PriorMotion: Generative Class-Agnostic Motion Prediction with Raster-Vector Motion Field Priors

Kangan Qian[1], Jinyu Miao[1], Xinyu Jiao[1,†], Ziang Luo[1], Zheng Fu[1], Yining Shi[1], Yunlong Wang[1], Kun Jiang[1,2,†], Diange Yang[1,2,†]

[1]School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China

[2]State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, 10084, China

{qka23, jiangkun, ydg}@mails.tsinghua.edu.cn

Abstract

Reliable spatial and motion perception is essential for safe autonomous navigation. Recently, class-agnostic motion prediction on bird's-eye view (BEV) cell grids derived from LiDAR point clouds has gained significant attention. However, existing frameworks typically perform cell classification and motion prediction on a per-pixel basis, neglecting important motion field priors such as rigidity constraints, temporal consistency, and future interactions between agents. These limitations lead to degraded performance, particularly in sparse and distant regions. To address these challenges, we introduce PriorMotion, an innovative generative framework designed for class-agnostic motion prediction that integrates essential motion priors by modeling them as distributions within a structured latent space. Specifically, our method captures structured motion priors using raster-vector representations and employs a variational autoencoder with distinct dynamic and static components to learn future motion distributions in the latent space. Experiments on the nuScenes dataset demonstrate that PriorMotion outperforms state-of-the-art methods across both traditional metrics and our newly proposed evaluation criteria. Notably, we achieve improvements of approximately 15.24% in accuracy for fast-moving objects, an 3.59% increase in generalization, a reduction of 0.0163 in motion stability, and a 31.52% reduction in prediction errors in distant regions. Further validation on FMCW LiDAR sensors confirms the robustness of our approach.

1. Introduction

Autonomous driving systems [1, 16, 18, 24, 33] demand highly precise spatial perception, including position information [11], semantic classification [36], and motion behavior [26], which are crucial for ensuring safety and comfort in open-world scenarios [1, 30, 44]. The traditional paradigm of object-centric methods models motion behavior estimation as a trajectory prediction task [5, 7, 9, 27, 55], which, despite its maturity, falters in open environments due to its reliance on detection-tracking-prediction pipelines that are ill-equipped to handle inexhaustible categories of traffic participants critical for system safety [51].

†Corresponding authors: Xinyu Jiao, Kun Jiang, Diange Yang

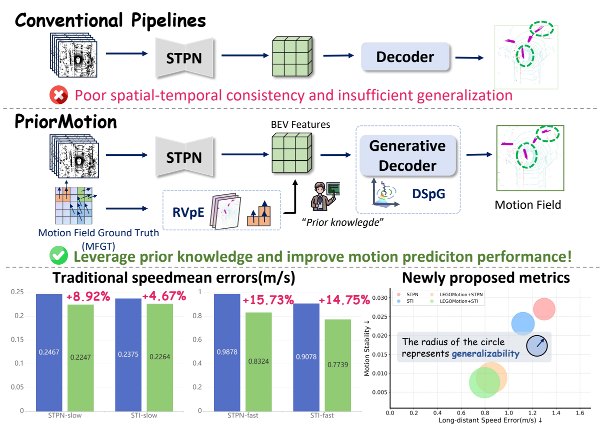

Figure 1. Comparison of PriorMotion with conventional pipelines. **Top row**: Previous SOTA framework, which adopts encoder-decoder pipeline. **Bottom**: Our approach leverages prior knowledge from the MFGT, modeling motion prediction as a future generation task in a structured latent space. Our methods outperforms the baseline across both (a) traditional mean speed error among both slow and fast moving objects and (b) newly proposed metrics. The radius of the circle represents generalization ability.

Recently, class-agnostic methods have emerged as a promising alternative by jointly performing semantic classification and motion prediction through BEV occupancy grids, removing the limitation of the requirement to detect the semantic object first [41]. However, these methods formulate motion prediction and cell classification tasks on the per-pixel basis, which suffer from poor spatial and temporal consistency as well as instance interaction. Traditional pipelines primarily focus on designing novel encoders and human-designed loss functions to extract spatial-temporal features from point clouds and ensure consistency [25, 45–49]. As depicted in Fig. 1, the discriminative approach with a simple regression decoder falls short in capturing structured physical laws, like rigid body characteristics and the continuous, linear evolution of motion, in BEV grid-based representations. This shortcoming is especially evident considering the sparse nature of point clouds [6, 46]. In conventional pipelines that directly regress outputs based on spatiotemporal features, while effective to some extent, the inherent sparsity of point clouds limits the optimization potential of feature extraction encoders, leaving fundamental challenges unaddressed.

We seek to address these challenges by drawing inspiration from a common phenomenon in our daily lives—teaching methodologies—where solutions can often be derived from the problem itself or inferred from standard answers. The Motion Field Ground Truth (MFGT), as a dense and structured representation, inherently contains implicit prior knowledge. Specifically, the motion within a single instance should exhibit consistent flow patterns. And there are global priors such as the motion field at crossroads differing from that on regular roads. Furthermore, high-order interactions, like the deceleration of a leading vehicle significantly influencing the behavior of following vehicles, exemplify the complex dependencies captured by MFGT. Indeed, prior-enhanced frameworks [19, 58] have already been successfully applied in online mapping. However, traditional motion prediction pipelines employing simple convolutional decoders fail to effectively leverage such priors. Motivated by these observations, two key questions naturally emerge: *(1) Can prior knowledge enhance class-agnostic motion prediction? (2) How can we integrate priors into traditional class-agnostic motion prediction frameworks?* To explore the first, we begin with a toy example. We evaluate previous SOTA methods MotionNet [51], on our private FMCW LiDAR benchmark, utilizing radial velocity information as additional input. Our results in Tab. 1 confirm that high-quality prior knowledge has a positive impact on the motion prediction task.

Method	**Static↓**	**Slow↓**	**Fast↓**
Base [35]	**0.0644**	0.5036	1.0654
Base + Radial_Velocity	0.0653	**0.4344**	**0.8897**

Table 1. Motion prediction mean error (m) comparison on FMCW.

