Vehicular Multimodal Motion Forecasting With Scenario-Guided Diffusion Probabilistic Models

Yi Zhou, *Member, IEEE*, Zhangyun Wang, *Student Member, IEEE*, Nianwen Ning, Ning Lu, *Member, IEEE*, Mianxiong Dong, *Senior Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—Ensuringthe safe driving and decision-making of autonomous vehicles requires reliable modeling of the multimodal motion of vehicles. In this paper, we propose a novel scenario-guided diffusion probabilistic model called RoadDiff to address these issues. By modeling and interacting with scene-contextual heterogeneous data at the feature level, we achieve lightweight and high-quality representations of traffic dynamics. We interpolate these representations as guiding conditions into a stochastic iterative denoising process of reverse diffusion, progressively transforming a known prior distribution into a multimodal trajectory distribution that includes the ground truth mode. Our well-designed Transformer-based neural network approximates the denoising matching function of the reverse diffusion process, and our region of interest-based training strategy ensures that the sampled multimodal trajectories conform to road and physical constraints to overcome mode blur. We carefully design two loss functions to ensure multimodal prediction for regions of interest and efficient trajectory guidance during this process. Experimental results on real-world motion datasets demonstrate that RoadDiff outperforms state-of-the-art methods in terms of prediction multimodality and accuracy, indicating its potential as a novel paradigm for vehicular multimodal motion prediction.**

***Index Terms*—Connected and automated vehicles, heterogeneous representation, multimodal motion prediction, diffusion probabilistic model.**

Received 7 August 2024; revised 23 January 2025, 15 March 2025, and 24 April 2025; accepted 22 May 2025. Date of publication 30 May 2025; date of current version 20 November 2025. This work was supported in part by the International Strategic Innovative Project of National Key Research & Development Program of China under Grant 2023YFE0112500, in part by the National Natural Science Foundation of China under Grant 62176088, in part by the Key Research and Development Project of Henan Province under Grant 251111211100, in part by the Program for Science & Technology Development of Henan Province under Grant 242102210195, in part by JSPS KAKENHI under Grant JP24K14910, and in part by the Key Technologies and Applications of Urban Spatiotemporal Foundation Models Integrated with Knowledge Graphs under Grant ghfund202407028284. The review of this article was coordinated by Prof. Zhaolong Ning. *(Corresponding author: Nianwen Ning.)*

Yi Zhou, Zhangyun Wang, and Nianwen Ning are with the School of Artificial Intelligence, Henan University, Zhengzhou 450046, China, and also with the International Joint Research Laboratory for Cooperative Vehicular Networks of Henan, Zhengzhou 450046, China (e-mail: zhouyi@henu.edu.cn; zhyun@henu.edu.cn; nnw@henu.edu.cn).

Ning Lu is with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: ning.lu@queensu.ca).

Mianxiong Dong is with the Department of Sciences and Informatics, Muroran Institute of Technology, Muroran 050-8585, Japan (e-mail: mx.dong@csse.muroran-it.ac.jp).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TVT.2025.3575478

I. Introduction

Autonomous driving is developing rapidly as an emerging field with a substantial potential social impact. Connected and automated vehicles (CAVs) will coexist in various traffic scenarios alongside human-driven vehicles for an extended period [1]. Understanding interactions between vehicles and anticipating their future motions are prerequisites for proactive hazard avoidance and one of the critical differences between machines and human drivers [2]. Developments of CAVs enable much more accurate and efficient acquisitions of scene information [3]. Inter-vehicle connectivity by dedicated short-range communications can extend the sensory coverage and reduce the response time of intelligent sensing devices, allowing CAVs to grasp continuous traffic dynamics and quasi-static high-definition maps comprehensively within the communication range. Thus, future motions of vehicles in dynamic traffic scenarios can be predicted by autonomous driving systems leveraging such real-time information.

The crucial challenge in predicting vehicle motions is the uncertainty of driving modes, such as lane-changing, yielding, or turning. Each mode requires reasoning with map semantics and observed features to ensure it is plausible to take over the coming period. Although recent studies have made significant progress in multimodal trajectory prediction, existing methods still suffer from mode collapse [4] or mode blur [5] problems due to their inherent modeling mechanisms and generative paradigms. As shown on the left of Fig. 1, mode collapse means all predicted trajectories converge to individual or finite modes within the multimodal one, which may cause all predictions to miss the future entirely, posing a significant collision risk. As shown on the right of Fig. 1, mode blur means part of the generated trajectories deviate from the lanes. CAVs are hardly obtaining valuable references from these predictions, which leads to the conservative driving of CAVs. Generative adversarial networks (GANs) [4] model highly multimodal trajectory distributions as Gaussian mixture models in an end-to-end manner. Social-GAN [4] and its improved architecture [6] alleviate mode collapse through variety-loss training and collaborative sampling strategies, respectively. However, they cannot simultaneously guarantee the diversity and quality of predicted trajectories. Conditional variational autoencoders (CVAEs) [7], [8], [9] learn mappings from samples to points in the latent space and generate reasonable samples aided by latent variables, which is revealed to be limited by mode blur in multimodal prediction. Choi et al. [5] introduced a hierarchical latent structure in CVAEs and utilized lane-level context vectors to modulate the latent

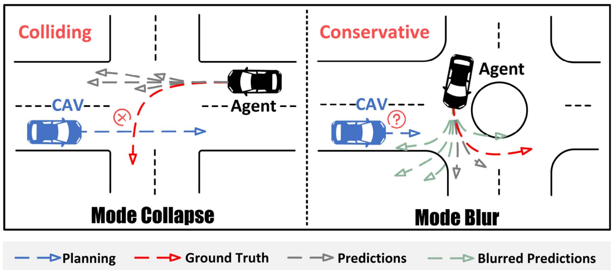


Fig. 1. Schematic illustration of mode collapse and mode blur in multimodal motion prediction. All predictions miss the future entirely when multimodal predictions collapse into just one mode (left), leading to dangerous decisions for CAVs that could collide. Most predictions are unrealistic or non-referential when multimodal predictions occur with mode blur (right). CAVs get confused and adopt more conservative driving modes, affecting traffic efficiency.

variables for each mode to mitigate mode blur. Proposal-based approaches introduce a set of target points to mitigate modal collapse. CTSGI [10] estimates target points for agents and neighbors employing specific latent distributions of CVAEs. The proposal-based paradigm leads to mode blur as the predicted trajectories exceed the drivable area on their paths to target points. In addition, the above methods rely on regularized features of scene contexts to implicitly generate trajectories while neglecting explicit constraints of reliable road regulations and driving areas, remaining limited in simultaneously solving mode collapse and mode blur problems.

Recent studies [11], [12], [13] show that diffusion probabilistic models (DPMs) achieve impressive results in image generation fields due to their stable multimodal modeling process, which is proven to be less prone to mode collapse. Nevertheless, extending the thought from image generation to vehicle motion prediction to address mode collapse and mode blur requires overcoming a series of challenges: (1) Implicit controllable generation of vehicle trajectories by scenario-guided DPMs. We are interested in learning parameter approximations of Gaussian transition kernels during reverse diffusion with the stochastic iterative refinement of scene context conditions mapped to agent trajectories. (2) Explicit constraint guidance to address mode blur while ensuring multimodal prediction capability. Pruning branches with mode blur directly to dynamically optimize prediction results can harm diversity. Because multiple pruning causes mode collapse, i.e., only the most rational modes are retained. Mode collapse and mode blur are mutually exclusive to some extent, making it difficult to simultaneously address both issues. Thus, we are motivated to answer the following question: can DPMs be improved in motion prediction to learn useful guidance representations and encourage generations of reasonable multimodal trajectories for greater accuracy and generality?

In this paper, these challenges are explicitly explored and tackled. We propose a novel end-to-end scenario-guided motion diffusion model called RoadDiff to model the multimodal trajectory distribution of vehicles accurately. To overcome the challenge (1), we naturally interpolate feature-level fusion representations of scene heterogeneous data into the reverse process of DPMs to guide trajectory generation. The cross-modal interactions of high-dimensional information prompt the model with a better understanding of the compatibilities and interactions between agents and map semantics to enrich the engagement and complementation of map resources. To overcome the challenge (2), we predict the multimodal region of interest (ROI) of agents and design two specific loss functions to encourage RoadDiff to conduct reverse diffusion within the region for efficient distribution propagations and reasonable trajectory generations. We perform extensive experiments to demonstrate the effectiveness of our method. The contributions of this paper are summarized as follows:

- We propose a novel paradigm of scenario-guided diffusion probabilistic model for vehicle multimodal motion prediction. RoadDiff simultaneously addresses the problems of mode collapse and mode blur in existing methods caused by the unbalanced distribution of data modes and generation mechanisms driven by implicit regularization features. We employ implicit high-quality representations and explicit regional constraints to co-guide multimodal predictions, controlling models to spread distributions via reverse diffusion within multimodal road-compliant ROIs.
- We learn scene context representations from various subspaces and integrate heterogeneous information at the feature level by cross-modal interactions, prompting models to perceive the compatibility and constraints of agent motions with maps. RoadDiff maps these high-quality representations to the stochastic iterative refinement process of trajectories to guide the parameter approximation learning of Gaussian transition kernels. We prove validity of heterogeneous representations, providing potential directions for representation learning of scene contexts.
- We design a diversity loss function for ROI prediction to eliminate the problems of positive and negative sample imbalance and mode imbalance during ROI prediction. It ensures that every complex or sparse mode is captured for robustly generating multimodal ROIs. We re-design the Hausdorff distance loss through distance transformations and inverse mapping of ROIs for the regional training strategy. It prevents enormous computational costs and unstable training and ensures efficient distribution transitions within ROIs during the reverse diffusion process.

The remainder of the paper is organized as follows. Section II outlines a comprehensive literature review of the latest research advances in vehicle motion prediction and DPMs. Section III expounds on the modeling procedure of our proposed scenario-guided DPMs-based vehicle motion prediction. The specific implementation of the critical components in RoadDiff is described in Section IV. Section V analyses the experiments conducted and the results. Finally, Section VI concludes and outlooks this paper.

II. Related Work

We focus on state-of-the-art methods of modeling scene context and multimodal motion prediction as well as research progress in DPMs, which leads us to propose RoadDiff.

A. Scene Context Representation

In early works on trajectory prediction, researchers utilized variants based on attention mechanisms to represent the degree of mutual influence and interactions between agents at different levels, such as spatial-temporal attention [14], social attention [15], graph attention [16], and agent-aware attention [8]. These approaches provide the underlying methodology for more complex forms of scene context modeling.

Pretraining-based methods establish specific loss functions against the inherent structure of spatio-temporal data, encouraging models to learn scalable and robust representations from the scenario context to accommodate various downstream tasks. BTCL [17] maximizes the similarity between different amplifications of the same trajectory while minimizing the similarity between different trajectory amplifications through a jointly constructed pretext task. Poibrenski et al. [18] proposed source-domain data augmentation and target-domain pseudo-label weighting strategies to improve the generalization of trajectory prediction models across different domains.

Growing researchers are beginning to focus on rendering, representing, and modeling scene contexts containing HD maps. MTP [19] and P2T [20] methods encode the contextual information of traffic scenes into raster images and leverage convolutional neural networks to extract spatial features and dynamic information. Trajectron++ [7] leverages graph structures to integrate agent dynamics and heterogeneous data, enabling future-conditional predictions that adhere to dynamic constraints while producing full probability distributions. Integrated graphical representation [21] quantifies the impact of elements in dynamic traffic scenarios by rendering color-scaled rasterized feature maps based on collision hazard levels. SGN [22] extracts generic static and dynamic representations of driving scenarios utilizing semantic and domain knowledge, followed by introducing semantic graphs to model the spatio-temporal structural relationships. Although the above methodologies have made great progress in modeling driving scenarios through single data forms, how to fuse heterogeneous data and whether they benefit from each other are rarely explored.

B. Multimodal Motion Prediction

Probabilistic generation methods support the end-to-end modeling of highly multimodal trajectory distributions, followed by the outputs of future trajectories by sampling [23]. Due to the training instability and mode collapse encountered during the process of reaching Nash equilibrium, GANs are revealed to be mainly constrained to relatively limited variation tasks. The variety loss of Social-GAN [4] only penalizes the samples closest to the ground truth and does not suppress low-quality samples during the training process. The ability of CVAEs models to approximate the ground truth posterior distribution relies on the choice of the posterior distribution and the design of the loss function, which suffers from poor sample accuracy and mode blur in multi-agent motion modeling. ScePT [9] captures the interactions between agents with Gibbs distributions and predicts mode probabilities for a whole clique. Social-CVAE [24] proposes a graph neural network architecture based on message-passing of sparse graph attention and a series of measures to correctly encode social interactions into latent spaces of variational self-encoders. Limited by the implicit generation of latent regularization features, such methods are inadequate in explicit guidance and still have limitations in overcoming model blur.

