G-VTM: A Multimodal Vision-Trajectory Model for Generalized Vehicle Trajectory Prediction

Xinyue Zhang[1], **Letian Gong**[1], **Yan Lin**[2], **Jinjun Cheng**[1], **Junlin Zhang**[1], **Guanyu Yao**[1], **Shengnan Guo**[1,3*], **Youfang Lin**[1,4], **Shaojiang Wang**[5*] and **Huaiyu Wan**[1,4]
[1]School of Computer Science and Technology, Beijing Jiaotong University, China
[2]Department of Computer Science, Aalborg University, Denmark
[3]Key Laboratory of Big Data & Artificial Intelligence in Transportation, Ministry of Education, China
[4]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, China
[5]Institute of AI for Industries, Chinese Academy of Sciences, Nanjing 211135, China
{zhangxinyue, gonglt, chengjinjun, junlinzhang, guanyuyao, guoshn, yflin, hywan}@bjtu.edu.cn,
lyan@cs.aau.dk, wangshaojiang@iaii.ac.cn

Abstract

Generalized vehicle trajectory prediction across diverse junctions, including urban intersections and roundabouts, remains a fundamental task in Cooperative Vehicle–Infrastructure Systems (CVIS). This study faces two key challenges: (1) Generalization across junctions with heterogeneous map semantics and traffic behavioral patterns, where the former arises from differences in road topologies and traffic regulations, and the latter reflects diverse behavioral intentions of road users; (2) Scenario-adaptive interaction modeling, where single-modality trajectory learning captures local spatio-temporal correlation, but lacks map constraint and direction-aware interaction contexts. To overcome these challenges, we propose G-VTM, a generalized vision-trajectory model. G-VTM models fine-grained behavioral patterns and relative spatial interaction from trajectory modality. At the vision modality, G-VTM captures global map semantics while modeling scenario- and direction-aware interaction based on intuitive visual perception. Experiments on multiple real-world datasets collected by unmanned aerial vehicles (UAVs) demonstrate that our method achieves strong generalized performance under heterogeneous traffic conditions. The code is provided at https://github.com/zxyhaclyon/G-VTM.

1 Introduction

Vehicle trajectory prediction at junctions remains a critical task in Cooperative Vehicle–Infrastructure Systems (CVIS). This study focuses on generalized vehicle trajectory prediction across diverse junctions, which has significant application in intelligent transportation systems (ITS) [Zhao *et al.*, 2020; Cai *et al.*, 2021; Schimpf *et al.*, 2023; Huang *et al.*, 2025]. For example, a model trained on a signalized four-way intersection can be efficiently transferred to an unseen roundabout with a few training data, thereby reducing data collection and model deployment costs. We analyse trajectories of road users derived from unmanned aerial vehicles (UAVs) recorded videos. These UAVs are employed as roadside perception devices. Existing methods are typically tailored to a specific junction [Liao *et al.*, 2024b; Wei *et al.*, 2024], failing to achieve generalization, because they cannot effectively address two key challenges.

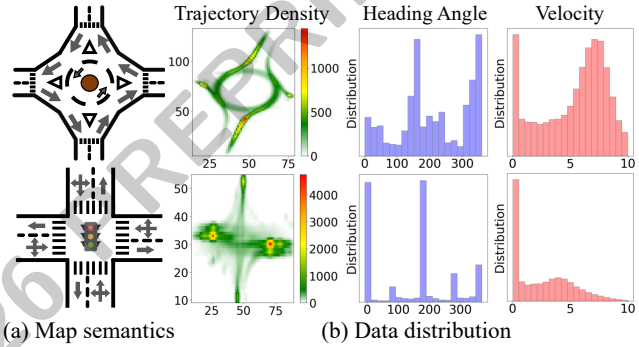

Figure 1: (a) illustrates distinct map semantics in the signaled four-way intersection and the roundabout. (b) presents heterogeneous data distributions across these junctions.