To address the second question, we propose **PriorMotion**, a generative design that seamlessly converts conventional motion regression decoders into probabilistic motion generators. The core idea is to formalize prior knowledge as probability distributions, allowing the model to capture the full range of plausible motions [17, 56]. Unlike discriminative models, which focus on conditional predictions, PriorMotion explicitly learns the underlying distribution of motion patterns. By integrating generative principles with regression architectures, our framework overcomes the limitations of deterministic approaches. This enables both accurate predictions and systematic exploration of diverse, high-fidelity motions based on the learned prior distribution. As illustrated in Fig. 1, our framework comprises two key modules, which helps to extract prior knowledge and guided motion prediction. The Raster-Vector Prior Encoder (RVpE) extracts prior knowledge from the MFGT using both rasterized and vectorized representations. Dynamic-Static prior Generation (DSpG) module draws inspiration from human visual processing of dynamic-static scene decomposition. We model structural priors from MFGT using a dual-branch Variational Autoencoder (VAE), followed by a Spatial GRU for autoregressive latent space modeling to refine motion prediction. During inference, we sample from the learned distributations conditioned on the BEV features and can thus predict future motion evolution. Our contributions are summarized as follows:

- Propose a novel paradigm that systematically integrates structured motion priors to address spatiotemporal inconsistencies and model high-order interactions within BEV cell grids, significantly enhancing motion prediction task.
- Reformulate motion prediction as a generative process where priors are modeled as distributions, allowing seamless integration into conventional pipelines. The **RVpE** and **DSpG** modules collaboratively extract and leverage these priors to enhance motion prediction tasks.
- Comprehensive experiments on the nuScenes dataset (as shown in Fig. 1) demonstrate that **PriorMotion** surpasses state-of-the-art methods on both traditional benchmarks and newly proposed evaluation metrics.

2. Related Work

2.1. Motion Prediction

Motion prediction methods aim to predict agents' future trajectories from past observations by extracting spatial and semantic information [23]. These approaches can be categorized into object-centric and class-agnostic (grid-centric) methods. **Object-centric** methods rely on a cascaded pipeline of detection [22, 31, 32, 40, 53], tracking [20, 38, 42], and prediction [8, 50, 52]. Each module sequentially depends on object detectors, making these methods less effective in open-set scenarios with unknown objects. **Class-agnostic** methods directly predict cell classification and motion fields in BEV without relying on detection [25, 34, 45, 46]. MotionNet [51] laid the foundation for grid-based spatial perception, while PillarMotion [29] improved motion accuracy using cross-sensor self-supervision. BE-STI [47] introduced a spatio-temporal encoder for separate decoding of grid semantics and motion, and ST-Transformer [49] enhanced feature extraction with

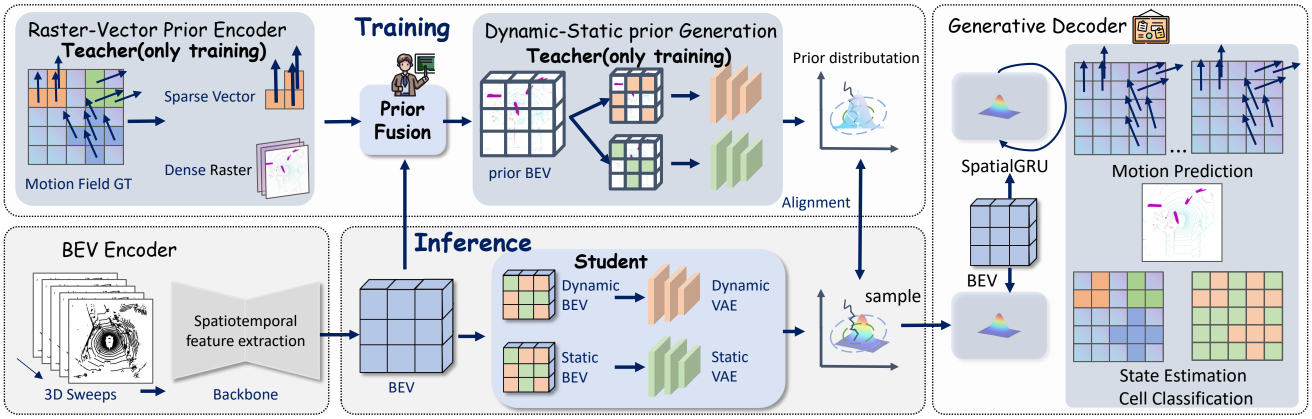

Figure 2. Architecture of **PriorMotion**. The architecture comprises several key components: a BEV Encoder, a Raster-Vector prior knowledge Encoder (RVpE) and a Dynamic-Static prior Generator (DSpG). RVpE is only used during training. The motion prior knowledge extracted by the RVpE module is used to guide the training of the latent space so that even without the RVpE module during inference, the structural priors of BEV motion can still be involved in the network.

spatio-temporal attention. Despite these advances, most methods focus on spatial-temporal feature extraction and struggle with the sparsity of point cloud data. Efforts to address this also include spatial and temporal consistency loss functions [25, 45, 46].

In contrast, our generative framework integrates structural prior knowledge from the MFGT, improving the decoder's ability to produce accurate and adaptable predictions across diverse motion patterns. To advance class-agnostic methods, we propose a comprehensive evaluation protocol with novel metrics that assess traditional criteria as well as motion stability and long-range prediction accuracy.

2.2. Generative Framework

Generative methods have been widely applied in various fields, including image processing [15, 37], text generation [2], and malware classification [3]. In autonomous driving, their utility in trajectory prediction is particularly noteworthy [43, 54, 56]. For example, Social GAN [13] integrates adversarial training for pedestrian motion forecasting, while TrajGen [54] generates diverse and realistic trajectories to enhance system adaptability in complex scenarios. DiffMap [17] is the prior work to use the generative method with diffusion models to improve map segmentation.

Our work presents a pioneering generative framework tailored for motion prediction, addressing a significant void in the utilization of generative models for class-agnostic motion prediction tasks.

3. The proposed Approach

3.1. Problem Formulation

Input data representation. The input to our model is a sequence of LiDAR point clouds, denoted as $\mathcal{P}_t = \{P_t^i\}_{i=1}^{N_t}$, where $P_t^i \in \mathbb{R}^3$ represents the coordinates of a point at time t, and N_t is the number of points in the cloud. These point clouds are synchronized to the ego vehicle's coordinate system. The point clouds are transformed into a voxelized representation $\mathcal{V}_t \in \{0,1\}^{H\times W\times C}$, where H, W, and C are the voxel grid dimensions along the X, Y, and Z axes. Non-empty voxels are set to 1, and empty ones to 0.

Output data representation. The model outputs cater to three tasks: motion prediction, cell classification, and state estimation. For motion prediction, we predict the future trajectory of each cell in the BEV pseudo-image as:$\{\mathcal{M}_t = (x_t, y_t) \mid \mathcal{M}_t \in \mathbb{R}^{H\times W\times 2}\}_{t=1}^{T}$ where $\mathcal{M}_t$ denotes the BEV motion field, and T is the number of predicted frames. For cell classification, we predict the class of each cell at the current time step, represented as $\mathcal{C}_t \in \mathbb{R}^{H\times W\times N_{\mathcal{C}}}$, where $\mathcal{C}_t$ contains the class labels of each cell, and $N_{\mathcal{C}}$ denotes the number of detectable cell categories. For state estimation, we predict the probability of each cell being static, represented as $\mathcal{S}_t \in \mathbb{R}^{H\times W}$, which denotes the static probability for each cell.