Proposal-based approach explicitly generates various candidate proposals, which are then regressed into multimodal motion trajectories [25]. Methods that leverage bridging intentions as proposals pursue multimodal results by traversing unimodal predictions over embedded combinations of intentions. CS-LSTM [26] fuses intention-specific features to output a multimodal trajectory distribution based on agent intention categories. iNATran [15] identifies crucial features strongly correlated with future trajectories from middle representations of intention estimation and generates queries in one step via an intention-aware decoder query generation module. This prediction paradigm embeds intentions across all modes in any scenario, leading to high diversity but inefficient, low-quality multimodal prediction samples and mode blur phenomenon. Methods that leverage target points or anchors as proposals [27] begin by predicting a set of turning points, endpoints, paths, etc. Real-Time [28] designs multi-task learning branches with dynamic adaptive anchors as part of end-to-end learning to improve coverage of multimodal trajectories. CoverNet [29] dynamically generates the trajectory set based on the current state of the agent and then transforms the trajectory prediction problem into the classification of possible physically feasible trajectories. LaPred [30] generates trajectory lane features for each lane candidate and then predicts multimodal future trajectories given multiple lane candidates. HOME [31] generates heatmaps representing probability distributions of agent locations through lane map representations and sparse projections of HD maps, followed by sampling possible endpoints from heatmaps. This prediction paradigm only constrains target points to drivable areas but not to the target-conditional motion estimation stage. The predicted trajectories may exceed drivable areas on paths to target points, leading to mode blur. Driving region is a robust map attribute, yet little work goes into applying it rationally. In this paper, we propose trajectory prediction methods with ROIs as proposals. We co-guide multimodal trajectory prediction through implicit regularization features and explicit multimodal region constraints to overcome mode collapse and mode blur problems.

C. Diffusion Probabilistic Models for Motion Prediction

MID [32] adjusts the length of the parameterized Markov chains to control the degree of uncertainty and balance the diversity and determinacy of the predictions. SingularTrajectory [33] constructs Singular spaces to unify human dynamics representations and utilizes environmentally adaptive anchors and cascading denoising diffusion predictors to improve cross-task trajectory prediction performance. LADM [34] extracts the pedestrian-group interaction features and then utilizes VAEs as generators and DPMs as refiners to improve the accuracy and real-time performance of predicted trajectories. MotionDiffuser [35] and OptTrajDiff [36] address the high computational cost and efficiency bottleneck in the controlled generation of joint trajectories by optimizing the initialization distribution and controlled sampling process of DPMs. TGD [37] inputs the BEV raster maps as conditions into DPMs and then extracts the rough trajectories from the generated raster maps, fed into the trajectory refiner to output the exact trajectories. Different from the above works, we design a ROI-based strategy to guide

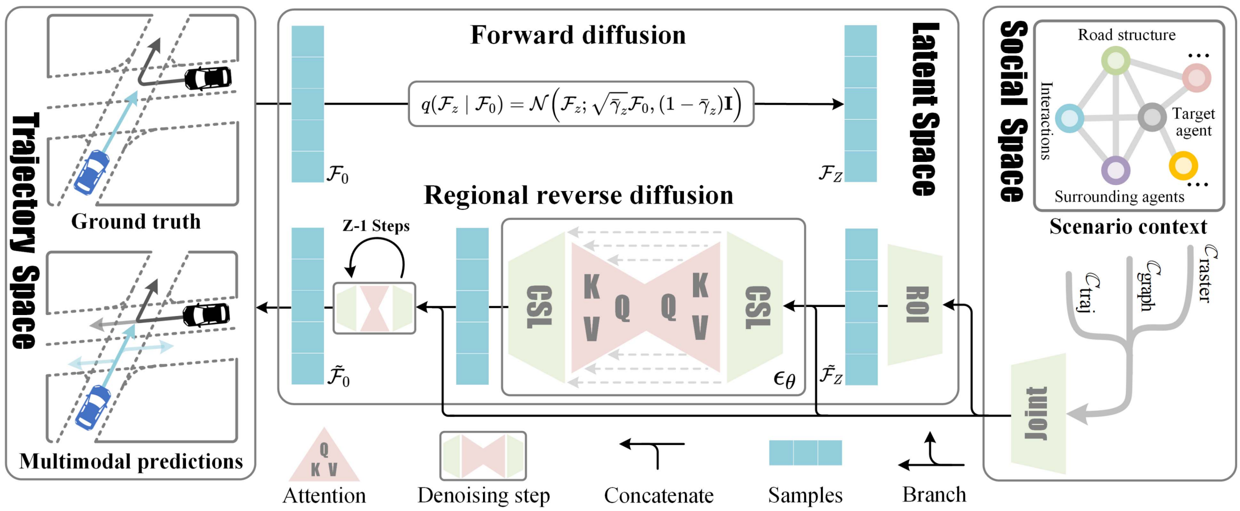


Fig. 2. Overview of RoadDiff.

reasonable distributional transitions during reverse diffusion. It explicitly considers the compatibility of mode blur and mode collapse while maintaining the strong diversity capability of DPMs.

III. Scenario-Guided Diffusion Probabilistic Models for Vehicle Motion Prediction

Creating noise from one trajectory is easy; creating multimodal trajectories from noise is RoadDiff. The diffusion process gradually transforms the complicated original trajectory distribution into the standard Gaussian distribution by progressively injecting tiny amounts of noise. The scenario-guided reverse diffusion process allows us to exploit useful information in the scenario contextual representation for accurate modeling. The stochastic Gaussian distribution is then transformed into a regionally constrained original data distribution by stepwise denoising, which contains multimodal results including the ground truth mode. The overview and core elements of RoadDiff architecture are shown in Fig. 2, and more details are presented in the following subsections.

A. Problem Formulation

With a known high-definition map $\mathcal{M}$ of urban roads, we describe the motion prediction problem as predicting the future trajectory $\mathcal{F}_0$ of the target agent in the future T_{pred} time period based on the observed trajectories $\mathcal{H}$ of the road participants in the past T_{obs} time period. $\mathcal{M}$ contains relevant semantic information that constrains and guides vehicle driving, such as drivable areas, angles, candidate lanes, turning signs, etc. Consider a driving scenario containing m agents, where number $i=1$ is the target agent and number $2\sim m$ is the surrounding road participants, including other vehicles and pedestrians, etc. We aim to predict n possible multimodal future trajectories $\tilde{\mathcal{F}}_0$ of the target agent. At any moment k, the agent position is defined as (x^k, y^k). In summary, the problem can be formulated as:

$$\mathcal{H}=\left\{(x_i^{k-T_{\text{obs}}+1}, y_i^{k-T_{\text{obs}}+1}), \ldots, (x_i^k, y_i^k) | i=1, \ldots, m\right\},$$

$$\begin{aligned}
\mathcal{F}_0 &= \left\{(x_i^{k+1}, y_i^{k+1}), \ldots, (x_i^{k+T_{\text{pred}}}, y_i^{k+T_{\text{pred}}}) | i=1\right\}, \\
\mathcal{M} &= \{drivablearea, angle, lanepolygon, \ldots\}, \\
\tilde{\mathcal{F}}_0 &= \Re(\mathcal{H}, \mathcal{M}; \vartheta) \\
&= \left\{(x_{i,j}^{k+1}, y_{i,j}^{k+1}), \ldots, (x_{i,j}^{k+T_{\text{pred}}}, y_{i,j}^{k+T_{\text{pred}}}) | i=1, j=1, \ldots, n\right\},
\end{aligned} \tag{1}$$

where $\Re$ denotes the trained multimodal motion prediction model and ϑ denotes the weights and parameters of the model.

B. Diffusion Processes for Vehicular Motions

Suppose the future trajectory $\mathcal{F}_0$ of the motion dataset obeys the unknown data distribution $q_{data}(\mathcal{F}_0)$. The diffusion process $\mathcal{F}_1, \mathcal{F}_2 \ldots, \mathcal{F}_Z$ transforms $\mathcal{F}_0$ into a prior Gaussian distribution by gradually injecting noise, of which the approximate posterior distribution can be expressed as a fixed Markov chain process without training parameters:

$$\begin{aligned}
q(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_Z \mid \mathcal{F}_0) &= \Pi_{z=1}^{Z} q(\mathcal{F}_z \mid \mathcal{F}_{z-1}), \\
q(\mathcal{F}_z \mid \mathcal{F}_{z-1}) &= \mathcal{N}(\mathcal{F}_z; \sqrt{1-\omega_z}\mathcal{F}_{z-1}, \omega_z \mathbf{I}),
\end{aligned} \tag{2}$$

where Z denotes the length of the longest Markov chain, $q(\mathcal{F}_z \mid \mathcal{F}_{z-1})$ denotes a Gaussian perturbation kernel, and $\{\omega_z \in (0, \tau)\}_{z=1}^{Z}$ denotes an increasing positive variance schedule satisfying $\frac{\omega_1}{\omega_2} = \ldots = \frac{\omega_{Z-1}}{\omega_Z} < 1$. The perturbation process becomes increasingly dramatic as the noise scale increases. The final perturbed variable $\mathcal{F}_Z$ of the ground truth trajectory $\mathcal{F}_0$ converges to an isotropic standard Gaussian distribution when $Z \to \infty$.

Furthermore, let $\gamma_z{=}1-\omega_z, \bar{\gamma}_z{=}\prod_{h=1}^{z}\gamma_h$. Based on the reparameterization [38] and the regeneration of the independent Gaussian distribution, the diffusion process defined in (2) allows us to directly sample any step of it with an analytic solution conditional on the input $\mathcal{F}_0$ [39]:

$$q(\mathcal{F}_z \mid \mathcal{F}_0) = \mathcal{N}(\mathcal{F}_z; \sqrt{\bar{\gamma}_z}\mathcal{F}_0, (1-\bar{\gamma}_z)\mathbf{I}). \tag{3}$$

Since the sampling action is discrete and non-differentiable, the perturbation variable $\mathcal{F}_z$ after adding z-step noise is obtained by employing the reparameterization technique again:

$$\mathcal{F}_z = \sqrt{\bar{\gamma}_z}\mathcal{F}_0 + \sqrt{1-\bar{\gamma}_z}\epsilon. \tag{4}$$

Therefore, the randomness from the computational graph is transferred to the Gaussian noise ϵ, guaranteeing that gradients are propagated.

C. Scenario-Guided Reverse Diffusion Processes

To distinguish the training process from the sampling process for the reverse diffusion process, we employ $\mathcal{F}_z$ without superscript to denote the perturbation variables in the training process that depend on the diffusion process, and $\tilde{\mathcal{F}}_z$ with superscript to denote the denoising variables sampled directly from the prior distribution. The reverse diffusion process $\tilde{\mathcal{F}}_Z, \tilde{\mathcal{F}}_{Z-1}, \ldots, \tilde{\mathcal{F}}_0$ is a Markov chain with learning parameters, starting from a standard Gaussian distribution $p(\tilde{\mathcal{F}}_Z) = \mathcal{N}(\tilde{\mathcal{F}}_Z; \mathbf{0}, \mathbf{I})$ of the same dimension as $\mathcal{F}_0$. The scenario-guided reverse diffusion process converts the standard Gaussian distribution $p(\tilde{\mathcal{F}}_Z)$ back to an approximate ground truth data distribution $q_{data}(\mathcal{F}_0)$ by gradually eliminating the noise added at each diffusion process step utilizing a parameterized Gaussian transition kernel:

$$p_\theta(\tilde{\mathcal{F}}_{z-1}|\tilde{\mathcal{F}}_z, \mathbb{C}) = \mathcal{N}(\tilde{\mathcal{F}}_{z-1}; \mu_\theta(\tilde{\mathcal{F}}_z, z, \mathbb{C}), \sigma_\theta(\tilde{\mathcal{F}}_z, z, \mathbb{C})\mathbf{I}), \tag{5}$$

where μ_θ is a learnable parameter conditioned on the denoising variable $\tilde{\mathcal{F}}_z$, the diffusion step z, and the scene context representation $\mathbb{C}$. Following recent empirical work [32], we focus the training parameters on μ_θ and thus set $\sigma_\theta(\tilde{\mathcal{F}}_z, z, \mathbb{C})\mathbf{I} = \omega_z\mathbf{I}$ to be time-dependent, which leads to more stable training. The final data distribution model $p_\theta(\tilde{\mathcal{F}}_0)$ is parsed by the following joint distribution:

$$p_\theta(\tilde{\mathcal{F}}_Z, \tilde{\mathcal{F}}_{Z-1}, \ldots, \tilde{\mathcal{F}}_0 \mid \mathbb{C}) = p(\tilde{\mathcal{F}}_Z)\Pi_Z^1 p_\theta(\tilde{\mathcal{F}}_{z-1} \mid \tilde{\mathcal{F}}_z, \mathbb{C}). \tag{6}$$