Generalizing across junctions with different map semantics and traffic behaviour patterns is the first challenge. The different map semantics primarily arise from distinct road topologies (e.g., global size, road geometry, and lane connectivity) and traffic regulations (signalized or unsignalized). The traffic behavioral patterns of road users are jointly influenced by road network constraints, traffic density, and traffic regulations, leading to heterogeneous data distributions across different junctions. Fig. 1 presents two different traffic scenarios and corresponding data distributions. At the roundabout, road users follow priority-based rules, and the traffic density exhibits a circular shape. The

*Corresponding author.

heading angles exhibit a dispersed distribution due to continuous steering adjustments at roughly constant velocity. In contrast, behavior patterns of road users are strictly regulated at the signalized four-way intersection, leading to a *deceleration–stop–restart* stage-wise movement. The distribution of heading angles is rather concentrated, as road users always execute left or right-angle turns.

Another challenge lies in modeling scenario-adaptive interactions among road users to facilitate generalization. Existing methods primarily employ Graph Neural Networks (GNNs) to learn interactive motion among road users based on fine-grained relative spatial features such as Euclidean distance and polar angle [Mo *et al.*, 2021; Westny *et al.*, 2023]. However, this single-modality trajectory learning lacks holistic perceptions of the surrounding scenario, which explicitly constrain interaction. Besides, it is limited in intuitively capturing direction-aware relative movement relations between the target and interacting road users, such as co-directional parallel or opposing bidirectional movement.

We observed that vision modalities could provide complementary advantages to the trajectory modality. Specifically, although coarse-grained visual images lack precise positional and movement details, they offer a holistic view that facilitates learning global map semantics with local relative spatial relations. Moreover, they provide intuitive perceptual fields to capture overall behavioral patterns of each trajectory and explicitly reveal relative movement relations between the target and interacting road users.

To address the above challenges, we propose a Generalized Vision-Trajectory Model (**G-VTM**) that effectively extracts and fuses multimodal information. G-VTM introduces two key components: (1) A *Map-aware Vision Learner* learns global map semantics with local relative spatial relations, and models scenario- and direction-aware interaction from coarse-grained visual features based on a multiscale strategy; (2) A *Spatio-temporal Trajectory Learner* captures fine-grained behavioral patterns of road users guided by visual priors. Finally, a Transformer encoder effectively fuses multimodal semantics, producing future trajectory predictions through a trajectory predictor. Our contributions are summarized as follows:

- We introduce G-VTM, a model that unifies vision and trajectory modality for generalized prediction of vehicle trajectories across diverse junctions.
- A Map-aware Vision Learner is designed to learn global map semantics for map images of arbitrary resolution, and model scenario- and direction-aware visual interaction between the target and interacting road users. Besides, A Spatio-temporal Trajectory Learner captures fine-grained behavioral patterns of road users guided by visual priors.
- We conduct extensive experiments on three real-world datasets. G-VTM achieves strong generalized performance in few-shot and zero-shot settings.

2 Related Work

Vehicle trajectory prediction aims to forecast the future motion of road users. It emphasizes modeling interactions among road users and capturing the surrounding scenario semantics. Traffic scenarios that existing methods focused on can be grouped into two types: (i) urban intersections; (ii) roundabouts and highway on-ramps and off-ramps.

Trajectory Prediction at Urban Intersections Urban intersections are among the most challenging traffic scenarios due to complex road topology, explicit traffic regulations, and multiple classes of road users. Early works primarily focus on trajectory prediction of a single road user based on RNN-based or physics-based models [Kawasaki and Tasaki, 2018; Abdelraouf *et al.*, 2022], but fail to capture the interactions among multiple road users. To address this limitation. GRIP++ [Li *et al.*, 2019] constructs a spatio-temporal graph network to model dynamic interactions between road users. FJMP [Rowe *et al.*, 2023] combines attention mechanisms with sparse interaction graphs based on influencer-reactor relationships to improve future interaction modeling. MTR [Shi *et al.*, 2022a] introduces learnable motion query pairs to model both global overall intention and local fine-grained trajectory dynamics, enabling efficient prediction of multimodal trajectories. To handle problems of non-uniform sampling and long-horizon uncertainty, continuous-time modeling has also been explored. LG-ODE [Huang *et al.*, 2020] models multi-agent trajectories in continuous time using Neural ODEs with a latent interaction graph. MTP-GO [Westny *et al.*, 2023] leverages Neural ODEs together with an Extended Kalman Filter (EKF) to facilitate multi-modal probabilistic trajectory prediction. Moreover, HPNet [Tang *et al.*, 2024] models the dynamic dependencies across continuous-time prediction steps via an attention mechanism, using historical predictions to guide the current prediction. Recognizing the impact of traffic signals, KI-GAN [Wei *et al.*, 2024] captures vehicle motion with signal states to generate accurate predictions under traffic regulations.