Problem formulation. Given a sequence of LiDAR point clouds $\{\mathcal{P}_t\}_{t=1}^{T}$, our goal is to predict future object positions, classify each cell, and estimate its state in the BEV map. We aim to learn a function f such that:

$$f(\{\mathcal{P}_t\}_{t=1}^{T}) \to (\mathcal{M}_t, \mathcal{C}_t, \mathcal{S}_t) \tag{1}$$

3.2. PriorMotion Network

We introduce **PriorMotion**, a prior-enhanced generative architecture designed to improve motion prediction in autonomous driving. As illustrated in Fig. 2, our architecture consists of three key components: a BEV Encoder (Sec. 3.2.1) that extracts spatial and temporal features from voxelized LiDAR data; RVpE (Sec. 3.2.2) that captures instance-level and grid-level priors in the motion field during training; and DSpG (Sec. 3.2.3) that learns latent structural priors by disentangling dynamic and static elements and generating future motion in this learned latent space.

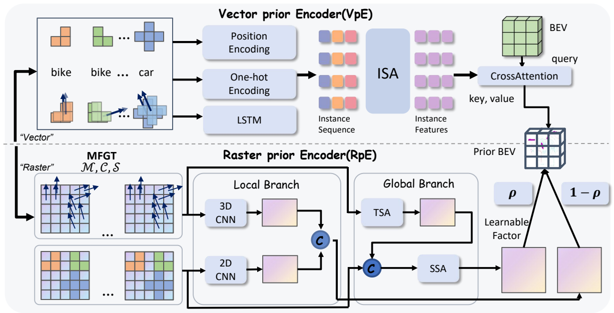

Figure 3. **RVpE** Module. The top row illustrates the Vector Prior Encoder (VpE), which captures sparse interactions between instances. The bottom part shows the Raster Prior Encoder (RpE), which encodes dense rasterized representations of the motion field to enhance class-agnostic grid predictions.

3.2.1. BEV Encoder

The BEV Encoder processes multiple frames of voxelized point clouds $\mathcal{V}_t$. The BEV feature encoder in our framework can be flexibly chosen from various state-of-the-art architectures, including the Spatio-Temporal Pyramid Network (STPN) [51], or the temporally and spatially bidirectional enhanced encoder (TeSE and SeTE) from BE-STI [47]. Using the backbone network, it outputs a BEV feature map $\mathcal{B} \in \mathbb{R}^{H \times W \times C'}$.

3.2.2. Raster-Vector prior Encoder

The prior knowledge extractor in PriorMotion is designed to distill valuable prior knowledge from the motion field, which is crucial for enhancing the network's performance. Given that tasks operate on grid data in the BEV pseudo-image, we naturally leverage dense rasterized representations to represent prior knowledge. Additionally, to account for the instance characteristics of objects, we also extract sparse vectorized representations to capture the interactions between instances. This dual approach ensures comprehensive integration of prior knowledge through two main components: the Raster Prior Encoder (RpE) and the Vector Prior Encoder (VpE), as shown in Fig. 3.

Raster prior Encoder. To effectively extract features from dense feature maps, convolutional operations are naturally employed [47, 51]. However, conventional encoders often struggle to capture essential global details, potentially overlooking critical scene-wide information. Such information is crucial for understanding complex traffic scenarios, such as those encountered at intersections. In this work, we treat the motion field as the primary feature map and introduce a raster prior encoder that incorporates both global and local processing branches. The CNNs enhance the extraction of local features in high-resolution images, while attention mechanisms are utilized on low-resolution inputs to efficiently gather global scene information, thereby reducing computational overhead.

Local Branch: For the ground-truth BEV motion field $\mathcal{M}$, we initially apply a 1×1 3D convolution operation to extract spatial and temporal features, subsequently integrating the temporal dimension into the channel dimension, denoted as $\mathcal{F}_{\mathcal{M}}^{local}$. Similarly, for the ground-truth cell classification C and state S, concatenation is first performed before applying a 1×1 2D convolution to extract their features, represented by $\mathcal{F}_{\mathcal{C},\mathcal{S}}^{local}$ in Eq. 2, then a concatenation operation is employed to form the final local features $\mathcal{F}_{\mathcal{M},C,S}^{local}$.

$$\mathcal{F}_{\mathcal{M}}^{local} = \texttt{3DConv}(\mathcal{M}), \quad \mathcal{F}_{\mathcal{C},\mathcal{S}}^{local} = \texttt{2DConv}([\mathcal{C}, \mathcal{S}]) \tag{2}$$

Here, $[\cdot]$ denotes the concatenation operation.

Global Branch: We start by reducing the resolution of the motion field $\mathcal{M}$ and then apply a Temporal Self-Attention (TSA) module to generate a new feature map $\mathcal{F}_{\mathcal{M}}^{global}$. After acquiring this temporally enriched global feature map, it is concatenated with the category map $\mathcal{C}$ and motion state map $\mathcal{S}$ and fused by a Spatial Self-Attention (SSA) module to form an integrated feature map $\mathcal{F}_{\mathcal{M},\mathcal{C},\mathcal{S}}^{global}$.

$$\mathcal{F}_{\mathcal{M}}^{global} = \text{TSA}(\mathcal{M}), \mathcal{F}_{\mathcal{M},\mathcal{C},\mathcal{S}}^{global} = \text{SSA}([\mathcal{F}_{\mathcal{M}}^{global}, \mathcal{C}, \mathcal{S}]) \tag{3}$$

Ultimately, we introduce a learnable factor ρ to dynamically merge global and local information into $\mathcal{P}_{\text{R}}$, facilitating the adaptive allocation of feature weights.

$$\mathcal{P}_{\text{R}} = \rho \times \mathcal{F}_{\mathcal{M},\mathcal{C},\mathcal{S}}^{global} + (1 - \rho) \times \mathcal{F}_{\mathcal{M},\mathcal{C},\mathcal{S}}^{local} \tag{4}$$

Vector prior Encoder. To capture instance-aware dynamics, we construct N_{ins} instance sequences through motion field sampling. For each instance, N grid points (x_i, y_i) are sampled along its trajectory, analogous to SD map polyline extraction, with 2D interpolation for underpopulated instances. Each grid is characterized by: 1) sinusoidal positional embeddings, 2) instance-specific displacement vector, and 3) one-hot category encoding.