The original data distribution $q_{data}(\mathcal{F}_0)$ can be learned by maximizing the log-likelihood utilizing the variational evidence lower bound (ELBO) of Jensen's inequality [39], whose Kullback-Leibler (KL) divergence form can be calculated by:

$$\begin{aligned} &\max_\theta \mathbb{E}_{q_{data}(\mathcal{F}_0)}[\log p_\theta(\mathcal{F}_0 \mid \mathbb{C})] \\ &= -\sum_{z=2}^{Z} D_{\text{KL1}}(q(\mathcal{F}_{z-1}|\mathcal{F}_z, \mathcal{F}_0)||p_\theta(\mathcal{F}_{z-1}|\mathcal{F}_z, \mathbb{C})) \\ &\quad - D_{\text{KL2}}(q(\mathcal{F}_Z \mid \mathcal{F}_0) \parallel p(\mathcal{F}_Z)) + \log p_\theta(\mathcal{F}_0 \mid \mathcal{F}_1, \mathbb{C}). \end{aligned} \tag{7}$$

Item D_{KL2} contains no training parameters, which can be ignored in the subsequent calculation process. $\log p_\theta(\mathcal{F}_0 \mid \mathcal{F}_1, \mathbb{C})$ can be regarded as a particular case of D_{KL1} when $z = 1$ and merged into D_{KL1}. The $q(\mathcal{F}_{z-1} \mid \mathcal{F}_z, \mathcal{F}_0)$ in D_{KL1} represents the posterior probability distribution in the diffusion process. According to $\mathcal{F}_0 = (\mathcal{F}_z - \epsilon\sqrt{1-\bar{\gamma}_z})/\sqrt{\bar{\gamma}_z}$ deduced from (4) and Bayes' rule, Ho et al. [12] proved that the mean of $q(\mathcal{F}_{z-1} \mid \mathcal{F}_z, \mathcal{F}_0)$ could also be described by Gaussian distribution as follows:

$$\tilde{\mu}_z(\mathcal{F}_z, \mathcal{F}_0) = \frac{1}{\sqrt{\gamma_z}}\left(\mathcal{F}_z(\mathcal{F}_0, \epsilon) - \frac{\omega_z}{\sqrt{1-\bar{\gamma}_z}}\epsilon\right). \tag{8}$$

Both the q-distribution and the p_θ-distribution are Gaussian distributions. According to the KL divergence formula for the Gaussian distribution, $\text{KL}(\mathcal{N}(\mu_1, \sigma_1^2) \parallel \mathcal{N}(\mu_2, \sigma_2^2)) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$, the term D_{KL1} as the objective function to be trained is parsed as:

$$\mathbb{E}_q\left[\frac{1}{2\sigma_\theta} \parallel \tilde{\mu}_z(\mathcal{F}_z, \mathcal{F}_0) - \mu_\theta(\tilde{\mathcal{F}}_z, z, \mathbb{C})\|^2\right] + C. \tag{9}$$

Convert the mean matching to noise matching of varying levels added during diffusion and reweight the objective function towards obtaining more stable gradient updates. The final loss function to be trained is formulated as:

$$\mathcal{L}_{DN} = \mathbb{E}_{\mathcal{F}_0\sim q(\mathcal{F}_0), \epsilon\sim\mathcal{N}(0,\text{I}), z\sim\mathcal{U}(0,Z)}[\eta(z)\|\epsilon - \epsilon_\theta(\mathcal{F}_z, z, \mathbb{C})\|^2], \tag{10}$$

where the perturbation variable $\mathcal{F}_z$ is resolved by (4), and $\mathbb{C}$ has a 10% probability of being a zero vector. We employ the classifier-free guidance technique [40] to trade off mode coverage and sample quality by adjusting the weight ∂:

$$\tilde{\epsilon}_z = (1+\partial)\epsilon_\theta(\tilde{\mathcal{F}}_z, z, \mathbb{C}) - \partial\epsilon_\theta(\tilde{\mathcal{F}}_z, z). \tag{11}$$

Utilizing $\mu_\theta(\tilde{\mathcal{F}}_z, z, \mathbb{C})$ with the reverse diffusion process defined in (5) & (6), we are able to obtain the denoised distribution after sampling an arbitrary steps on the prior distribution by reparameterization technique:

$$\tilde{\mathcal{F}}_{z-1} = \frac{1}{\sqrt{\gamma_z}}\left(\tilde{\mathcal{F}}_z - \frac{\omega_z}{\sqrt{1-\bar{\gamma}_z}}\tilde{\epsilon}_z\right) + \sqrt{\omega_z}\epsilon, \tag{12}$$

where ϵ is a standard Gaussian distribution but takes zero when sampled to $\tilde{\mathcal{F}}_0$. We introduce the concept of Hausdorff distance (HD) [41], widely applied to evaluate image segmentation methods, to constrain the reverse sampling process within ROI, which can be calculated by:

$$hd(\tilde{\mathcal{F}}_z, \text{ROI}) = \max_{f\in\tilde{\mathcal{F}}_z} \min_{a\in\text{ROI}} \parallel f - a\|_2. \tag{13}$$

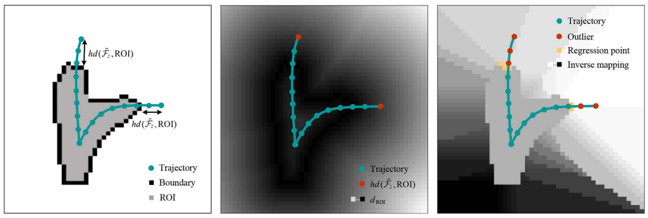
Employing HD directly as a loss function is intractable and leads to large computations as well as unstable training. Karimi et al. [42] successfully reduced the HD loss by a distance transform method of segmentation boundaries. We extend this concept naturally to distance transformations between trajectories and ROI with appropriate refinements:

$$\mathcal{L}_{HD}(\tilde{\mathcal{F}}_z, \text{ROI}) = \frac{1}{\Omega}\sum((\tilde{\mathcal{F}}_z \ominus \text{ROI}) \circ d_{\text{ROI}}^\beta), \tag{14}$$

where the symbol $\ominus$ denotes the difference set between sets $\tilde{\mathcal{F}}_z$ and ROI, the symbol $\circ$ denotes the Hadamard product, and Ω denotes the number of outliers. β determines the penalty degree for out-of-region trajectories, which means we systematically correct all predictions that violate the ROI rule, rather than focusing only on the maximum prediction error defined by (13). d_{ROI} is the dimensionless distance from the trajectory outside the region to the boundary. However, continuous trajectories need to be filled into the rasterized map with integer coordinate forms, which is a non-differentiable discrete process. To solve this problem, we transfer the computational graph to an inverse mapping of ROI, which is then mapped to the dimension of $\tilde{\mathcal{F}}_z$ through the query way. Finally, we can constrain $\tilde{\mathcal{F}}_z$ within ROI through regression. The modified Hausdorff distance (MHD) loss can be given by:

$$\mathcal{L}_{MHD}(\tilde{\mathcal{F}}_z, \overleftrightarrow{\text{ROI}}) = \mathcal{L}_{reg}(\overleftrightarrow{\text{ROI}}(\tilde{\mathcal{F}}_z), \tilde{\mathcal{F}}_z \circ \text{mask}), \tag{15}$$

where $\overleftrightarrow{\text{ROI}}$ denotes the inverse mapping of ROI, i.e., the map outside ROI is filled with boundary coordinates with d_{ROI} distance from the ROI boundary, and the map within ROI is loaded with zero. The symbol mask is a matrix with one as the outlier and zero as the internal points. Please refer to Fig. 3



(a) (b) (c)

Fig. 3. Illustration of the ROI distance transformation and inverse mapping processes. (a) The background is the predicted rasterized map of ROI. The foreground is HD representations $hd(\tilde{\mathcal{F}}_z, \mathrm{ROI})$ between the samples and ROI. (b) The background is a distance-transformed map of the ROI boundary, where each pixel value equals its shortest distance to the boundary. The foreground is the corresponding mapping of the samples on the map. The maximum d_{ROI} value outside ROI is indicated by red points, corresponding to $hd(\tilde{\mathcal{F}}_z, \mathrm{ROI})$. (c) The background is an inverse mapping of the ROI, and each pixel outside the ROI stores the boundary coordinates with the d_{ROI} distance from the ROI boundary. Here, we visualize the sum of the coordinates in each pixel. Outliers will be regressed to the ROI boundary.

for detailed descriptions of ROI distance transformation and inverse mapping. Once ROI prediction network is trained, the ROI inverse mapping only requires to be computed once. Then, the regression coordinates can be derived employing the query method, significantly reducing calculation costs. In summary, the final objective function for training the reverse diffusion process is defined as follows:

$$\mathcal{L}_{ID} = \mathcal{L}_{DN}(\mathcal{F}_z, z, \mathbb{C}) + \lambda \mathcal{L}_{MHD}(\tilde{\mathcal{F}}_z, \overset{\leftrightarrow}{\mathrm{ROI}}). \quad (16)$$

Algorithm 1 summarizes the core training process of RoadDiff. As illustrated by the reverse diffusion process in Fig. 2, once ϵ_θ training is completed, the reverse sampling process starts from the prior distribution $\tilde{\mathcal{F}}_Z \sim \mathcal{N}(0, I)$. Following Z iterations of (12), noise is gradually removed to generate the predicted trajectory that conforms to ROI constraint.

IV. Proposed Model

This section describes the specific implementation of the critical components in RoadDiff, namely the scene context representation, the denoising matching function ϵ_θ, and ROI prediction decoder.

A. Scene Context Representation

High-quality conditional representations are essential for the controllable generation of DPMs. High-definition maps containing extensive semantic information guiding and limiting vehicle movements and the observed trajectories of road participants are highly coupled to the future trajectory of the target agent. Fig. 4 gives a detailed unfolding of the social space in Fig. 2. We model maps and trajectory information through representations of three distinct attributes. For undifferentiated training, we rotate the interaction region to establish a coordinate system centered on the target agent. ROI and the denoising matching function selectively apply these heterogeneous representations to further generate high-quality cross-modal fusion representations through the complementation and interaction between high-dimensional information.

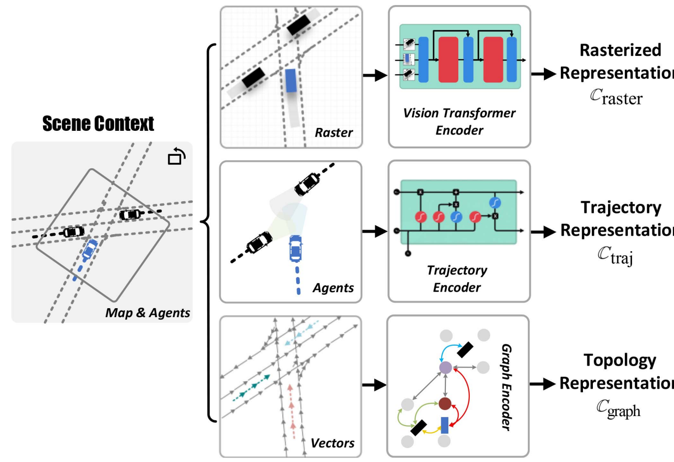


Fig. 4. Heterogeneous representations of scene contexts.

1) Rasterized Representation of Scene Context: The rasterized representation allows for explicitly modeling the interactions and positional relationships between agents and maps. We render dynamic contextual scenes of the target agent within T_{obs} of the past period, including static semantic maps, dynamic agent behaviors, etc., as a set of continuous multi-dimensional images. The size of each image frame is $(288, 288)$, where each raster represents a realistic scene of $0.5 \times 0.5\ m^2$ area, i.e., it covers a $144 \times 144\ m^2$ area of interaction centered on the target agent. Vision Transformer (ViT) [43] boasts remarkable performance in cross-modal representation learning due to its excellent computational efficiency and extensibility [44]. As in the standard ViT, our encoder embeds the chunked raster map through linear projections with positional embedding information, yielding $\mathbb{C}_{\mathrm{raster}}$ following a series of Transformer Blocks.