Trajectory Prediction at Highway On-Ramps and Off-Ramps and Roundabouts In highway on-ramps and off-ramps scenarios, existing studies mainly focus on forecasting vehicle behaviors such as lane changes and acceleration/deceleration. CS-LSTM [Deo and Trivedi, 2018] and PiP [Song *et al.*, 2020] introduce convolutional social pooling to capture fine-grained interactions. WSiP [Wang *et al.*, 2023] advances this concept by modeling interactions as wave superposition through a wave-pooling mechanism, which. C2F-TP [Wang *et al.*, 2025] adopts a diffusion-based framework, achieving strength in long-horizon prediction. BAT [Liao *et al.*, 2024b] models driving behavior using dynamic geometric graphs. HLTP [Liao *et al.*, 2024a] incorporates vision-aware pooling with a knowledge distillation framework to refine interaction representation while reducing inference latency. Roundabouts often have more complex road topologies and priority-based traffic regulations, which result in more complicated interactions between road users. MMH-STA [Sun *et al.*, 2023] introduces a macro–micro hierarchical spatio-temporal attention structure to model multi-agent trajectory prediction. FSD-Transformer [Yang *et al.*, 2025] leverages traffic states and driving styles to adapt predictions under varying traffic conditions.

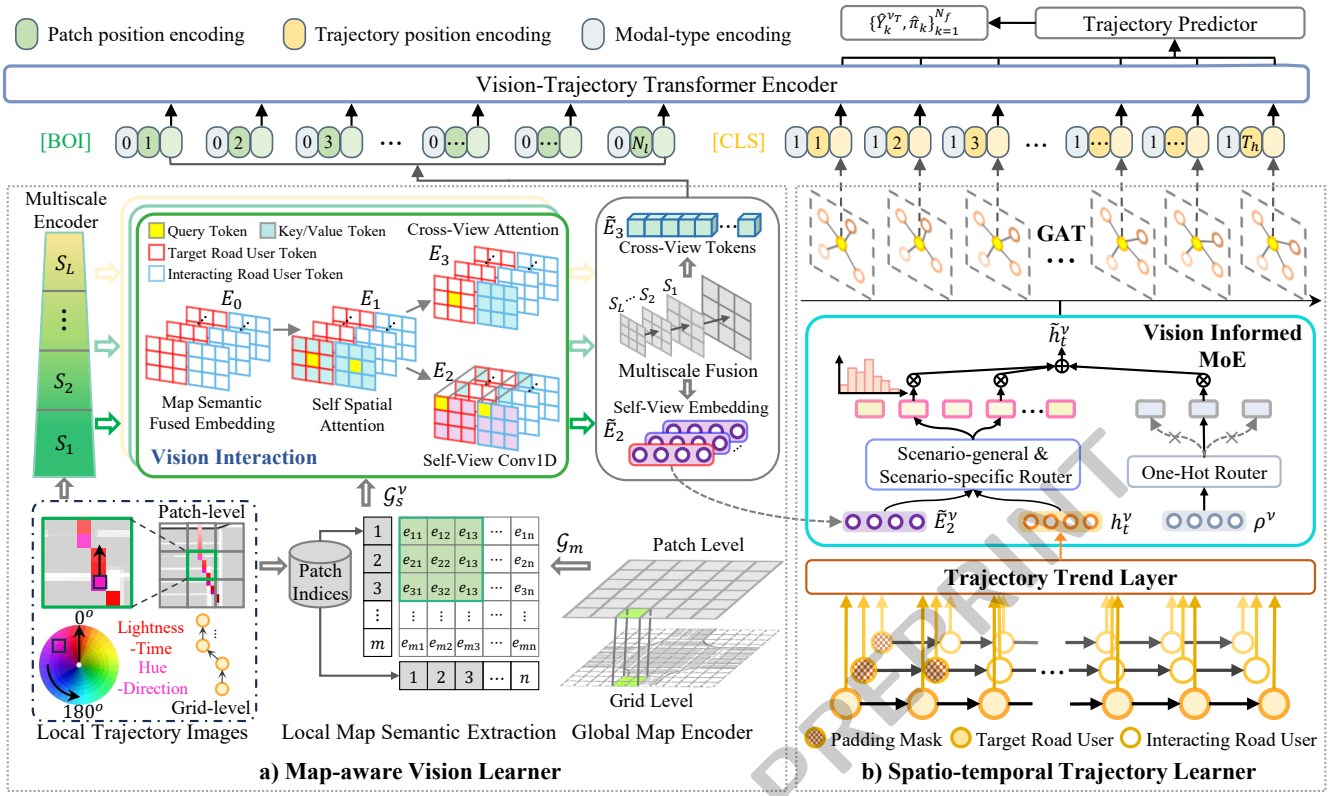

Figure 2: Overall framework of G-VTM.

3 Preliminaries

3.1 Definition

Definition 1 (Trajectory Features)**.** *Features of a road user ν at time t include the current position $\tau_t^\nu = (x_t^\nu, y_t^\nu)$ and motion information (e.g., velocity, acceleration, and heading angle). Road users include both the target road user ν_T and interacting road users $\nu_i \in \mathcal{V}_I$, such as bicycles, motorcycles, and pedestrians. Each road user has its class attribute represented as a one-hot vector $\boldsymbol{\rho}_v$ of length N_c. We select six road users geographically closest to the target road user, forming the set $\mathcal{V}_I$ of interacting road users.*