The temporal evolution of the instance-aware displacement $\mathcal{M}^{\mathcal{I}} \in \mathbb{R}^{T \times 2}$ is captured through an LSTM layer.

The final instance representation $\mathcal{I} \in \mathbb{R}^{N_{\text{ins}} \times D_{\mathcal{I}}}$ integrates spatial interaction features:

$$D_{\mathcal{I}} = \underbrace{N \cdot d_{\text{pos}}}_{\text{geometry}} + \underbrace{N_{\mathcal{C}}}_{\text{semantics}} + \underbrace{T \cdot d_{\mathcal{M}}}_{\text{motion}} \tag{5}$$

where d_{pos} is the dimension of positional embedding, $d_{\mathcal{M}}$ is the dimension of LSTM output, and $D_{\mathcal{I}}$ is the dimension of the instance representation.

Transformer encoder for vector priors. Given the instance sequence representation of motion field, we wish to use a Transformer encoder to learn a feature representation. The Transformer Encoder consists of L layers of self-attention($L = 6$ in this work), allowing the network to model global relationships between instances with their corresponding motion across the entire scene. The output $\mathcal{P}_V$ is a vector prior knowledge representation of shape

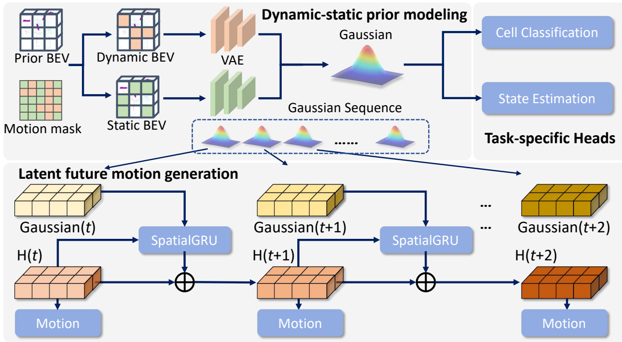

Figure 4. **DSpG** Module. The top part illustrates the VAE-based latent prior modeling, where prior BEV are projected into a Gaussian latent space. The bottom part shows the decoding process with spatialGRU, ensuring temporal consistency in the generated motion fields for motion field generation.

$N_{ins} \times D_{\mathcal{I}}$ produced by the instance-level self-attention (ISA) mechanism, which can be formulated as follows:

$$\mathcal{P}_V = \text{ISA}(\mathcal{I}) \tag{6}$$

Prior fusion. After extracting raster and vector prior knowledge, we apply Prior Cross-Attention (PCA) between the BEV features and vector priors. We then concatenate the enriched BEV features with raster priors, resulting in vector prior-enhanced BEV features. The final fused features $\mathcal{B}_{\text{prior}} \in \mathbb{R}^{H\times W\times C'}$ combine BEV features from voxelized point clouds and prior knowledge, offering a comprehensive representation for downstream tasks.

$$\mathcal{B}_{\text{prior}} = [\mathcal{B}, \mathcal{P}_{\text{R}}, \text{PCA}(q = \mathcal{B}, k = v = \mathcal{P}_{\text{V}})] \tag{7}$$

3.2.3. Dynamic-Static prior Generator

Considering that motion fields often exhibit structured patterns such as instance consistency and rigid dynamics, we formulate the prediction tasks as generating realistic motion and classification maps $\mathcal{M}$ and $\mathcal{C}, \mathcal{S}$ in BEV space, conditioned on features $\mathcal{B}$ to leverage these structured priors. Unlike conventional convolutional decoders, we model this process as $\mathcal{M} \sim p(\mathcal{M}|\mathcal{B})$ and $\mathcal{C}, \mathcal{S} \sim p(\mathcal{C}, \mathcal{S}|\mathcal{B})$.

Dynamic-static prior modeling. Drawing inspiration from how humans perceive the environment with varying focus on dynamic and static elements, we utilize an off-the-shelf perception model to initially generate dynamic and static masks. Our VAE framework projects BEV features into a structured latent space using dynamic/static masks from a pretrained perception model. The encoder outputs distribution parameters:

$$p(\mathcal{Z}|\mathcal{M}(\mathcal{B}_{\text{dynamic}}, \mathcal{B}_{\text{static}})) \sim \mathcal{N}(\mu_f, \sigma_f^2) \tag{8}$$

yielding compressed latent representation in BEV space $\mathcal{Z} \in \mathbb{R}^{H/4\times W/4\times D_{\mathbf{Z}}}$ through spatial downsampling.

Latent future motion generation. Once the latent prior distribution of the future motion field is obtained, we decode the latent representations into explicit motion fields in the BEV space. A direct approach would involve feeding the latent features into a decoder to generate the motion field, along with cell classification and state estimation. However, such an approach neglects the temporal evolution and interaction dynamics of grids. To address this concern, we decode latent states $\mathcal{Z}_0$ using a Spatial GRU (SGRU) that models temporal evolution and :

$$\mathcal{Z}_{t+1} = \text{SGRU}(\mathcal{Z}_t), \quad \mathcal{M}_{t+1} = \text{FSD}(\mathcal{Z}_{t+1}) \tag{9}$$

Subsequently, we decode the motion field for the next time step, $\mathcal{M}_{T+1}$, using a Feature Selection Decoder (FSD) with SE layer [14](explained in appendix). This process models the joint conditional probability:

$$\begin{aligned} p(\mathcal{M}(\mathcal{B}) \mid \mathcal{Z}_0) = p(\mathcal{M}_{t+1} \mid \mathcal{Z}_0) \cdot p(\mathcal{M}_{t+2} \mid \mathcal{M}_{t+1}, \mathcal{Z}_0) \cdots \\ \cdot\, p(\mathcal{M}_{t+T} \mid \mathcal{M}_{t+1}, \cdots, \mathcal{M}_{t+T-1}, \mathcal{Z}_0) \end{aligned} \tag{10}$$

t denotes the current timestamp and T is the prediction horizon, which is set to 20 in this work. For classification tasks, we decode through:

$$p(\mathcal{C}, \mathcal{S}|\mathcal{Z}_0) = \mathcal{B}' \oplus \mathcal{Z}_0 \tag{11}$$

where $\oplus$ denotes feature fusion. The complete architecture of DSpG module is shown in Fig. 4.

3.2.4. Loss Function

The PriorMotion model is trained using a composite loss function that optimizes motion prediction, cell classification, state estimation, and motion field priors. For motion prediction, classification, and state estimation, we follow standard practices from [51], with details in the appendix.