2) Observed Trajectory Representation: We represent the observed trajectories of road participants with lightweight LSTM encoders to better exploit their interactions in downstream tasks. Specifically, we select surrounding agents within a 72 m radius of the target agent as the critical factors influencing its motions, reducing the redundancy of parameters and computational complexity. The target agent is encoded independently from the surrounding agents but shares the identical structure for representation, including an LSTM network (LS) and a fully connected layer (FC). The network output utilizes layer normalization (LN) to ensure the stability of the data feature distribution, and ultimately the non-linear representation of the encoder is enhanced by the ELU activation function σ. Described by the formula as:

$$\mathbb{C}_{\mathrm{traj}} = \sigma(LN(FC(LS(\mathcal{H})))). \quad (17)$$

3) Topological Graph Representation of Scene Context: The advantage of the graph topology over the rasterized map is more sensitive to the connectivity between lanes and the extension tendency of agent motion. We vectorize the dynamic actions and scene contexts of agents, followed by aggregating the node features of individual polyline subgraphs at the vector level utilizing hierarchical graph networks [45]. Global aggregation of trajectories and structured maps is then performed by means of the graph convolutional network (GCN):

$$\{g_i^{(\ell+1)}\} = GCN(\{g_i^{(\ell)}\}, Adjacency), \quad (18)$$

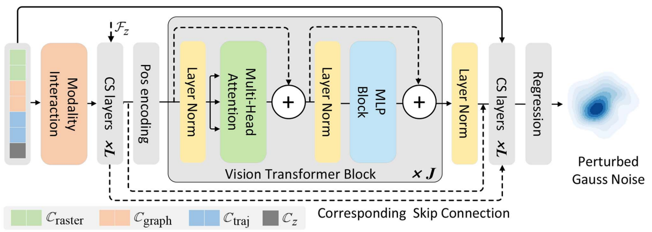


Fig. 5. Overview of the network fitted to the denoising matching function ϵ_θ.

where $\{g_i^{(\ell)}\}$ and $Adjacency$ correspond to the features and adjacency matrix of the polyline node set. We model the high-order interactions employing multi-head attention calculations:

$$\mathbb{C}_{\text{graph}} = MultiHead(LN(FC(\{g_i^{(\ell+v)}\}))), \tag{19}$$

where v denotes the number of layers of GCN networks.

B. Denoising Matching Function

The diffusion model relies on the solid fitting ability of the neural network to approximate the denoising matching function ϵ_θ. Transformer has a natural potential for capturing temporal correlation and dependent coupling between inputs and outputs due to its theoretical signal transmission distance of $O(1)$. Hence, we follow MID with Transformer as the core for the network structure of the denoising matching function.

Fig. 5 gives a detailed unfolding of the denoising matching function ϵ_θ in Fig. 2. The diffusion step encoding $\mathbb{C}_z$ is employed to indicate the execution progression of the sampling process. The cross-modal heterogeneous representations with embedded diffusion step features are fed into the model, guaranteeing that RoadDiff extends trajectory diversity through coarse sampling of the prior distribution in the early stages and further improves trajectory generation accuracy through fine-grained noise in the later stages. Unlike pre-training methods that depend heavily on the feature extraction process, our modal interaction strategy focuses on the lightweight and rapid embedding of various modalities. Specifically, we employ a combination of self-attention computation and fully-connected layers as the modality interaction module, modeling optional inter-modal and intra-modal interactions. Contributions of cross-modal information are adaptively distinguished, and more modalities could be integrated than cross-attention. Modal interaction representations and perturbation variables $\mathcal{F}_z$ are fused and dimensionally enhanced through a series of concat squash (CS) [46], [47] operations:

$$X_{\ell+1} = (X_\ell W_1 + b_1) \circ \sigma(\mathbb{C}W_2 + b_2) + \mathbb{C}W_3, \tag{20}$$

where W_i and b_i are weights and bias vectors to control output dimensions. X_ℓ and $X_{\ell+1}$ are inputs and outputs of the concat squash layer, and X_0 is the perturbation variable $\mathcal{F}_z$. σ denotes the sigmoid activation function. Latent dependencies and spatio-temporal cues between inputs and outputs are efficiently learned through a series of ViT blocks in high-dimensional spaces. Correspondingly, the outputs are gradually mapped to the noise dimension through the symmetrical CS dimension reduction operations and a regression layer.

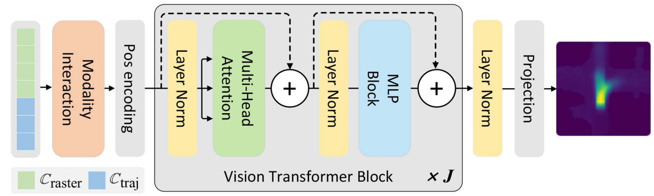


Fig. 6. Overview of the ROI decoder $\boldsymbol{r}_\theta$. Legend shared with Fig. 5.

C. Region of Interest Prediction

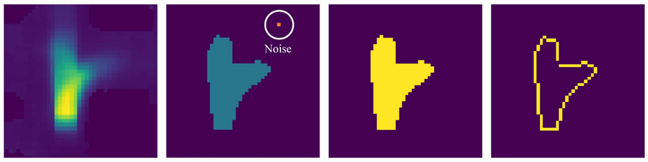
The motion dataset provides the drivable area of the vehicle, which is large and discontinuous. Therefore, we further narrow down driving range of the vehicle from the drivable area to ROI, i.e., the possible activity region of the vehicle for a period of time T_{pred} in the future. Fig. 6 provides an overview of our proposed ROI decoder $\boldsymbol{r}_\theta$. The trajectory representation $\mathbb{C}_{\text{traj}}$ is merged into the rasterized representation $\mathbb{C}_{\text{raster}}$ through the modality interaction module. We employ standard 1-D position embedding with gradient updates, and another series of stacked ViT blocks performing decoding operations. The projection layer combines a fully connected layer and a sigmoid activation function that aligns the output dimension with the ROI patches.

ROI constraining motion trajectories is multimodal due to the complexities of driving scenarios and agent interactions, which presents two unbalanced challenges. ROI (foreground) is much smaller than the background region, resulting in gradients being concentrated in the background region, which causes an imbalance in the distribution of the gradient norm. This phenomenon, which causes poor predictions of foreground areas, is referred to as the positive and negative sample imbalance. The agent intentions are unevenly distributed, e.g., most agents on regular roads are keeping straight, with individual agents choosing to change lanes. The model tends to optimize towards the direction with high frequency of intentions, as this leads to an overall gradient update gain [48]. Minority modes are neglected to the detriment of multimodal prediction, referred to as mode imbalance. Inspired by the focal loss [49], we design a diversity loss function for ROI prediction to compensate for both imbalances:

$$\mathcal{L}_{ROI} = \frac{\wp}{S} \cdot \sum_s (P_s - \hat{P}_s)^2 \cdot \begin{cases} -\alpha(1-\hat{P}_s)^\delta \log(\hat{P}_s) & \text{if } P_s > 0 \\ -(1-\alpha)\hat{P}_s^\delta \log(1-\hat{P}_s) & \text{otherwise} \end{cases}, \tag{21}$$

where the weight modulation hyperparameters α and δ regulate sample and mode imbalances, respectively. S is the pixel count of the output image. P_s represents the ground truth of the probabilistic raster map, which is the image of the trajectory mapped onto the rasterized map after standard Gaussian convolution. The range of ROI is modulated by the hyperparameters of the Gaussian convolution. $\hat{P}_s$ is the prediction corresponding to P_s. $\wp$ is the mask matrix representing the drivable area attribute in the high-definition map with the same size as P_s, which avoids unnecessary gradient calculations.

Please refer to Fig. 7 for an overview of our process for robustly extracting ROI and its boundaries from the predicted probabilistic raster map via the disjoint set algorithm. The pre-processing stage converts a probabilistic raster map in Fig. 7(a) into a binarized image. Iterate over all valued pixels in the



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

Fig. 7. Schematic illustration of the process to extract ROI boundaries via the disjoint set algorithm. (a) The predicted probabilistic raster map. (b) Connected regions (or noise) found by the disjoint set algorithm. (c) Region of interest after noise removal. (d) The extracted boundaries.

binarized image. If the pixel values in the historical neighborhood are all zero, the location is given a new label. Otherwise, if there are various labels in the historical neighborhood, the location is marked as the smallest label, and all the labels are connected utilizing the disjoint set. Next, iterate through all non-zero labels. Utilizing the find method of constructed disjoint sets, we update each label with a constant time complexity in the connected domain to the value corresponding to the root node, i.e., the smallest label in the set. Fig. 7(b) shows the final connected regions found. As shown in Fig. 7(c), we filter the anomalous regions or noise and take the connected region with the highest probability as the region of interest. Therefore, our method has an anti-interference ability and can automatically shield abnormal prediction regions. Fig. 7(d) shows the contours extracted for the ROI inverse mapping algorithm.

D. RoadDiff Training and Complexity Analysis

The main steps of the training process for the proposed RoadDiff are shown in Algorithm 1. We first give the time complexity of basic components in RoadDiff. The time complexity of multi-head attention can be computed as $\kappa_{\text{head}}(s_{\text{in}}, e_{\text{emd}}) = \sum_h (4s_{\text{in}}e_{\text{emd}}^2 + 2s_{\text{in}}^2 e_{\text{emd}})$, where s_{in} and e_{emd} denote the input sequence length and embedding dimension, respectively, and h represents the number of attention heads. The time complexity of ViT modules can be computed as $\kappa_{\text{vit}}(s_{\text{in}}, e_{\text{emd}}) = \sum_J (\kappa_{\text{head}}(s_{\text{in}}, e_{\text{emd}}) + 2e_{\text{emd}}e_{\text{ff}})$, where e_{ff} represents the hidden dimension of the feedforward network and J denotes the number of encoder blocks. We simply compute the time complexity of CS layers as $\kappa_{\text{cs}}(e_{\text{emd}}, e_{\text{exp}}) = 2\sum_L (re_{\text{exp}} + 2e_{\text{emd}}e_{\text{exp}})$, where r denotes the dimension of the inputs, e_{exp} denotes the hidden dimension, and e_{emd} denotes the embedding dimension of condition $\mathbb{C}_{\theta 2}$.

The HD map features and agent trajectories involved in each scene context are extracted after translating and rotating the interaction region (Line 1). These different forms of information are constructed as trajectory representations, raster representations, and topology representations, as shown in Fig. 4, followed by building them as heterogeneous representation wrappers to give compatible cross-modal fusion representation schemes for different downstream tasks (Lines 2∼3). The time complexity of $\mathbb{C}_{\text{traj}}$ is $O(\sum_m (T_{\text{obs}}(c_{\text{in}}e_{\text{hid}} + e_{\text{hid}}^2) + e_{\text{hid}}c_{\text{out}}))$, where c_{in} denotes the input feature dimension, e_{hid} denotes the hidden dimension of LSTM, and c_{out} means the output dimension of the fully connected layer. The time complexity of $\mathbb{C}_{\text{graph}}$ is linear in the number of edges E, which can be

---

Algorithm 1: Model Training Process with RoadDiff.