Definition 2 (Global Map Image)**.** *For a given junction, we visualize its road topology and traffic elements using the provided OpenStreetMap format map. First, the global map is split into 1 m × 1 m grids. Each grid corresponds to a fixed number of pixels and is further grouped into patches. Each patch includes a fixed number of grids. Thereby, ensuring the resolution of the map image is positively correlated with the physical dimensions of the junction.*

Definition 3 (Local Trajectory Image)**.** *Since the trajectory of road users is concentrated within a local region of the global map, we extract a fixed-resolution local image for each road user to facilitate generalization. We visualize dynamic behaviour patterns of road users on the static map image based on the grid level. The relative position offset between τ_{t-1} and τ_t is used to compute the direction angle, which is mapped onto the hue wheel. With saturation fixed, the lightness is determined by the ratio of time t in the historical horizon t_f, representing the temporal evolution of the trajectory.*

3.2 Problem Statement

Trajectory Prediction. Given the observed trajectory $\mathcal{H}_\nu = \langle \tau_{t_0-t_h+1}^\nu, \tau_{t_0-t_h+2}^\nu, \cdots, \tau_{t_0}^\nu \rangle, \nu \in \{\nu_T\} \cup \mathcal{V}_I$ of road users in the past t_h timesteps, the modeling objective of trajectory prediction is to estimate the future trajectory $\langle \tau_{t_0+1}^{\nu_T}, \tau_{t_0+2}^{\nu_T}, \cdots, \tau_{t_0+t_f}^{\nu_T} \rangle$ of the target road user, where t_f is the prediction horizon and t_0 is the current time.

Generalization. The objective of generalization is to learn a robust model that performs well across diverse scenarios. Once pretrained on the source dataset, this model could effectively transfer to the unseen target dataset, predicting accurate future trajectories based on few-shot inputs.

3.3 Rotary Position Embedding

Positional encoding is essential for capturing order- and spatial-aware relationships. Rotary Position Embedding (RoPE) represents relative positional information by rotating feature vectors in the embedding space.