Prior loss. We introduce a Kullback-Leibler (KL) divergence loss to model motion field priors. This loss ensures the predicted BEV distribution $p(\mathcal{Z}|B)$ matches the ground truth $p(\mathcal{Z}|B')$, enforcing consistency with learned motion priors:

$$L_{\text{prior}} = \text{KL}\left(p(\mathcal{Z}|\mathcal{B}) \parallel p(\mathcal{Z}|\mathcal{B}_{\text{prior}})\right) \tag{12}$$

The total loss is a weighted sum of individual terms, balancing their contributions during training:

$$L = \lambda_{\text{mot}} \cdot L_{\text{mot}} + \lambda_{\text{state}} \cdot L_{\text{state}} + \lambda_{\text{cls}} \cdot L_{\text{cls}} + \lambda_{\text{prior}} \cdot L_{\text{prior}} \tag{13}$$

where λ_{mot}, λ_{state}, λ_{cls}, and λ_{prior} are hyperparameters controlling the importance of each loss term.

4. Experiment

In this section, we conduct experiments to address the following questions: (1) *Does our generative framework improve the performance of class-agnostic motion prediction algorithms across different backbones?* (2) *How do RVpE and DSpG influence learning performance and enhance the motion field with prior knowledge?* (3) *Does incorporating prior knowledge through our method unlock new capabilities such as motion stability, distant region perception with sparse point cloud inputs, and improved generalization?*

Method	Backbone	Static		Speed $\leq$ 5m/s		Speed $>$ 5m/s	
		Mean↓	Median↓	Mean↓	Median↓	Mean↓	Median↓
StaticModel	Rules	0	0	0.6111	0.0971	8.6517	8.1412
FlowNet3D[28]	PointNet	0.0410	0	0.8183	0.1782	8.5261	8.0230
HPLFlowNet[12]	BCL	0.0041	0.0002	0.4458	0.0960	4.3206	2.4881
PointRCNN[57]	PointNet	**0.0204**	0	0.5514	0.1627	3.9888	1.6252
LSTM-EM[34]	LSTM	0.0358	0	0.3551	0.1044	1.5885	1.0003
Pillar.M(L&I)[29]	Pillar.E	0.0245	0	0.2286	0.0930	0.7784	0.4685
SelfBEV(L&I)[10]	STPN	0.0514	0	0.4212	0.1073	2.0766	1.3226
WeakMotion[25]	STPN	0.0426	0	0.4009	0.1195	2.1342	1.2061
SelfMotion[45]	STPN	0.0419	0	0.3213	0.1061	2.2943	1.0508
MotionNet[51]	STPN	0.0262	0	0.2467	0.0961	0.9878	0.6994
MotionNet[51]†	STPN	0.0201	0	0.2292	0.0952	0.9454	0.6180
MotionNet[46]‡	STPN	0.0271	0	0.2267	0.0945	0.8427	0.5173
STPN /w(Ours)	STPN	0.0251(↓**4.20%**)	0	0.2247(↓**8.92%**)	0.0949	**0.8324**(↓**15.73%**)	0.6069
STI[47]	STI	0.0244	0	0.2375	0.0950	0.9078	0.6262
BE-STI[47]†	STI	0.0220	0	**0.2115**	**0.0929**	0.7511	0.5413
STI /w(Ours)	STI	0.0239(↓**2.05%**)	**0**	0.2264(↓**4.67%**)	0.0882	**0.7739**(↓**14.75%**)	**0.5772**

Table 2. Comparison with State-of-the-Art Results on nuScenes. We report the mean errors for static grids, slow moving grids with speed $\leq$ 5 m/s, and fast moving grids with speed $>$ 5 m/s. Pillar.M(I&L)[29] is the only method trained using both camera and LiDAR modalities. †: MGDA [35]. ‡: Data augmentation from [46]. STPN: MotionNet backbone. STI: BE-STI backbone.

4.1. Experimental Setup

Dataset. We conduct our experiments on the **nuScenes** dataset [4], a large-scale autonomous driving benchmark that provides comprehensive sensor data, including a 360-degree LiDAR, cameras, and radars. The dataset consists of 1000 scenes, with 850 scenes for training and validation, and 150 scenes for testing. Following the standard split, we use 500 scenes for training, 100 scenes for validation, and 250 scenes for testing. Each scene lasts approximately 20 seconds, with annotations provided at 2Hz and LiDAR point clouds captured at 20Hz.

Implementation details. For fair comparison, the same data preprocessing pipeline is adopted as in [51] with 3D bounding box annotations. Input point clouds are cropped to the range of $[-32m, 32m] \times [-32m, 32m] \times [-3m, 2m]$ and voxelized with a resolution of $0.25m \times 0.25m \times 0.4m$. Each sequence consists of 5 frames, with the last frame corresponding to the current time and the previous 4 frames from past timestamps. During training, we supervise the model with both motion prediction and cell classification tasks. We use the Adam optimizer [21] with an initial learning rate of 0.0016, which is decayed by a factor of 0.5 at epochs 10, 20, 30, and 40. The network is trained for 45 epochs with a batch size of 4 on a single Tesla A100 GPU.

Evaluation metrics. To evaluate our model, we follow the protocol in [51], dividing non-empty cells into three speed groups (static: speed ≤ 0.2 m/s, slow: speed ≤ 5 m/s, fast: speed > 5 m/s) and reporting mean and median prediction errors for each group. We also report overall accuracy (OA) and mean category accuracy (MCA) for cell classification. Additionally, we propose novel metrics for a more comprehensive evaluation:

Generalization metric. We assess the model's ability to generalize to unseen categories by masking the loss for a specific category during training. At test time, we compute the displacement error for the masked category. Let M be the set of masked cells, and $\hat{v}_i$ and v_i be the predicted and ground truth velocities for cell i. The error and generalization index (GI) are defined as:

$$L_{\mathrm{M}_c} = \frac{1}{|M|} \sum_{i \in M} \|\hat{v}_i - v_i\|_2, \quad GI = \frac{L_{\mathrm{M}_c}^{\text{Mask}}}{L_{\mathrm{M}_c}} \tag{14}$$

where $L_{\mathrm{M}_c}^{\text{Mask}}$ is the error for the masked category during masked training, and L_{M_c} is the error without masking. A higher GI indicates better generalization.

Motion stability metric. This metric measures the stability of motion for grids within the same instance. Let $\mathcal{I}$ be the set of instances, and D_i the set of displacement vectors for instance i. The stability for instance i is:

$$\sigma_i^2 = \frac{1}{|D_i|} \sum_{d \in D_i} \|d - \bar{d}_i\|_2^2 \tag{15}$$

where $\bar{d}_i$ is the mean displacement vector for instance i.