Input: Observed scene contexts $\{\mathcal{H}, \mathcal{M}\}$;
Ground truth future trajectories $\mathcal{F}_0$
Output: Trained network weights of $\mathbb{C}_{\theta 1}$, $\mathbb{C}_{\theta 2}$, $\boldsymbol{r}_\theta$, and $\boldsymbol{\epsilon}_\theta$

1: Preprocess HD map and trajectories in the scene contexts
2: Construct representations of trajectory $\mathbb{C}_{\text{traj}}$ by (17), raster $\mathbb{C}_{\text{raster}}$ by ViT blocks, and topology $\mathbb{C}_{\text{graph}}$ by (19)
3: Build heterogeneous representation wrappers $\mathbb{C}_{\theta 1}$ and $\mathbb{C}_{\theta 2}$
4: Set model training and diffusion process hyperparameters
5: Initialize the network weights of $\mathbb{C}_{\theta 1}$, $\mathbb{C}_{\theta 2}$, $\boldsymbol{r}_\theta$, and $\boldsymbol{\epsilon}_\theta$
6: **while** No early stopping **do**
7: **repeat**
8: Take $\mathbb{C}_{\theta 1}$ as the input to ROI decoder $\boldsymbol{r}_\theta$ in Fig. 6
9: $Generate$ $\hat{P}_s$ through the forward propagation
10: $Compute$ the loss function $\mathcal{L}_{ROI}(P_s, \hat{P}_s)$ in (21)
11: Take gradient step on $\nabla_{\boldsymbol{\theta}}(\mathcal{L}_{ROI})$
12: $\mathbb{C}_{\theta 1}, \boldsymbol{r}_\theta \leftarrow Backward(\nabla_{\boldsymbol{\theta}})$
13: **until** Model convergence
14: **end while**
15: Extract ROI from $\hat{P}_s$ by disjoint set algorithm in Fig. 7
16: Build the inverse mapping $\overleftrightarrow{\text{ROI}}$ via the process in Fig. 3
17: **while** No early stopping **do**
18: **repeat**
19: Take $\mathbb{C}_{\theta 2}$ as the input to $\boldsymbol{\epsilon}_\theta$ in Fig. 5
20: $\mathcal{F}_0, z \leftarrow Sample(q_{\text{data}}(\mathcal{F}_0)), \text{Uniform}(\{1, \ldots, Z\})$
21: $Calculate$ $\mathcal{F}_z$ by (4) and $\tilde{\mathcal{F}}_z$ by (12)
22: $Generate$ noise through the forward propagation
23: $Compute$ loss functions by (10) and Eq.(15):
$\mathcal{L}_{DN} = \eta(z) \parallel \epsilon - \boldsymbol{\epsilon}_\theta(\mathcal{F}_z, z, \mathbb{C}_{\theta 2})\|^2$
$\mathcal{L}_{MHD} = \mathcal{L}_{reg}(\overleftrightarrow{\text{ROI}}(\tilde{\mathcal{F}}_z), \tilde{\mathcal{F}}_z \circ \text{mask})$
24: Take gradient step on $\nabla_{\boldsymbol{\theta}}(\mathcal{L}_{DN} + \mathcal{L}_{MHD})$
25: $\mathbb{C}_{\theta 2}, \boldsymbol{\epsilon}_\theta \leftarrow Backward(\nabla_{\boldsymbol{\theta}})$
26: **until** Model convergence
27: **end while**

---

denoted as $O(v|E| + \kappa_{\text{head}}(s_{\text{in}}, e_{\text{emd}}))$. $\mathbb{C}_{\text{raster}}$ has time complexity $O(\sum_{s_{\text{in}}} u^2 w^2 e_{\text{in}} e_{\text{emd}} + \kappa_{\text{vit}}(s_{\text{in}}, e_{\text{emd}}))$, where u denotes the size of output feature maps. w, e_{in}, and e_{emd} indicate the convolution kernel size, and the number of input and output channels, respectively. Here s_{in} is the number of patches in the rasterized map. In the initialization stage of models (Lines 4∼5), we initialize the hyperparameters for model training, including batch size, number of data-loading subprocesses, and optimizer settings. We initialize non-training parameters of the diffusion process, including the variance schedule ω_z, Markov chain length Z, and γ_z. The weights of $\mathbb{C}_{\theta 1}$, $\mathbb{C}_{\theta 2}$, $\boldsymbol{r}_\theta$, and $\boldsymbol{\epsilon}_\theta$ are also randomly initialized, where $\mathbb{C}_{\theta 1} = \{\mathbb{C}_{\text{traj}}, \mathbb{C}_{\text{raster}}\}$ and $\mathbb{C}_{\theta 2} = \{\mathbb{C}_{\text{traj}}, \mathbb{C}_{\text{raster}}, \mathbb{C}_{\text{graph}}, \mathbb{C}_z\}$. $\boldsymbol{r}_\theta$ denotes the ROI decoder, and $\boldsymbol{\epsilon}_\theta$ denotes the denoising matching function.

The ROI decoder employs a dynamic learning rate combined with an early-stopping training strategy (Lines 6 ∼ 14). It takes

$\mathbb{C}_{\theta 1}$ as input and generates the probabilistic raster map $\hat{P}_s$ by forward propagation with the time complexity $O(\kappa_{\text{head}}(s_{\text{in}}, e_{\text{emd}}) + \kappa_{\text{cs}}(e_{\text{emd}}, e_{\text{exp}}) + \kappa_{\text{vit}}(s_{\text{in}}, e_{\text{exp}}))$ (Lines 8 $\sim$ 9). Gradients are computed as the partial derivatives of the ROI diversity loss function $\mathcal{L}_{ROI}$ with respect to model parameters (Lines 10 $\sim$ 11). Adam optimizer backpropagates loss function gradients to each layer of models through the chain rule to progressively update network weights of $\mathbb{C}_{\theta 1}$ and $\boldsymbol{r}_\theta$ (Line 12). Finally, the ROI decoder will be kept training until it converges. ROI and its boundaries are extracted from the probabilistic raster map $\hat{P}_s$ by the disjoint set algorithm after the ROI prediction (Line 15). We build the ROI inverse mapping $\overset{\leftrightarrow}{\text{ROI}}$ to serve the trajectory generation (Line 16).

The denoising matching function $\boldsymbol{\epsilon}_\theta$ takes $\mathbb{C}_{\theta 2}$ as input (Line 19). We obtain training samples from the trajectory dataset (Line 20), where $q_{\text{data}}(\mathcal{F}_0)$ denotes the original unknown distribution of the dataset, and $\mathcal{F}_0$ denotes a randomly selected sample from the dataset. z obeys the uniform distribution, which is randomly selected from the variance schedule $\{\omega_z \in (0, \tau)\}_{z=1}^{Z}$. The noise added at each step is predicted by the forward propagation with the time complexity $O(\kappa_{\text{head}}(s_{\text{in}}, e_{\text{emd}}) + \kappa_{\text{vit}}(s_{\text{in}}, e_{\text{emd}}))$, followed by the calculation of $\mathcal{L}_{DN}$ and $\mathcal{L}_{MHD}$ loss functions (Lines 21 $\sim$ 23). Gradient updating and backpropagation are done by Adam optimizer until the model converges (Lines 24 $\sim$ 27).

V. Experimental Evaluations

A. Experimental Settings

Extensive experiments have been performed on real-world datasets from various perspectives.

1) Dataset: The Argoverse dataset [50] tracks more than 300,000 driving scenes in Miami and Pittsburgh over a distance of 290 kilometers. Each sample consists of 2 s of historical states and 3 s of ground truth future trajectory, sampled at 10 Hz. The nuScenes dataset [51] is a large-scale public autonomous driving dataset with support for multimodal perception developed by the Motional team, containing 1000 complex driving scenarios from Boston and Singapore. Each scene lasts 20 s, and keyframes are annotated at 2 Hz. Following the existing work [24], [28], we conduct evaluations utilizing the Argoverse and nuScenes validation sets.

2) Evaluation Criteria: Average displacement error (ADE) and final displacement error (FDE) are two specialized measures of prediction accuracy that are widely used to evaluate the performance of various trajectory prediction models:

$$ADE = \frac{1}{T_{\text{pred}}} \sum\nolimits_{t=1}^{T_{\text{pred}}} \parallel \tilde{\mathcal{F}}_0^t - \mathcal{F}_0^t \parallel,$$
$$FDE = \parallel \tilde{\mathcal{F}}_0^{T_{\text{pred}}} - \mathcal{F}_0^{T_{\text{pred}}} \parallel . \tag{22}$$

We estimate the accuracy of multimodal trajectory predictions by applying the following four metrics: (1) Minimum final displacement error (minFDE): the minimum endpoint L_2 distance between the ground truth and predicted trajectory. (2) Minimum average displacement error (minADE): the average L_2 distance between the minFDE trajectory and the ground truth over all time steps. (3) Miss rate (MR): percentage of scenes where none of the FDEs of the predicted trajectories are within 2.0 meters of the ground truth. (4) Drivable area compliance (DAC): $(A - G)/A$, where A and G denote the number of predicted trajectories generated by the model and those that are not exactly in the drivable area, respectively.

3) Implementation Details: The batch size and the number of data-loading subprocesses are set to 128 and 8, respectively. The weight modulation hyperparameters of the diversity loss function in (21) are set to $\alpha = 0.75$ and $\delta = 2$. The hidden dimension of the LSTM network is 64 and the output dimension of the fully connected layer is 128 in $\mathbb{C}_{\text{traj}}$. The number of GCN layers υ and attention heads are 4 and 14 in $\mathbb{C}_{\text{graph}}$, respectively. The Gaussian convolution kernel size for modulating the ROI range is 6 and the variance is 3. The modality interaction module has four attention heads, and the fully connected layer maps the output dimension to 256. In the denoising matching function ϵ_θ, the number of CS layers $L = 2$, whose hidden dimension is 512. The number of attention heads is 8, the vector representation dimension is 512, and the hidden dimension of the feed-forward network is 2048. We employ gradient clipping techniques to restrict the max norm for the gradients of $\boldsymbol{r}_\theta$ and $\boldsymbol{\epsilon}_\theta$ networks to 1.0, maintaining the stability of the model parameter updates.

B. Comparison With State-of-The-Art

1) Baseline Methods: We compare models based on different modeling mechanisms from recent years to more fully validate the effectiveness of RoadDiff for trajectory prediction.

Methods evaluated on Argoverse: 1) CSDI [52]: a masked self-supervised learning approach based on attention mechanism and DPMs. *2) TNC* [53]: an unsupervised representation learning approach designed to learn potential dynamics of non-smooth data via adaptive temporal neighborhood coding. *3) BTCL* [17]: a self-supervised bidirectional trajectory comparison learning model based on jointly constructed pretext tasks. *4) CS-LSTM* [26]: a LSTM encoder-decoder based model for multimodal trajectory prediction of vehicles with maneuver categories as proposals. *5) DESIRE* [25]: the method generates trajectories using CVAE, which are then ranked and refined by inverse optimal control RNN. *6) HOME* [31]: the method samples endpoint coordinates from the projected heat map to generate multimodal trajectories. *7) Social-GAN* [4]: a generative adversarial model that models interactions between agents through an introduced pooling module. *8) R2P2* [23]: the method balances trajectory diversity and accuracy by parameterizing the model distribution as the push forward of a simple underlying distribution under the simulation operator. *9) Social-CVAE* [24]: the method improves trajectory prediction accuracy by mitigating the social posterior collapse problem in the variational autoencoder for interaction modeling. *10) MID* [32]: a model that performs the trajectory prediction by modeling the reverse process of motion uncertainty diffusion.

Methods evaluated on nuScenes: 1) MTP [19]: a method that uses raster images to encode the surroundings and CNN models for trajectory prediction. *2) P2T* [20]: a maximum entropy inverse reinforcement learning and grid sampling based trajectory prediction approach. *3) CoverNet* [29]: an approach to solving the trajectory prediction problem by classifying a diverse set of trajectories. *4) LaPred* [30]: a multimodal trajectory prediction method based on instance-level lane information.

TABLE I
RESULTS OF DIFFERENT METHODS FOR VEHICLE TRAJECTORY PREDICTION ON THE ARGOVERSE DATASET

	Method	$K=1$			$K=6$			
		minADE↓	MR↓	minFDE↓	minADE↓	MR↓	minFDE↓	DAC↑
Pretrain	CSDI [52]	2.07	0.67	4.19	1.36	0.42	2.57	0.87
	TNC [53]	1.90	0.64	3.82	1.23	0.43	2.34	0.87
	BTCL [17]	1.61	0.59	3.55	1.12	0.34	2.13	0.91
Proposal	CS-LSTM [26]	2.79	0.69	5.33	1.51	0.52	2.92	0.70
	DESIRE [25]	2.38	-	4.64	1.09	-	1.89	-
	HOME [31]	1.64	$\underline{0.58}$	$\underline{3.50}$	$\underline{0.97}$	**0.15**	$\underline{1.71}$	$\underline{0.98}$
Probabilistic	Social-GAN [4]	2.34	0.71	5.18	1.49	0.44	2.91	0.78
	R2P2 [23]	3.02	-	5.41	1.40	-	2.35	-
	Social-CVAE [24]	$\underline{1.59}$	-	3.52	1.15	-	1.98	-
	MID [32]	1.67	$\underline{0.58}$	3.59	1.14	0.38	2.08	0.89
	RoadDiff (Ours)	$\mathbf{1.52}_{0.07\downarrow}$	$\mathbf{0.56}_{0.02\downarrow}$	$\mathbf{3.39}_{0.11\downarrow}$	$\mathbf{0.86}_{0.11\downarrow}$	$\underline{0.18}_{0.03\uparrow}$	$\mathbf{1.36}_{0.35\downarrow}$	$\mathbf{0.99}_{0.01\uparrow}$

5) Real-Time [28]: a heterogeneous agent trajectory prediction method for hierarchical modeling via global scene extraction and local agent learning. *6) ScePT* [9]: a scene-consistent trajectory prediction model based on policy planning. *7) Trajectron++* [7]: a generative trajectory prediction method coupled with dynamics and graph structure. *8) AgentFormer* [8]: an approach to simultaneously modeling the temporal and social dimensions of agent trajectories.