Taking 1D RoPE as an example. Given embedding $\boldsymbol{e}_m \in \mathbb{R}^d$ of position m, we calculate query vector $\boldsymbol{q}_m = \boldsymbol{W}_q \boldsymbol{e}_m$ and key vector $\boldsymbol{k}_m = \boldsymbol{W}_k \boldsymbol{e}_m$. The rotary transformation is

calculated by rotation Matrixs $\boldsymbol{R}(m)$ which is defined as:

$$\boldsymbol{R}(m) = \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{bmatrix}, \tag{1}$$

where $\theta_k = 10000^{-2k/d}$ are frequency weights. The inner product in the attention between $\boldsymbol{q}_m$ at position m and $\boldsymbol{k}_n$ at position n enbale relative position-aware interactions:

$$\begin{aligned} \Phi(m,n) &= (\boldsymbol{R}(m)\boldsymbol{q}_m)^T \cdot \boldsymbol{R}(n)\boldsymbol{k}_n, \\ &= \boldsymbol{e}_m^T \boldsymbol{W}_q^T \boldsymbol{R}(m)^T \boldsymbol{R}(n) \boldsymbol{W}_k \boldsymbol{e}_n, \\ &= \boldsymbol{e}_m^T \boldsymbol{W}_q^T \boldsymbol{R}_{(m-n)} \boldsymbol{W}_k \boldsymbol{e}_n. \end{aligned} \tag{2}$$

4 Methodology

We proposed G-VTM to integrate the complementary strengths of vision and trajectory modalities. As illustrated in Figure 2, G-VTM comprises two key components: a Map-aware Vision learner and a Spatio-temporal trajectory learner.

4.1 Map-aware Vision Learner

General Global Map Encoder. When processing images with variable resolutions, Vision Transformer (ViT) generally employs linear interpolation for absolute positional embeddings, which is inadequate for capturing relative spatial geometric relationships between map patches. Therefore, we incorporate adaptive-resolution rotary positional encoding into the ViT. Specifically, ViT first projects the global map image into patch embeddings $\boldsymbol{\mathcal{G}}_m = \langle \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_{N_g} \rangle$. We assign two-dimensional positional indices to each patch, facilitating the modeling of spatial topological relationships. To support images of arbitrary resolutions, we transform positional indices to a new coordinate based on a scaling factor, which is calculated by the ratio between the current image resolution and the reference resolution (i.e., image size in the pre-trained setting of ViT). Next, we introduce rotation matrices $\boldsymbol{R}_x$ and $\boldsymbol{R}_y$ in the self-attention as definated in Section 3.3:

$$\Phi(m,n) = \boldsymbol{e}_m^T \boldsymbol{W}_q^T (\boldsymbol{R}_{x_m - x_n} + \boldsymbol{R}_{y_m - y_n}) \boldsymbol{W}_k \boldsymbol{e}_n, \tag{3}$$

Local Map Semantic Extraction. Based on the local trajectory image of each road user, we extract corresponding local patch indices $I^\nu = \{i_1, i_2, \cdots, i_{N_s}\}$ from the global map. The local map patch embeddings can be retrieved from $\boldsymbol{\mathcal{G}}_m$, denoted as $\boldsymbol{\mathcal{G}}_s^\nu = \langle \boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \cdots, \boldsymbol{e}_{i_{N_s}} \rangle$.

MultiScale Vision Interaction. This module is designed to model scenario- and direction-aware interaction among road users. We employ a multiscale strategy to extract valuable information from coarse-grained visual features. At the ℓ level, the local trajectory images of the target and interacting road users are hierarchically encoded by convolutional layers. Incorporating $\boldsymbol{\mathcal{G}}_s^\nu$, map semantics fused embeddings $\{\boldsymbol{E}_0^\nu | \nu \in \{\nu_T\} \cup \mathcal{V}_I\}$ are obtrained by a linear combination.

A decomposed attention mechanism is employed to model vision interaction. First, the spatial self-attention mechanism learn the spatial correlation between patch embeddings from $\boldsymbol{E}_0^\nu$ for each road user, which is formulated as follows:

$$\boldsymbol{E}_1^\nu = \text{softmax}\left(\frac{\boldsymbol{Q}^\nu (\boldsymbol{K}^\nu)^T}{\sqrt{d_s}}\right) \boldsymbol{V}^\nu, \tag{4}$$

where $\boldsymbol{Q}^\nu$, $\boldsymbol{K}^\nu$ and $\boldsymbol{V}^\nu$ are derived from $\boldsymbol{E}_0^\nu$.