Distance-based displacement error metric. We evaluate displacement prediction performance across three distance ranges: $[0, 10]$, $[10, 20]$, and $[20, \infty)$ meters. For each range R, the mean displacement error is computed.

Method	Classification Accuracy(%)↑						
	Bg	Vehicle	Ped.	Bike	Others	MCA	OA
PointRCNN[39]	**98.4**	78.7	44.1	11.9	44.0	55.4	96.0
LSTM-ED[34]	93.8	91.0	73.4	17.9	71.7	69.6	92.8
MotionNet[51]	97.6	90.7	77.2	25.8	65.1	71.3	**96.3**
MotionNet†[51]	97.0	90.7	77.7	19.7	66.3	70.3	95.8
BE-STI[47]	97.3	91.1	78.6	24.5	66.5	71.6	96.0
BE-STI[47]†	94.6	**92.5**	82.9	25.9	77.3	74.7	93.8
STPN /w(Ours)	94.6	92.1	**86.9**	**27.3**	**80.4**	**76.3**	93.6

Table 3. Performance on the cell classification task on nuScenes.

4.2. Main Results

Comparison with SOTA methods. We show our performance on the nuScenes dataset in Tab. 2, comparing results of motion prediction models with and without PriorMotion. For fairness, no data augmentation or multi-gradient descent algorithm (MGDA) is used.

Our generative prior-enhanced framework significantly boosts performance compared to traditional models without such enhancements, especially in predicting both slow and fast-moving objects. It is compatible with multiple backbone architectures, including STPN and STI, consistently improving results across different setups. Specifically, when compared to MotionNet [51], PriorMotion reduces the mean prediction error by approximately 8.92% for slow-moving objects and 15.73% for fast-moving objects. Compared to STI, the reductions are about 4.67% and 14.75%, respectively. In cell classification tasks, our method also shows superior accuracy, particularly in identifying dynamic objects such as bicycles.

Generalization analysis. Object-level methods struggle under masking, producing random-like outcomes due to their reliance on detection, tracking, and prediction pipelines, which are highly sensitive to masked objects. In contrast, class-agnostic methods demonstrate stronger generalization, effectively handling incomplete or partially obscured data even when loss is not propagated through masked regions during training. Our generative architecture further enhances this generalization by capturing abstract prior knowledge. By leveraging structured priors and a rich latent space representation, the model effectively generalizes to unseen scenarios, including masked regions. This capability improves adaptability and reliability, ensuring accurate predictions in novel or partially occluded environments, making it well-suited for real-world applications.

Prediction performance at distant regions. We evaluate our framework across different distance ranges and observe a significant reduction in prediction errors at longer distances compared to other methods. Traditional approaches often suffer in distant regions due to the sparsity of point cloud data, as they mainly focus on encoder design. In contrast, our generative model leverages motion field priors and a structured latent space, effectively addressing point cloud sparsity and maintaining high prediction accuracy. These results highlight the robustness of **PriorMotion** in challenging scenarios involving distant objects.

Method	Backbone	Mean Speed↓		Generalization(%)↑
		≤ 5(m/s)	> 5(m/s)	
MotionNet	STPN	0.0704	0.2579	81.6
MotionNet(†)	STPN	0.0927	0.3159	
STPN /w(Ours)	STPN	0.0674	0.1969	**86.4**(↑5.88%)
STPN /w(Ours)(†)	STPN	0.0669	0.2278	
BE-STI	STI	0.0736	0.2077	84.3
BE-STI(†)	STI	0.0744	0.2463	
STI /w(Ours)	STI	0.0615	0.1672	**85.4**(↑1.30%)
STI /w(Ours)(†)	STI	0.0637	0.1956	

Table 4. Evaluation of the proposed generalization metric. The † symbol denotes that the "other" category cells are masked.

Method	Backbone	Static↓	Slow↓	Fast↓	Motion S. ↓
MotionNet	STPN	**0.0224**	0.2587	1.2990	0.0267
STPN /(Ours)	STPN	0.0263	**0.2207**	**0.8549**	**0.0088**
STI	STI	**0.0215**	0.2784	1.1200	0.0221
STI /(Ours)	STI	0.0254	**0.1922**	**0.7962**	**0.0075**

Table 5. Comparison of SOTA methods on long-distance(over 20 meters) speed error and motion stability metrics.

Runtime analysis. For real-time autonomous driving, LiDAR point cloud processing must stay within 100ms. Our model runs at 69ms(in Tab. 6), with 12ms for point cloud voxelization and 57ms for the forward pass.

Params	RVpE	DSpG	Static↓	Slow↓	Fast↓	Infer.T↓
(a)8.0M	✗	✗	**0.0240**	0.2467	1.0109	**19ms**
(b)9.2M	✗	✗	0.0244	**0.2375**	**0.9078**	45ms
(c)8.3M	✓	✗	0.0274	0.2273	0.9028	**24ms**
(d)11.5M	✓	✓	**0.0251**	**0.2247**	**0.8318**	69ms

Table 6. Results with different backbones: (a) MotionNet, (b) STI, Baseline variants: Our approaches: **(c) RVpE module /w simple decoder**, **(d) Ours**. Inference time measured on an RTX 3090.

Qualitative results. Results are shown in Fig. 5. Our framework accurately predicts motion across diverse object categories and improves motion stability and prediction accuracy in distant regions. This provides robust and reliable perception information for downstream tasks. More qualitative results are available in supplementary material.

4.3. Ablation Study

We conduct extensive ablation studies to evaluate the contribution of different components in our framework. The baseline model we used is MotionNet.

Effect of RVpE. To validate the effectiveness of the RVpE, we experiment with a simple generative framework(Simple.G), using concatenation to fuse prior knowledge. Incorporating the Rasterized and Vectorized prior