2) Results on Argoverse: Table I summarizes the results of evaluating multimodal trajectory predictions for all methods on the Argoverse dataset, selecting the most likely trajectories for $K=1$ and $K=6$. The best results for each metric are highlighted in bold and the suboptimal results are underlined.

RoadDiff achieves significantly better results compared to pre-training-based approaches. The CSDI, BTCL, and TNC methods obtain information-rich low-dimensional representations from the inherent potential dynamic structure of motion trajectories employing mask prediction or contrastive learning. Although these representations are more generalizable and robust, their specially tailored deployment patterns and loss functions limit the ability of the model to capture agent-map interactions and performance on individual supervised tasks. Instead of training cumbersome pre-training modules, our approach adopts learnable heterogeneous representations to simulate higher-order interactions between agents and maps.

RoadDiff outperforms the proposal-based multi-phase model in terms of composite metrics. CS-LSTM outputs multimodal trajectory distributions based on agent intention categories. Its prediction diversity is the same in each scenario, leading to low efficiency and prediction quality. DESIRE regressively refines the highly rated candidate trajectories through an iterative feedback mechanism. Nevertheless, the inference process encounters cumulative errors at various stages. RoadDiff directly models highly multimodal ground truth trajectory distributions in an end-to-end manner. Furthermore, it restricts the trajectories to ROI that strictly obeys road rules and physical constraints. Cumulative errors and unreasonable predicted trajectories are thus avoided. HOME predicts vehicle endpoints in the form of probabilistic heat maps to pursue the extreme MR ($K=6$) metric. Although it is ahead of us in this metric, the method is limited to the resolution of the heat maps. We outperform HOME on remaining metrics, which is sufficient to prove the competitiveness of our approach.

TABLE II
RESULTS OF DIFFERENT METHODS FOR VEHICLE TRAJECTORY PREDICTION ON THE nuScenes DATASET

	Method	$K=5$		$K=10$	
		minADE↓	minFDE↓	minADE↓	minFDE↓
Reg	MTP [19]	2.22	4.83	1.74	3.54
	P2T [20]	$\underline{1.51}$	**3.16**	1.25	2.35
Proposal	CoverNet [29]	2.54	5.08	1.97	3.39
	LaPred [30]	1.56	3.41	$\underline{1.23}$	2.39
	Real-Time [28]	1.60	3.34	$\underline{1.23}$	$\underline{2.32}$
Probabilistic	ScePT [9]	1.90	3.99	1.71	3.54
	Trajectron++ [7]	1.88	3.87	1.51	3.07
	AgentFormer [8]	1.86	3.89	1.45	2.86
	RoadDiff (Ours)	$\mathbf{1.47}_{0.04\downarrow}$	$\underline{3.23}_{0.07\uparrow}$	$\mathbf{1.12}_{0.11\downarrow}$	$\mathbf{2.29}_{0.03\downarrow}$

RoadDiff is more efficient in generating highly accurate motion trajectories than other methods based on probability distribution modeling, which is reflected in the overwhelming advantage of RoadDiff on different metrics. Social-CVAE alleviates the social posterior collapse phenomenon of conditional variational autoencoders but still has limitations in improving prediction ability. While Social-GAN mitigates mode collapse through the variety loss function, it ultimately learns to be uniformly distributed, i.e., generating high-diversity and low-quality predictions. MID explores the social interactions of agents while ignoring the constraining effect of lane graphs on motion trajectories. In this way, mode blur is easily triggered, limiting the high accuracy of trajectory prediction.

3) Results on nuScenes: Table II summarizes the results of evaluating all methods for multimodal trajectory prediction on the nuScenes dataset. K is taken to be 5 and 10 for a better comparison of multimodal prediction performance. We can observe that RoadDiff provides optimal trajectory inference performance with all metrics within acceptable limits.

RoadDiff outperforms MTP and P2T methods in comprehensive metrics. MTP focuses the predictions on main modes when calculating the probability distributions and has limitations in more modal predictions. P2T performs better in finite modes and slightly outperforms us in minFDE (K=5). We surpass it on the remaining metrics, especially with more modal predictions (K = 10).

TABLE III
SAMPLES GENERATED BY DPM-SOLVER WITH DIFFERENT NUMBER OF FUNCTION EVALUATIONS ON THE ARGOVERSE DATASET

NFE	5	10	20	50
minADE_6	0.93	0.90	0.88	0.87
minFDE_6	1.50	1.42	1.39	1.40
Real-Time (s)	0.25	0.47	0.93	2.16

RoadDiff outperforms proposal-based multimodal trajectory prediction methods. Compared to CoverNet, LaPred, and Real-Time, our model performance improves by an average of 63.5%, 6.5%, and 5.8% on four metrics, respectively. By addressing mode collapse and mode blur, RoadDiff can predict diverse and accurate trajectories more robustly in complex scenarios. CoverNet selects fixed modes of customized scenes from a set of static anchors, and the results are primarily limited to the spatial coverage and quality of the anchors. Although the Real-Time method enhances the scene adaptation of multimodal prediction results by constructing dynamic adaptive anchors, it is prone to suffer from the mode blur problem during anchor generations. LaPred predicts future trajectories with lane candidates as proposals. The complex behaviors of target agents are usually challenging to represent by single-lane instances fully. This coarse-grained modeling approach limits model performance in dealing with diverse behaviors, especially in complex or unstructured scenarios.

The prediction performance of RoadDiff beats the CVAE-based multimodal probabilistic generation models. Compared to ScePT, Trajectron++, and AgentFormer, our model performance improves by an average of 40.0%, 29.1%, and 25.3% on four metrics, respectively. ScePT introduces the method of conditional value at risk to avoid mode collapse. However, ScePT cannot exploit the sparsity of the interaction graph to consider only the neighbors of each node, which leads to ignoring the interactions between nodes in different cliques and decreasing accuracy in congested traffic scenarios. Trajectron++ employs an autoregressive decoder to generate trajectories. Predictions at each step depend on the results of the previous step and have limited effectiveness in handling long-time dependencies. AgentFormer can capture temporal and social dimensional dependencies but has limited ability to model complex map interactions at a fine-grained level. In addition, Trajectron++ and AgentFormer are prone to mode collapse when dealing with complex traffic scenarios, retaining only a few high-probability predictions, leading to a decrease in the diversity and accuracy of predictions.

4) Sampling Efficiency: Sampling from DPMs is slow as the reverse diffusion process involves iterating highly computationally expensive denoising matching networks. This process can be viewed as solving the corresponding diffusion ordinary differential equation (ODE). Our method supports any training-free accelerated DPM sampler. We select DPM-Solver [54], a diffusion ODE solver with convergence order guarantee, to observe the sampling speed of the existing trained DPMs. To fulfill the requirements of autonomous vehicles, we test the average real-time performance when predicting 32 agents on the Argoverse dataset. Table III records the inference results when sampling from DPMs with uniform time steps of 5, 10, 20, and 50 function evaluations. Specifically, the real-time consumed by trajectory sampling grows approximately linearly with the number of function evaluations (NFE). The DPM-Solver can converge to a satisfactory result in 5 to 20 function evaluations. Continuing to increase NFE optimizes partial performance but also brings several folds of time consumption. Generally, DPM-Solver allows our models to maintain an acceptable range of sampling efficiency and inference performance.

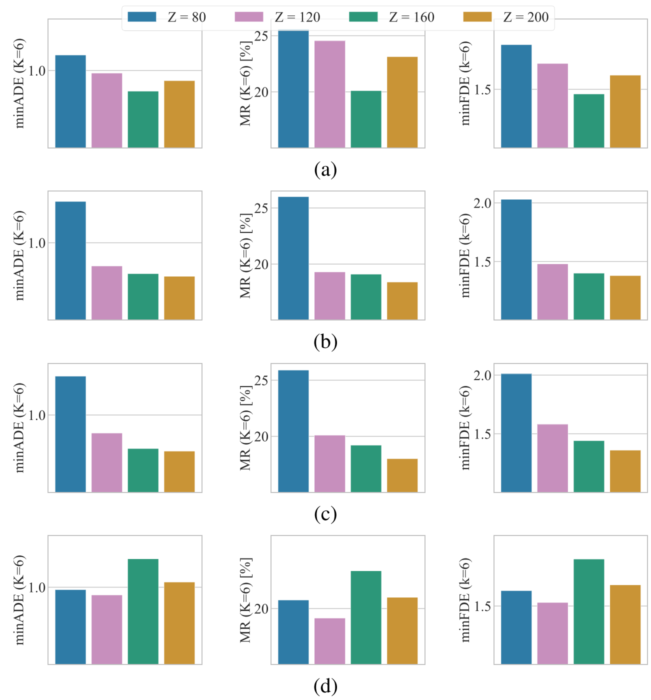


Fig. 8. Analytical results of two critical hyperparameters in RoadDiff. (a) Cosine schedule with $s = 0.008$. (b) Linear schedule with $\tau = 0.05$. (c) Linear schedule with $\tau = 0.1$. (d) Linear schedule with $\tau = 0.2$.

C. Parameter Sensitivity

We report the effects of two crucial hyperparameters in the diffusion process: (1) the Markov chain length Z and (2) the variance schedule ω on the performance of multimodal motion prediction ($K=6$). Fig. 8 illustrates the effect of the hyperparameters adjustment on performance under control variables, with lower bars representing better model performance.

We perform a grid search over the hyperparameters. In terms of variance schedules, performances of cosine variance schedules [55] and linear variance schedules [12] with three various configurations are reported. The cosine variance schedule gradually increases the noise addition variance ω_z at a non-linear rate throughout the diffusion process:

$$\omega_z = 1 - f(z)/f(z-1),$$
$$f(z) = \cos\left[(z/Z + s)/(1+s) \cdot (\pi/2)\right]^2, \qquad (23)$$

where the offset $s = 8e^{-3}$ is employed to prevent ω_z from being too small when close to $z = 0$. For the linear variance schedule, we set $\omega_1 = 1e^{-4}$, and ω_Z from the search space $\tau \in$ [0.05, 0.1, 0.2]. In terms of Markov chain length, we restrict the search

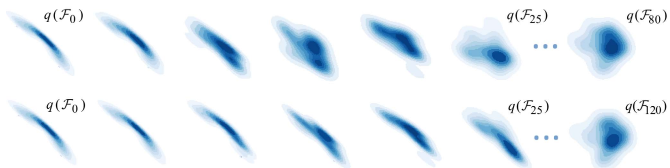


Fig. 9. Visualize kernel density distributions of sample diffusion processes when Z is taken to various values. Latent samples are from $Z = 80$ (top) and $Z = 120$ (bottom), respectively.

space to $Z \in [80, 120, 160, 200]$ based on practical experience from recent work [32].

Different parameter combinations coupled with each other jointly determine the quality and speed of the diffusion process. As indicated by the brown bars in Fig. 8(c), the parameter combination with the best model performance is $\tau = 0.1$ and $Z = 200$. All noise shares a uniform denoising matching function, while the non-linear noise addition makes it harder to align various noise levels. As shown in Fig. 8(a), the performance under the cosine variance schedule mode is much lower than that of the linear variance schedule. Noise is injected smoothly and evenly under the linear variance schedule, with performance influenced by noise granularity. As shown in Fig. 8(b), the slow diffusion process contains redundant steps when the noise granularity is overly fine ($\tau = 0.05$). The model has difficulty in learning valuable information with limited steps. As shown in Fig. 8(d), coarse-grained noise ($\tau = 0.2$) reduces the sampling accuracy of the reverse diffusion process, leading to a degradation of the trajectory generation performance.

As shown by blue bars in Fig. 8, we note that in most cases the model performance at $Z = 80$ is much lower than the performance under other parameters. To analyze the reason behind this phenomenon, we focus on visualizing the diffusion process in Fig. 9 at $Z = 80$ and $Z = 120$ when $\tau = 0.1$. It can be observed that at $Z = 80$, the original data distribution is quickly perturbed to a Gaussian-like distribution as the noise addition variance rapidly expands. Since these different magnitudes of noise share a denoising matching network, the model has difficulty adapting to this drastic evolutionary process. When $Z = 120$, the transition between modes is more natural and smooth, allowing the model to learn fine-grained noise, which brings an accuracy gain in trajectory generation.

D. Ablation Study

We conduct ablation experiments on RoadDiff with the Argoverse dataset to explore the impact of three vital components (1) denoising matching network, (2) heterogeneous representation, and (3) loss function for ROI prediction.

1) Denoising Matching Network: We replace three commonly used diffusion model backbone networks to approximate the denoising matching function ϵ_θ, respectively, verifying the effectiveness of our designed network. U-Net [56] consists of systolic paths that capture context and symmetric extended paths that support exact localization. We convert the Conv2d to Conv1d convolution in U-Net to better fit the trajectory data. Pointwise-Net [46] comprises stacked CS layers, and the network configuration recommended in the original paper is adopted.