To capture overall behaviour representation for each road user, a self-view convolution is applied in $\boldsymbol{E}_1^\nu$ to aggregate multi-channel features into a single-channel representation $\boldsymbol{E}_2^\nu$ with a kernel of size 1×1. The cross-view attention aims to model the interaction between the target and interacting road users within the same feature channel. The interaction embedding $\boldsymbol{E}_3^{\nu_T}$ is obtained as follows:

$$\boldsymbol{E}_3^{\nu_T} = \text{softmax}\left(\frac{\boldsymbol{Q}^{\nu_T} (\boldsymbol{K}^{\nu_I})^T}{\sqrt{d_s}}\right) \boldsymbol{V}^{\nu_I}, \tag{5}$$

where $\boldsymbol{Q}^{\nu_T}$ from $\boldsymbol{E}_1^{\nu_T}$, $\boldsymbol{K}^{\nu_I}$ and $\boldsymbol{V}^{\nu_I}$ from $\boldsymbol{E}_1^{\nu_I}$.

Multiscale Feature Fusion. Given self-view embeddings $\boldsymbol{E}_2^{(\ell)}$ and $\boldsymbol{E}_2^{(\ell-1)}$ at adjacent scale level, the high-level embedding $\boldsymbol{E}_2^{(\ell)}$ is first upsampled by bilinear interpolation, and then fused to $\boldsymbol{E}_2^{(\ell-1)}$ using a one-dimentional convolution $\mathcal{P}(\cdot)$. This coarse-to-fine aggregated method is formulated as follows, especially $\tilde{\boldsymbol{E}}_2^{(L)} = \boldsymbol{E}_2^{(L)}$:

$$\tilde{\boldsymbol{E}}_2^{(\ell-1)} = \mathcal{P}(\text{Upsample}(\tilde{\boldsymbol{E}}_2^{(\ell)}), \boldsymbol{E}_2^{(\ell-1)}). \tag{6}$$

We fuse multiscale cross-view patch embeddings with the same calculation in Eq. (6). Finally, we obtain scenario- and direction-aware vision interaction patch embeddings $\tilde{\boldsymbol{E}}_3$ with the token number of N_s.

4.2 Spatio-temporal Trajectory Learner

Trajectory Trend Layer. Given historical trajectories of all road users, we compute each point's position relative to the first point to preserve scale variations of different junctions. In terms of interacting road users appearing at different times, we temporally align the trajectory features by padding all sequences to the same length t_f. Continuous features (e.g., position and velocity) are embedded by a linear layer, and discrete features (e.g., road user class) are fed into index-fetching embedding layers. The embedded vector of τ_t^ν is obtained by summing all feature embeddings and then fed into a Long Short-Term Memory network to capture temporal dependencies, resulting in $\boldsymbol{h}_t^\nu$.

Vision Informed MoE. The behavioral patterns of road users are strongly constrained by the surrounding scenario semantics. Therefore, we introduce a Vision Informed Mixture-of-Experts that integrates gating networks with a vision-guided bias compensation mechanism, capturing scenario-general and scenario-specific behavioral patterns for the target and interacting road users. Specifically, we apply the multiscale aggregated self-view embedding $\tilde{\boldsymbol{E}}_2^\nu$ as visual prior information, integrating the fine-grained trajectory movement features embedding $\boldsymbol{h}_t^\nu$, the output is defined as follows:

$$\tilde{\boldsymbol{h}}_t^\nu = \sum_{n=1}^{N_\alpha} g_n(\boldsymbol{h}_t^\nu, \tilde{\boldsymbol{E}}_2^\nu) f_n^{(a)}(\mathbf{h}_t^\nu) + \sum_{m=1}^{N_c} \boldsymbol{\rho}_\nu f_m^{(c)}(\boldsymbol{h}_t^\nu), \tag{7}$$

