/2022/ITP, 0081/2022/A2, 001/2024/SKL), and University of Macau (SRG2023-00037-IOTSC).

References

- [Alahi *et al.*, 2016] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE CVPR*, pages 961–971, 2016.
- [Alth   and de La Fortelle, 2017] Florent Alth   and Arnaud de La Fortelle. An lstm network for highway trajectory prediction. In *IEEE 20th ITSC*, pages 353–359. IEEE, 2017.
- [Cai *et al.*, 2021] Yingfeng Cai, Zihao Wang, Hai Wang, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li. Environment-attention network for vehicle trajectory prediction. *IEEE Transactions on Vehicular Technology*, 70(11):11216–11227, 2021.
- [Chen *et al.*, 2022] 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.
- [Cong *et al.*, 2023] Peichao Cong, Yixuan Xiao, Xianquan Wan, Murong Deng, Jiaying Li, and Xin Zhang. Dacrampt: adaptive multi-modal vehicle trajectory prediction for dynamic drivable areas based on collision risk. *IEEE Transactions on Intelligent Vehicles*, 2023.
- [Deo and Trivedi, 2018a] Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE CVPR Workshops*, pages 1468–1476, 2018.
- [Deo and Trivedi, 2018b] Nachiket Deo and Mohan M Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *IEEE IV*, pages 1179–1184. IEEE, 2018.
- [Gao *et al.*, 2023] Kai Gao, Xunhao Li, Bin Chen, Lin Hu, Jian Liu, Ronghua Du, and Yongfu Li. Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment. *IEEE Transactions on Intelligent Transportation Systems*, 2023.
- [Gupta *et al.*, 2018] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE CVPR*, pages 2255–2264, 2018.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [Ho *et al.*, 2022] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.
- [Hu, 2020] Dichao Hu. An introductory survey on attention mechanisms in nlp problems. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 432–448. Springer, 2020.
- [Huang *et al.*, 2022] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, 2022.
- [Kim *et al.*, 2017] ByeoungDo Kim, Chang Mook Kang, Jaekyum Kim, Seung Hi Lee, Chung Choo Chung, and Jun Won Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In *IEEE ITSC*, pages 399–404. IEEE, 2017.
- [Kim *et al.*, 2020] Hayoung Kim, Dongchan Kim, Gihoon Kim, Jeongmin Cho, and Kunsoo Huh. Multi-head attention based probabilistic vehicle trajectory prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1720–1725. IEEE, 2020.
- [Krajewski *et al.*, 2018] Robert Krajewski, Julian Bock, Laurent Kloecker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *21st ITSC*, pages 2118–2125. IEEE, 2018.
- [Kuefler *et al.*, 2017] Alex Kuefler, Jeremy Morton, Tim Wheeler, and Mykel Kochenderfer. Imitating driver be-

- havior with generative adversarial networks. In *IEEE IV*, pages 204–211. IEEE, 2017.
- [Lee *et al.*, 2017] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE CVPR*, pages 336–345, 2017.
- [Lefkopoulos *et al.*, 2020] Vasileios Lefkopoulos, Marcel Menner, Alexander Domahidi, and Melanie N Zeilinger. Interaction-aware motion prediction for autonomous driving: A multiple model kalman filtering scheme. *IEEE Robotics and Automation Letters*, 6(1):80–87, 2020.
- [Liao *et al.*, 2024a] Haicheng Liao, Yongkang Li, Zhenning Li, Chengyue Wang, Zhiyong Cui, Shengbo Eben Li, and Chengzhong Xu. A cognitive-based trajectory prediction approach for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, pages 1–12, 2024.
- [Liao *et al.*, 2024b] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10332–10340, 2024.
- [Liao *et al.*, 2024c] 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, 2024.
- [Liao *et al.*, 2024d] Haicheng Liao, Zhenning Li, Chengyue Wang, Bonan Wang, Hanlin Kong, Yanchen Guan, Guofa Li, Zhiyong Cui, and Chengzhong Xu. A cognitive-driven trajectory prediction model for autonomous driving in mixed autonomy environment. *arXiv preprint arXiv:2404.17520*, 2024.
- [Liao *et al.*, 2024e] Haicheng Liao, Shangqian Liu, Yongkang Li, Zhenning Li, Chengyue Wang, Bonan Wang, Yanchen Guan, and Chengzhong Xu. Human observation-inspired trajectory prediction for autonomous driving in mixed-autonomy traffic environments. *arXiv preprint arXiv:2402.04318*, 2024.
- [Liao *et al.*, 2024f] Haicheng Liao, Huanming Shen, Zhenning Li, Chengyue Wang, Guofa Li, Yiming Bie, and Chengzhong Xu. Gpt-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Communications in Transportation Research*, 4:100116, 2024.
- [Messaoud *et al.*, 2019] Kaouthar Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. Non-local social pooling for vehicle trajectory prediction. In *IEEE IV*, pages 975–980. IEEE, 2019.
- [Messaoud *et al.*, 2021] Kaouthar Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. Attention based vehicle trajectory prediction. *IEEE Transactions on Intelligent Vehicles*, 6(1):175–185, 2021.
- [Pan, 2020] Yunhe Pan. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216–217, 2020.
- [Poole *et al.*, 2022] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [Prevost *et al.*, 2007] Carole G Prevost, Andre Desbiens, and Eric Gagnon. Extended kalman filter for state estimation and trajectory prediction of a moving object detected by an unmanned aerial vehicle. In *2007 American control conference*, pages 1805–1810. IEEE, 2007.
- [Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF CVPR*, pages 10684–10695, 2022.
- [Tang and Salakhutdinov, 2019] Charlie Tang and Russ R Salakhutdinov. Multiple futures prediction. *Advances in neural information processing systems*, 32, 2019.
- [Wang *et al.*, 2022] Yu Wang, Shengjie Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):236–248, 2022.
- [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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4685–4692, 2023.
- [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 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13693–13700. IEEE, 2021.
- [Zhao *et al.*, 2019] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF CVPR*, pages 12126–12134, 2019.
- [Zhou *et al.*, 2021] Hao Zhou, Dongchun Ren, Huaxia Xia, Mingyu Fan, Xu Yang, and Hai Huang. Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction. *Neurocomputing*, 445:298–308, 2021.
- [Zuo *et al.*, 2023] Zhiqiang Zuo, Xinyu Wang, Songlin Guo, Zhengxuan Liu, Zheng Li, and Yijing Wang. Trajectory prediction network of autonomous vehicles with fusion of historical interactive features. *IEEE Transactions on Intelligent Vehicles*, 2023.