TABLE IV
RESULTS OF AN ABLATION STUDY OF DENOISING MATCHING NETWORK

Alternative Model	$\text{minADE}_6\downarrow$	$\text{MR}_6\downarrow$	$\text{minFDE}_6\downarrow$	$\text{DAC}_6\uparrow$
U-Net	1.17	0.28	2.19	0.96
Pointwise-Net	1.49	0.35	2.88	0.90
Transformer	1.02	0.27	1.93	0.97
Ours	**0.86**	**0.18**	**1.36**	**0.99**

TABLE V
RESULTS OF AN ABLATION STUDY OF HETEROGENEOUS REPRESENTATIONS

Ablation Item	$\text{minADE}_6\downarrow$	$\text{MR}_6\downarrow$	$\text{minFDE}_6\downarrow$	$\text{DAC}_6\uparrow$
$\mathbb{C}_{\text{raster}}$	0.88	0.18	1.38	0.98
$\mathbb{C}_{\text{graph}}$	0.89	0.19	1.42	0.98
$\mathbb{C}_{\text{raster}}$ & $\mathbb{C}_{\text{graph}}$	0.93	0.20	1.49	0.98
No Ablation	**0.86**	**0.18**	**1.36**	**0.99**

The best results for each experiment are summarized in Table IV. Compared to the best performance in the comparison network backbone, RoadDiff leads in each metric by 29.46% (minFDE_6), 32.24% (MR_6), 15.93% (minADE_6), and 2.06% (DAC_6), respectively. Although U-Net possesses elegantly simple architecture and resilient deformation data augmentation, it does not consider how to learn temporal correlations, performing poorly in predicting motion trajectories with rich spatio-temporal information. The CS layers of Pointwise-Net provide well integrated representation of scene context and perturbation distribution $q(\mathcal{F}_z)$. However, it is a variant of the MLP, with a shallow network layer and without considering temporal correlation. Pointwise-Net thus performs the worst on all metrics. RoadDiff provides better performance than Transformer because we build on it by properly encoding heterogeneous representations into the latent space of RoadDiff through self-attention and upsampling operations. The role of conditional representations in the latent space is strengthened, and the problem of social posterior collapse [24] is avoided.

2) Heterogeneous Representation: We explore the impact of heterogeneous representations on RoadDiff performance by ablating the single scene contextual representation, respectively. As seen from the experimental results recorded in Table V, removing the heterogeneous representation from the RoadDiff results in considerable degradation. The performance of RoadDiff decreases by 9.77% (minFDE_6), 11.01% (MR_6), 8.22% (minADE_6), and 1.01% (DAC_6), respectively. The high-quality representation improves performance overall as the well-designed denoising network is carried through to each training step in the reverse diffusion process. Heterogeneous representations enrich the resources for scene contextual representation learning, allowing models to better understand the compatibility and higher-order interactions of agents and map semantics. The ablated graph topology representation $\mathbb{C}_{\text{graph}}$ degrades the model performance more significantly than the ablated raster representation $\mathbb{C}_{\text{raster}}$. This is attributed to the ability of the graph topology to more compactly represent the mapping structure between map-lane-agent and to be more sensitive in sensing the connectivity between lanes and the extensional tendency of agent motion. Additionally, the fusion of more heterogeneous

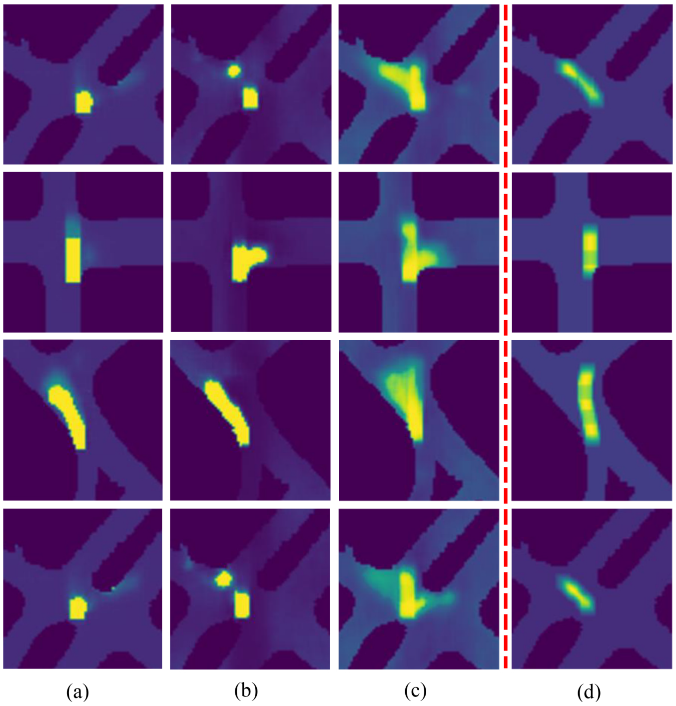


Fig. 10. Effectiveness of ROI prediction by training the model with different loss functions. (a) Binary cross-entropy loss function. (b) Regression loss function. (c) Our proposed ROI diversity loss function. (d) Ground truth.

information does not exacerbate performance gains. While heterogeneous representations enable complementary information, they also introduce redundancy and duplication of information.

3) Loss Function for ROI Prediction: We replace two commonly used loss functions to observe the prediction results, respectively. As shown in Fig. 10, compared to the binary cross-entropy and regression loss functions, our approach can capture the multimodal properties of the ROI, and the predictions do not miss the ground truth.

The inability to overcome mode imbalance is the main reason for the failed predictions of binary cross-entropy and regression loss functions. The frequency of various modes taken by agents in the same scenario may vary tens of times, while model gradients tend to move in the direction of more frequent modes that make the overall gradient update gain large. Modes with lower frequencies are swamped, leading to the phenomenon of mode collapse in ROI prediction. As shown in Fig. 10(a) and (b), the binary cross-entropy and regression loss functions can only focus predictions on the most frequent modes in the dataset, thus failing to capture multimodality. They miss ground truth that may be sparse instances and even present small or isolated prediction regions. Our method can modulate mode imbalance, which dynamically reduces the loss weights assigned to the numerous easy instances during training to focus learning efforts on difficult or sparse instances. As shown in the light-colored part in Fig. 10(c), our method captures the modes of the sparse instances so that the ground truth can be covered. In other words, if we do not modulate the mode imbalance, the lighter-colored predictions will disappear, leading to prediction failures similar to other loss functions. This phenomenon in turn demonstrates the effectiveness of our proposed diversity loss function. The success of multimodal ROI prediction allows us to proceed to the next stage of multimodal trajectory prediction robustly.

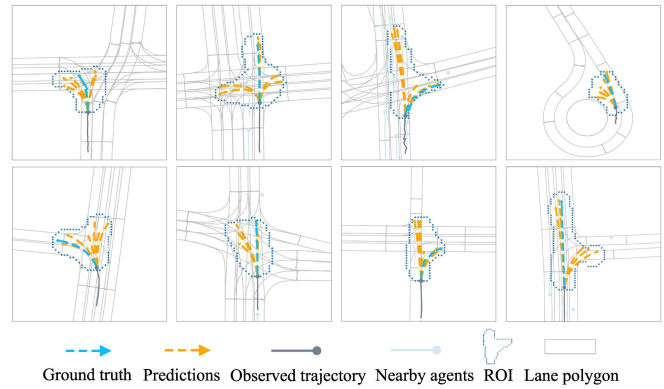


Fig. 11. Qualitative examples of multimodal trajectories generated by RoadDiff in several typical scenarios covering various modes.

E. Qualitative Analysis

We give qualitative examples in this section to demonstrate the performance of RoadDiff in typical scenarios and its effectiveness in improving mode blur and mode collapse.

1) Typical Scenarios: Typical examples include the driving modes of going straight, turning, acceleration and deceleration, and overtaking in complicated intersections and roundabouts scenarios. To visually reproduce the real-world situation as much as possible and to more clearly represent our work, we visualize the lane structure and the ROI contours in Fig. 11. The predicted ROI is an excellent reflection of the probable activity range of the agent in which the multimodal trajectories are successfully generated. Each mode is as likely to be taken by the agent in the current maneuvering state and the traffic environment, of which one or more predictions can cover the ground truth trajectory. These impressive results are attributed to the excellent ability of RoadDiff to capture the contextual interactions of the scene and the robustness of modeling multimodal motions.

2) Mode Blur: To avoid redundant and cumbersome comparisons, we chose HOME, the best performer in the comparison method, as our comparator to make our experiment more convincing. As observed in Fig. 12(a), HOME suffers from mode blur. Since HOME only restricts the endpoints to the drivable area and does not constrain the target conditional motion estimation phase. This results in the predicted trajectories being out of the drivable area on the path to the endpoints. As shown in Fig. 12(b), our method performs better for the identical scene. Different from the proposal-based prediction model typified by HOME, our approach provides a more robust ROI prediction strategy and constrains the entire trajectory generation process within it, avoiding mode blur.

Fig. 13(a) depicts the mode blur phenomenon of the HOME method due to anomalous endpoints caused by noise or other factors (unstable training, sampling method, etc.). As seen in Fig. 13(b), our method still has correct results in the identical scenario. ROI is the element with the highest probability in the set of connected regions generated by the disjoint set algorithm.

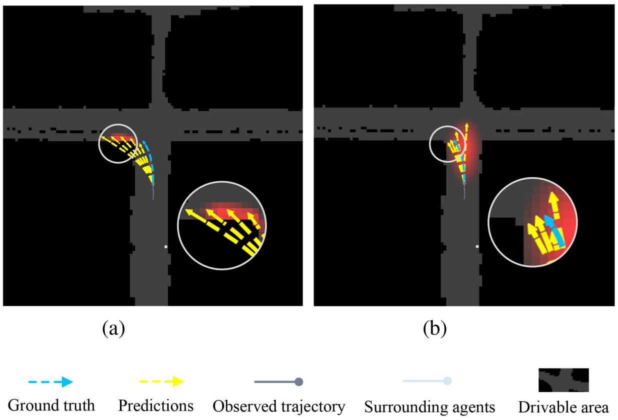


Fig. 12. The phenomenon of mode blur due to neglecting the trajectory constraints and the corresponding performance of our method. The probabilistic raster maps represent the endpoint regions (a) and ROI (b), respectively. (a) HOME. (b) Ours.

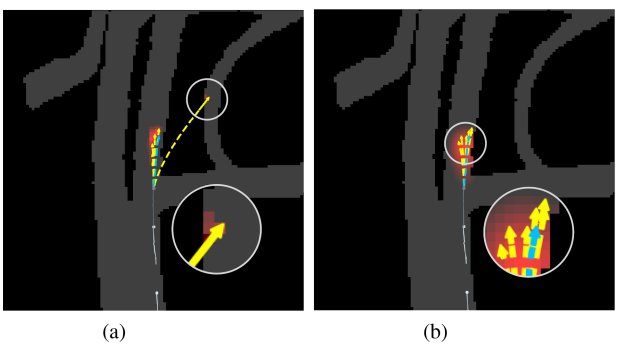


Fig. 13. The phenomenon of mode blur due to anomalous endpoints and the corresponding performance of our approach. Legend shared with Fig. 12. (a) HOME. (b) Ours.

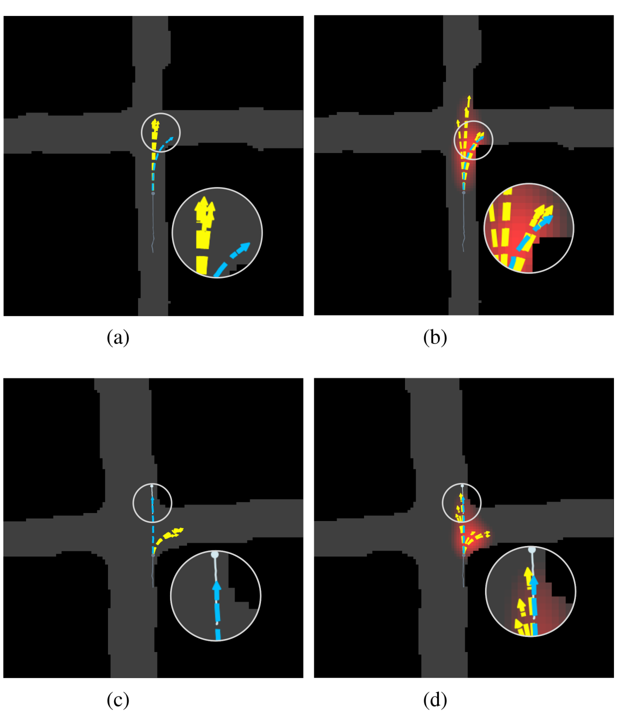


Fig. 14. The phenomenon of mode collapse in both scenarios (a) & (c) and the corresponding performances (b) & (d) of our approach. Legend shared with Fig. 12. (a) BTCL. (b) Ours. (c) BTCL. (d) Ours.