where $\boldsymbol{\rho}_\nu$ is a one-hot vector indicating the gating network selects road user class indexed expert network output. g_n denote behaviour-awre gating networks where scenario-general patterns are extracted by $\boldsymbol{h}_t^\nu$ and scenery-specific patterns are captued by $\tilde{\boldsymbol{E}}_2^\nu$. $f^{(a)}$ and $f^{(c)}$ denote the output of the expert network, which is implemented by two MLP layers. To promote different experts to learn diverse behaviour patterns and provide scenario-adaptive selection of expert subsets. We apply the Top-$\mathcal{K}_\gamma$ gating strategy in gating networks, where $\Psi(\cdot)$ denotes the Softmax function:

$$g(\boldsymbol{h}_t^\nu, \tilde{\boldsymbol{E}}_2^\nu) = \Psi(\text{TopK}(\text{MLP}_1(\boldsymbol{h}_t^\nu) + \text{MLP}_2(\tilde{\boldsymbol{E}}_2^\nu), K_\gamma)),$$
$$\text{TopK}(\boldsymbol{\beta}, \mathcal{K}_\gamma) = \begin{cases} \boldsymbol{\beta}_j & \text{if } \beta_j \text{ in top } \mathcal{K}_\gamma \text{ subset of } \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Trajectory Interaction Embeddings. Conditioned on the behavioral representation $\tilde{\boldsymbol{h}}_t^\nu$, we apply a Graph Attention Network to model spatial interactions between the target and interacting road users. Formally, we compute edge weight with polar diameter(relative distance) and polar angle:

$$\boldsymbol{\epsilon}_t^{\nu_I,\nu_T} = \boldsymbol{W}_\epsilon^T(\tilde{\boldsymbol{h}}_t^{\nu_T} \| \tilde{\mathbf{h}}_t^{\nu_I} \| \boldsymbol{\eta}_t^{\nu_I,\nu_T}),$$
$$\boldsymbol{\alpha}_t^{\nu_I,\nu_T} = \frac{\exp(\sigma(\boldsymbol{\epsilon}_t^{\nu_I,\nu_T})}{\sum_{\nu_I \in \mathcal{V}_t^I} \exp(\sigma(\boldsymbol{\epsilon}_t^{\nu_I,\nu_T}))}, \tag{9}$$

where $\boldsymbol{W}_\epsilon$ is learning paramter, and σ is the LeakyReLU activation function. We explicitly encode polar diameter d_t and polar angle θ_t features as $\boldsymbol{\eta}_t^{\nu_T,\nu_I} = \text{MLP}_3(d_t^{\nu_T,\nu_I}, \theta_t^{\nu_T,\nu_I})$ between ν_T and ν_I. Finally, the interaction embeddings $\boldsymbol{O} = [\boldsymbol{o}_{t_0-t_f+1}^{\nu_T} \| \boldsymbol{o}_{t_0-t_f+2}^{\nu_T} \| \cdots \| \boldsymbol{o}_{t_0}^{\nu_T}]$ is obtained as the output of this module:

$$\boldsymbol{o}_t^{\nu_T} = \boldsymbol{W}_1 \tilde{\boldsymbol{h}}_t^{\nu_T} + \sum_{\nu \in \mathcal{V}_t^I} \boldsymbol{\alpha}_t^{\nu_I,\nu_T} \tilde{\boldsymbol{h}}_t^{\nu_I}. \tag{10}$$

4.3 Multimodal Fusion and Optimization

We first separately project vision and trajectory interaction embeddings $\tilde{\boldsymbol{E}}_3$ and $\boldsymbol{O}$ into a shared embedding space integrated with position embeddings and corresponding modality type embeddings. Then, the multimodal tokens are concatenated into a combined sequence, which is fed into shared N_d-depth Transformer Encoder layers to learn cross-modal interaction semantics. Finally, we obtain outputting embeddings $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{O}}$. Due to the larger number of vision tokens compared to trajectory tokens, this dominates the sparse visual attention distribution and hinders effective intra-trajectory modeling. We introduce a perceiver resampler (trainable cross-attention) to reduce the redundant number of visual tokens while preserving critical context information.