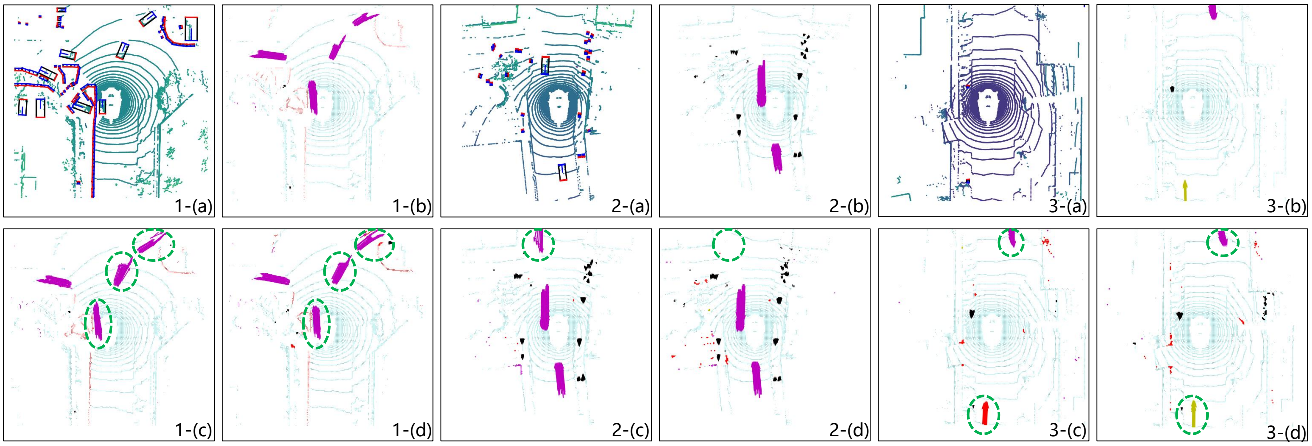

Figure 5. Qualitative comparison between PriorMotion and the baseline model. **Top row:** (a) object-level ground truth (BEV); (b) grid-level ground truth. **Bottom row:** (c) baseline predictions; (d) PriorMotion predictions. Motion directions are indicated by arrows; colors represent object categories: **Cyan**: background; **purple**: vehicle; **black**: pedestrian; **green**: bike; **red**: others.

Method	Module				Static		Speed≤5m/s		Speed > 5m/s	
	VpE	RpE	DSpM	LFmG	Mean↓	Median↓	Mean↓	Median↓	Mean↓	Median↓
Simple.G	✗	✗	✗	✗	0.0255	0	0.2477	0.0974	0.9733	0.7052
(a)	✓	✗	✗	✗	0.0318	0	0.2464	0.0960	0.9580	0.7003
(b)	✗	✓	✗	✗	0.0267	0	0.2356	0.0955	0.9351	0.6380
(c)	✓	✓	✗	✗	0.0274	0	**0.2273**	0.0953	0.9028	0.6216
(d)	✗	✗	✓	✗	**0.0235**	0	0.2360	0.0985	0.9564	0.6548
(e)	✗	✗	✓	✓	0.0249	0	0.2256	**0.0930**	0.8943	0.6892
(f)	✓	✓	✓	✓	0.0251	**0**	**0.2247**	0.0949	**0.8324**	**0.6069**

Table 7. Performance comparison of our models with different combinations of components on nuScenes.

representations individually and jointly—improves performance(shown in Tab. 7). Specifically, adding VpE alone reduces the mean prediction error for fast-moving objects by 1.57%, as it captures instance-level prior knowledge from vectorized representations, enhancing dynamic motion modeling. However, static cells show a slight performance drop due to the sparse nature of vector priors.

RpE further reduces the mean prediction error by 4.88% for slow-moving and 3.92% for fast-moving objects. By extracting prior knowledge from rasterized representations, RpE offers global-local, spatially consistent cues that complement VpE. The combination of both modules yields the most significant overall improvement, especially in motion stability, highlighting the benefit of integrating raster and vector priors for comprehensive scene understanding.

Effect of DSpG. We also investigate the role of the DSpG, with dynamic-static prior modeling (DSpM) and latent future motion generation (LFmG). As shown in Tab. 7, when using LFmG alone, we directly use a SGRU to gradually produce motion prediction given the simple prior injection. While SGRU effectively considers the temporal dynamics of motion prediction, it may introduce some complexity, leading to potential overfitting in static regions. With only the DSpM module, we directly decode the entire motion from the latent space, the dual modeling with dynamic and static avoid performance reduction of static cell grids. We see that both modules are effective and improve the performance. Combining the two modules further improves the performance by a large margin.

Performance on our private dataset. We further evaluate PriorMotion's performance on a private dataset collected using FMCW LiDAR. Our method shows consistent improvements in motion stability and speed accuracy in distant regions. Detailed results are provided in the appendix.

5. Conclusion

We presented PriorMotion, a generative framework enhancing class-agnostic motion prediction through the integration of motion priors modeled as distributions within a structured latent space. Our method includes a RVpE to extract motion field priors and a DSpG to model dynamic-static Gaussian distributions, enabling high-order interactions and producing more accurate future motion fields using learned structural priors. Experiments on the nuScenes dataset demonstrated that PriorMotion outperforms SOTA methods, achieving notable improvements in accuracy for fast-moving objects, enhanced generalization, increased motion stability, and reduced errors in distant regions.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (52372414, 52394264, 52472449).

References

[1] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018. 1

[2] Jonas Becker, Jan Philip Wahle, Bela Gipp, and Terry Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges. *arXiv preprint arXiv:2405.15604*, 2024. 3

[3] Rishiraj Biswas, Thirumurugan Shanmugam, Rajiv Vincent, Arun Kumar Sivaraman, Janakiraman Nithiyanantham, and Priya Ravindran. Gan-enhanced multiclass malware classification with deep convolutional networks. In *Applications and Techniques in Information Security*, pages 244–255, Singapore, 2025. Springer Nature Singapore. 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6

[5] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 1

[6] Yubo Cui, Jiayao Shan, Zuoxu Gu, Zhiheng Li, and Zheng Fang. Exploiting more information in sparse point cloud for 3d single object tracking. *IEEE Robotics and Automation Letters*, 7(4):11926–11933, 2022. 2

[7] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104, 2020. 1

[8] Scott Ettinger, Kratarth Goel, Avikalp Srivastava, and Rami Al-Rfou. Scaling motion forecasting models with ensemble distillation. *arXiv preprint arXiv:2404.03843*, 2024. 2

[9] Liangji Fang, Qinhong Jiang, Jianping Shi, and Bolei Zhou. Tpnet: Trajectory proposal network for motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6797–6806, 2020. 1

[10] Shaoheng Fang, Zuhong Liu, Mingyu Wang, Chenxin Xu, Yiqi Zhong, and Siheng Chen. Self-supervised bird's eye view motion prediction with cross-modality signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1726–1734, 2024. 6

[11] Tai Fei, SC Mukhopadhyay, João Paulo Javidi Da Costa, Chirasree RoyChaudhuri, Lan Lan, and Nevine Demitri. Spatial environment perception and sensing in automated systems: A review. *IEEE Sensors Journal*, 2024. 1

[12] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3254–3263, 2019. 6

[13] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018. 3

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[15] Xirui Hu, Jiahao Wang, Hao Chen, Weizhan Zhang, Benqi Wang, Yikun Li, and Haishun Nan. Dynamicid: Zero-shot multi-id image personalization with flexible facial editability. *arXiv preprint arxiv:2503.06505*, 2025. 3