Thus, RoadDiff has the natural ability to resist interference, automatically masking anomalous prediction regions and avoiding incorrect prediction results.

3) Mode Collapse: To validate the strength of RoadDiff in mitigating mode collapse, the well performing BTCL model is selected for comparison. Fig. 14 shows the different performance of the two methods in the specific scenarios of turning right and going straight. Predictions of BTCL collapse on one mode and therefore miss the future entirely. Our method achieves multimodal predictions and successfully covers the ground truth trajectory. RoadDiff is able to preserve the semantic structure of the data, allowing a highly multimodal distribution of trajectories to be modeled directly and sampled accurately with useful information learned from contextual scenarios. Hence, the trajectories generated by RoadDiff are more referential than BTCL and can effectively facilitate the security of autonomous vehicle decisions.

VI. Conclusion

In this paper, we have proposed RoadDiff which enables the controllable generation of multimodal trajectories by leveraging beneficial information from scene-contextual heterogeneous representations. The ROI-based training strategy and specialized loss functions ensure stable and efficient distribution propagation during the reverse diffusion process.

CAV platooning can exploit RoadDiff to anticipate the cut-in maneuvers or collision hazards of non-platooning vehicles in mixed traffic flows, guaranteeing vehicle-following stability and proactive security through lead control. In future work, we plan to incorporate risk-sensitive reinforcement learning or safety optimization algorithms to improve the generalizability and robustness of models in complex or extreme traffic scenarios. In addition, we aim to conduct online interaction and closed-loop validation on CAVs to evaluate the practical performance of models under real-world conditions, including communication uncertainty, perception disturbances, and dynamic behavioral feedback.

References

[1] W. Yue, C. Li, P. Duan, and F. R. Yu, “Revolution on wheels: A survey on the positive and negative impacts of connected and automated vehicles in era of mixed autonomy,” *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21820–21835, Dec., 2023.

[2] R. Song and B. Li, “Surrounding vehicles' lane change maneuver prediction and detection for intelligent vehicles: A comprehensive review,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6046–6062, Jul., 2022.

[3] W. Yue, C. Li, S. Wang, N. Xue, and J. Wu, “Cooperative incident management in mixed traffic of CAVs and human-driven vehicles,” *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12462–12476, Nov., 2023.

[4] A. Gupta, J. Johnson, F. Li, S. Savarese, and A. Alahi, “Social GAN: Socially acceptable trajectories with generative adversarial networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2255–2264.

[5] D. Choi and K. Min, “Hierarchical latent structure for multi-modal vehicle trajectory forecasting,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 129–145.

[6] P. Kothari and A. Alahi, “Safety-compliant generative adversarial networks for human trajectory forecasting,” *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4251–4261, Apr., 2023.

[7] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 683–700.

[8] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, "AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9813–9823.

[9] Y. Chen, B. Ivanovic, and M. Pavone, "ScePT: Scene-consistent, policy-based trajectory predictions for planning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17103–17112.

[10] J. Lian, F. Yu, L. Li, and Y. Zhou, "Causal temporal–spatial pedestrian trajectory prediction with goal point estimation and contextual interaction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24499–24509, Dec., 2022.

[11] F. -A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep., 2023.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[13] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10346–10357, Aug. 2023.

[14] Y. Sun, T. Xu, J. Li, Y. Chu, and X. Ji, "MMH-STA: A macro-micro-hierarchical spatio-temporal attention method for multi-agent trajectory prediction in unsignalized roundabouts," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11237–11250, Sep. 2023.

[15] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang, and Q. Ye, "Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for internet of vehicles," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2513912.

[16] X. Chen, H. Zhang, F. Deng, J. Liang, and J. Yang, "Stochastic non-autoregressive transformer-based multi-modal pedestrian trajectory prediction for intelligent vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 3561–3574, May 2024.

[17] Y. Zhou, H. Wang, N. Ning, Z. Wang, Y. Zhang, and F. Liu, "A bidirectional trajectory contrastive learning model for driving intention prediction," *Complex Intell. Syst.*, vol. 9, pp. 4301–4315, 2023.

[18] A. Poibrenski, F. Nozarian, F. Rezaeianaran, and C. Müller, "Uncertainty-aware pseudo labels for domain adaptation in pedestrian trajectory prediction," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst.*, 2023, pp. 5771–5777.

[19] H. Cui et al., "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 2090–2096.

[20] N. Deo and M. M. Trivedi, "Trajectory forecasts in unknown environments conditioned on grid-based plans," 2020, *arXiv:2001.00735*.

[21] L. Hou, S. E. Li, B. Yang, Z. Wang, and K. Nakano, "Integrated graphical representation of highway scenarios to improve trajectory prediction of surrounding vehicles," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1638–1651, Feb. 2023.

[22] Y. Hu, W. Zhan, and M. Tomizuka, "Scenario-transferable semantic graph reasoning for interaction-aware probabilistic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23212–23230, Dec. 2022.

[23] N. Rhinehart, K. M. Kitani, and P. Vernaza, "R2P2: A reparameterized pushforward policy for diverse, precise generative path forecasting," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 772–788.

[24] C. Tang, W. Zhan, and M. Tomizuka, "Exploring social posterior collapse in variational autoencoder for interaction modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 8481–8494.

[25] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "DESIRE: Distant future prediction in dynamic scenes with interacting agents," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 336–345.

[26] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1468–1476.

[27] Y. Zhou, Z. Wang, N. Ning, Z. Jin, N. Lu, and X. Shen, "I2T: From intention decoupling to vehicular trajectory prediction based on prioriformer networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9411–9426, Aug. 2024.

[28] L. Li, X. Wang, D. Yang, Y. Ju, Z. Zhang, and J. Lian, "Real-time heterogeneous road-agents trajectory prediction using hierarchical convolutional networks and multi-task learning," *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 4055–4069, Feb. 2024.

[29] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff, "CoverNet: Multimodal behavior prediction using trajectory sets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14074–14083.

[30] B. Kim et al., "LaPred: Lane-aware prediction of multi-modal future trajectories of dynamic agents," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14636–14645.

[31] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "HOME: Heatmap output for future motion estimation," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2021, pp. 500–507.

[32] T. Gu et al., "Stochastic trajectory prediction via motion indeterminacy diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17113–17122.

[33] I. Bae, Y. -J. Park, and H. -G. Jeon, "SingularTrajectory: Universal trajectory predictor using diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17890–17901.

[34] K. Lv, L. Yuan, and X. Ni, "Learning autoencoder diffusion models of pedestrian group relationships for multimodal trajectory prediction," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2511412.

[35] C. Jiang et al., "MotionDiffuser: Controllable multi-agent motion prediction using diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9644–9653.

[36] Y. Wang et al., "Optimizing diffusion models for joint trajectory prediction and controllable generation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 324–341.

[37] J. Wang, J. Guo, M. Feng, C. Li, X. Xue, and J. Pu, "Trajectory grid diffusion for multimodal trajectory prediction in autonomous vehicles," *IEEE Trans. Intell. Veh.*, early access, Nov., 11, 2024, doi: 10.1109/TIV.2024.3495037.

[38] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.

[39] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[40] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. Workshop Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–8.

[41] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.

[42] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 499–513, Feb. 2020.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and D. Weissenborn, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–12.

[44] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2022, pp. 16000–16009.

[45] J. Gao et al., "VectorNet: Encoding hd maps and agent dynamics from vectorized representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11522–11530.

[46] S. Luo and W. Hu, "Diffusion probabilistic models for 3D point cloud generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2836–2844.

[47] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, "FFJORD: Free-form continuous dynamics for scalable reversible generative models," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–13.

[48] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8577–8584.

[49] T. -Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[50] M. F. Chang et al., "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8748–8757.

[51] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.

[52] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24804–24816.

[53] S. Tonekaboni, D. Eytan, and A. Goldenberg, "Unsupervised representation learning for time series with temporal neighborhood coding," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–17.

[54] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 5775–5787.

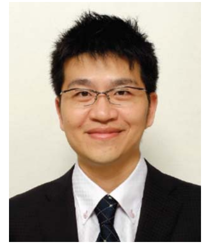
[55] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.
[56] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. 18th Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2015, pp. 234–241.

Yi Zhou (Member, IEEE) received the B.S. degree in electronic engineering from the First Aeronautic Institute of Air Force, Harbin, China, in 2002, and the Ph.D. degree in control system and theory from Tongji University, Shanghai, China, in 2011. He is currently a Full Professor and the Deputy Dean with the School of Artificial Intelligence, Henan University, Zhengzhou, China. He is also the Director of International Joint Research Laboratory for Cooperative Vehicular Networks, Zhengzhou. His research interests include vehicular cyber physical systems and multi-agent collaboration.

Zhangyun Wang (Student Member, IEEE) is currently working toward the M.S. degree with Henan University, Zhengzhou, China. His research interests include graph neural networks, probabilistic diffusion modeling, and self-supervised learning for multimodal motion forecasting of intelligent agents.

Nianwen Ning received the Ph.D. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2021. He is currently an Associate Professor with the School of Artificial Intelligence with Henan University, Zhengzhou, China. His research interests include graph data mining and intelligent transportation.

Ning Lu (Member, IEEE) received the B.Eng. and M.Eng. degrees in electrical engineering from Tongji University, Shanghai, China, in 2007 and 2010, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 2015. In 2009, he was as an Intern with the National Institute of Informatics, Tokyo, Japan. From 2015 to 2016, he was a Postdoctoral Fellow with Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Champaign, IL, USA. He was an Assistant Professor with the Department of Computing Science, Thompson Rivers University, Kamloops, BC, Canada. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering with Queen’s University, Kingston, ON. He has authored or coauthored more than 60 papers in top IEEE journals and conferences, including IEEE/ACM Transactions on Networking, IEEE Journal on Selected Areas in Communications, ACM MobiHoc, and IEEE INFOCOM. His research interests include scheduling, distributed algorithms, and reinforcement learning for wireless communication networks.



Mianxiong Dong (Senior Member, IEEE) was a JSPS Research Fellow with the School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan, and a Visiting Scholar with BBCR Lab, University of Waterloo, Waterloo, ON, Canada, supported by the JSPS Excellent Young Researcher Overseas Visit Program from 2010 to 2011. He is currently the Vice President and a Professor with the Muroran Institute of Technology, Muroran, Japan. Dr. Dong was selected as a Foreigner Research Fellow (a total of 3 recipients all over Japan) by NEC C&C Foundation in 2011. He was the recipient of the 12th IEEE ComSoc AsiaPacific Young Researcher Award 2017, Funai Research Award 2018, NISTEP Researcher 2018 (one of only 11 people in Japan), Young Scientists’ Award from MEXT in 2021, SUEMATSU-Yasuharu Award from IEICE in 2021, IEEE TCSC Middle Career Award in 2021. He was selected as a Highly Cited Researcher by Clarivate Analytics (WoS) in 2019, 2021, 2022, and 2023. And he is a Foreign Fellow of the Engineering Academy of Japan.

Xuemin Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular networks. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, and an International Fellow of the Engineering Academy of Japan. He was the recipient of the West Lake Friendship Award from Zhejiang Province in 2023, President’s Excellence in Research from the University of Waterloo in 2022, Canadian Award for Telecommunications Research from the Canadian Society of Information Theory in 2021, R.A. Fessenden Award in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society (ComSoc), and Technical Recognition Award from Wireless Communications Technical Committee in 2019 and AHSN Technical Committee in 2013. He was also the recipient of the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier’s Research Excellence Award in 2003 from the Province of Ontario, Canada. He is/was the General Chair for the 6G Global Conference’23 and ACM Mobihoc’15, Technical Program Committee Chair/Co-Chair for IEEE Globecom’24, 16 and 07, IEEE Infocom’14, IEEE VTC’10 Fall, and the Chair for the IEEE ComSoc Technical Committee on Wireless Communications. Dr. Shen is the Past President of the IEEE ComSoc, Vice President for Technical & Educational Activities, Vice President for Publications, Member-at-Large on the Board of Governors, Chair of the Distinguished Lecturer Selection Committee, and a Member of IEEE Fellow Selection Committee of the ComSoc. Dr. Shen was the Editor-in-Chief of the IEEE Internet of Things Journal, IEEE Network, and *IET Communications*.