[16] Zilin Huang, Zihao Sheng, Yansong Qu, Junwei You, and Sikai Chen. Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving. *arXiv preprint arXiv:2412.15544*, 2024. 1

[17] Peijin Jia, Tuopu Wen, Ziang Luo, Mengmeng Yang, Kun Jiang, Ziyuan Liu, Xuewei Tang, Zhiquan Lei, Le Cui, Bo Zhang, et al. Diffmap: Enhancing map segmentation with map prior using diffusion model. *IEEE Robotics and Automation Letters*, 2024. 2, 3

[18] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. A survey on vision-language-action models for autonomous driving. *arXiv preprint arXiv:2506.24044*, 2025. 1

[19] Zhou Jiang, Zhenxin Zhu, Pengfei Li, Huan-ang Gao, Tianyuan Yuan, Yongliang Shi, Hang Zhao, and Hao Zhao. P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. *IEEE Robotics and Automation Letters*, 2024. 2

[20] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):140–153, 2018. 2

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[22] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[23] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH journal*, 1:1–14, 2014. 2

[24] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011. 1

[25] Ruibo Li, Hanyu Shi, Ziang Fu, Zhe Wang, and Guosheng Lin. Weakly supervised class-agnostic motion prediction for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17599–17608, 2023. 2, 3, 6

[26] Jinhao Liang, Chaopeng Tan, Longhao Yan, Jingyuan Zhou, Guodong Yin, and Kaidi Yang. Interaction-aware trajectory prediction for safe motion planning in autonomous driving: A transformer-transfer learning approach. *arXiv preprint arXiv:2411.01475*, 2024. 1

[27] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020. 1

[28] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019. 6

[29] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Self-supervised pillar motion learning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3183–3192, 2021. 2, 6

[30] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021. 1

[31] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[32] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[33] Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, et al. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *arXiv preprint arXiv:2505.15298*, 2025. 1

[34] Marcel Schreiber, Stefan Hoermann, and Klaus Dietmayer. Long-term occupancy grid prediction using recurrent neural networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9299–9305. IEEE, 2019. 2, 6, 7

[35] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 6

[36] Amirreza Shaban, Xiangyun Meng, JoonHo Lee, Byron Boots, and Dieter Fox. Semantic terrain classification for off-road autonomous driving. In *Conference on Robot Learning*, pages 619–629. PMLR, 2022. 1

[37] Pourya Shamsolmoali, Masoumeh Zareapoor, Eric Granger, Huiyu Zhou, Ruili Wang, M Emre Celebi, and Jie Yang. Image synthesis with adversarial networks: A comprehensive survey and case studies. *Information Fusion*, 72:126–146, 2021. 3

[38] Sarthak Sharma, Junaid Ahmed Ansari, J Krishna Murthy, and K Madhava Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3508–3515. IEEE, 2018. 2

[39] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 7

[40] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 2

[41] Yining Shi, Kun Jiang, Jiusi Li, Zelin Qian, Junze Wen, Mengmeng Yang, Ke Wang, and Diange Yang. Grid-centric traffic scenario perception for autonomous driving: A comprehensive review. *arXiv preprint arXiv:2303.01212*, 2023. 1

[42] Jeany Son, Mooyeol Baek, Minsu Cho, and Bohyung Han. Multi-object tracking with quadruplet convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5620–5629, 2017. 2

[43] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 34(4):1972–1987, 2021. 3

[44] Dequan Wang, Coline Devin, Qi-Zhi Cai, Philipp Krähenbühl, and Trevor Darrell. Monocular plan view networks for autonomous driving. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2876–2883. IEEE, 2019. 1

[45] Kewei Wang, Yizheng Wu, Jun Cen, Zhiyu Pan, Xingyi Li, Zhe Wang, Zhiguo Cao, and Guosheng Lin. Self-supervised class-agnostic motion prediction with spatial and temporal consistency regularizations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14638–14647, 2024. 2, 3, 6

[46] Kewei Wang, Yizheng Wu, Zhiyu Pan, Xingyi Li, Ke Xian, Zhe Wang, Zhiguo Cao, and Guosheng Lin. Semi-supervised class-agnostic motion prediction with pseudo label regeneration and bevmix. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5490–5498, 2024. 2, 3, 6

[47] Yunlong Wang, Hongyu Pan, Jun Zhu, Yu-Huan Wu, Xin Zhan, Kun Jiang, and Diange Yang. Be-sti: Spatial-temporal integrated network for class-agnostic motion prediction with bidirectional enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17093–17102, 2022. 2, 4, 6, 7

[48] Zhensong Wei, Xuewei Qi, Zhengwei Bai, Guoyuan Wu, Saswat Nayak, Peng Hao, Matthew Barth, Yongkang Liu, and Kentaro Oguchi. Spatiotemporal transformer attention

network for 3d voxel level joint segmentation and motion prediction in point cloud. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1381–1386. IEEE, 2022.

[49] Zhensong Wei, Xuewei Qi, Zhengwei Bai, Guoyuan Wu, Saswat Nayak, Peng Hao, Matthew Barth, Yongkang Liu, and Kentaro Oguchi. Spatiotemporal transformer attention network for 3d voxel level joint segmentation and motion prediction in point cloud. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1381–1386. IEEE, 2022. 2

[50] Sungmin Woo, Minjung Kim, Donghyeong Kim, Sungjun Jang, and Sangyoun Lee. Fimp: Future interaction modeling for multi-agent motion prediction. *arXiv preprint arXiv:2401.16189*, 2024. 2

[51] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11385–11395, 2020. 1, 2, 4, 5, 6, 7

[52] Guipeng Xin, Duanfeng Chu, Liping Lu, Zejian Deng, Yuang Lu, and Xigang Wu. Multi-agent trajectory prediction with difficulty-guided feature enhancement network. *arXiv preprint arXiv:2407.18551*, 2024. 2

[53] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 2

[54] Qichao Zhang, Yinfeng Gao, Yikang Zhang, Youtian Guo, Dawei Ding, Yunpeng Wang, Peng Sun, and Dongbin Zhao. Trajgen: Generating realistic and diverse trajectories with reactive and feasible agent behaviors for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23 (12):24474–24487, 2022. 3

[55] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904. PMLR, 2021. 1

[56] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. *arXiv preprint arXiv:2402.11502*, 2024. 2, 3

[57] Qiang Zhou and Chaohui Yu. Point rcnn: An angle-free framework for rotated object detection. *Remote Sensing*, 14 (11):2605, 2022. 6

[58] Xiyue Zhu, Vlas Zyrianov, Zhijian Liu, and Shenlong Wang. Mapprior: Bird's-eye view perception with generative models, 2023. 2