The model is trained using multimodal loss. For the trajectory modality, $\hat{\boldsymbol{O}}$ is fed into multi-head decoder layers to produce diverse future trajectories with corresponding probability $\{\hat{Y}_k, \hat{\pi}_k\}_{k=1}^{N_f}$. We optimize the most probable prediction $k^* = \arg\max_{k \in N_f}(\hat{\pi}_k)$ using RMSE loss. Besides, the predicted probability distribution is supervised using the cross-entropy loss, where the target distribution is calculated based on a softmax over the negative final displacement error:

$$\mathcal{L}_{\text{reg}} = \sum_{t=t_0}^{t_0+t_f} \sqrt{(\hat{\boldsymbol{\tau}}_{t,k^*}^{\nu_T} - \boldsymbol{\tau}_t^{\nu_T})^2}, \mathcal{L}_{\text{cls}} = \sum_{k=1}^{N_f} -\pi_k^{\nu_T} \log(\hat{\pi}_k^{\nu_T}). \tag{11}$$

For the vision modality, we obtain label patch embeddings $\bar{\boldsymbol{E}}$ by applying a pre-trained vision encoder to the future local trajectory image of ν_T. The KL-Divergence is employed to encourage consistency between $\bar{\boldsymbol{E}}$ and $\hat{\boldsymbol{E}}$:

$$\mathcal{L}_{vis} = D_{KL}(\hat{\boldsymbol{E}} \| \bar{\boldsymbol{E}}). \tag{12}$$

5 Experiments

Datasets. We use three real-world datasets: SinD-Tianjin [Xu *et al.*, 2022], InD-Bendplatz [Bock *et al.*, 2020], and RounD-Neuweiler [Krajewski *et al.*, 2020]. The SinD-Tianjin dataset is collected at a signalized intersection in Tianjin, China, recorded at 10 Hz. The InD-Bendplatz dataset is recorded at an unsignalized intersection in Germany at 25 Hz. The RounD-Neuweiler dataset contains trajectories collected from a roundabout in Germany with the same sampling frequency as InD. We observe 20 trajectory points (4s) to predict the future 20 positions (4s), where trajectories from different datasets are resampled at adaptive frequencies to enable uniform temporal intervals. We also filter trajectories with long stationary periods or negligible movement, and split each dataset into training, validation, and test sets in a 6:2:2 ratio based on recording time.

Baselines and Metrics. We compare G-VTM with state-of-the-art trajectory prediction models that focus on different traffic scenarios. First, we include junction-oriented trajectory prediction models: MTR [Shi *et al.*, 2022b], HPNet [Tang *et al.*, 2024], BAT [Liao *et al.*, 2024b], and KIGAN [Wei *et al.*, 2024], which explicitly model road topologies and interactions of multitype road users. Second, we evaluate several methods designed for highway scenarios that excel at capturing temporal consistency and behaviour-awareness, such as CS-LSTM [Deo and Trivedi, 2018], WSiP [Wang *et al.*, 2023], STDAN [Chen *et al.*, 2022], HLTP [Liao *et al.*, 2024a], and C2F-TP [Wang *et al.*, 2025]; Finally, we also include pedestrian trajectory prediction baselines: SocialCircle [Wong *et al.*, 2024], AdapTraj [Qian *et al.*, 2024], and UniEdge [Li *et al.*, 2025], which emphasize complex interaction modeling and multi-modal(diverse trajectory predictions) behavior representation.

Evaluation Metrics. We use four types of basic metrics, the same as MTP-GO [Westny *et al.*, 2023]: ADE (Average Displacement Error), FDE (Final Displacement Error), MR (Miss Rate), and APDE (Average Path Displacement Error). For multi-modal K predictions, the minimum ADE (MinADE) and minimum FDE (MinFDE) are evaluated, which are equal to the deterministic ADE and FDE when $K = 1$.

Setting. In generalization experiments, we employ a unified set of model parameters for all datasets. Specifically, we configure the hyperparameter combinations $\{N_\alpha, \mathcal{K}_\gamma, N_d\}$ as $\{16, 4, 6\}$. Our method is trained with the AdamW optimizer (learning rate 10-4), a batch size of 32, an early stopping patience of 8, and a maximum of 100 epochs on the training set. The cosine learning rate scheduler with warmup is employed to stabilize the early-stage training and gradually anneal the learning rate for improved convergence, utilizing the first 5% of training steps for linear warmup. All experiments are conducted on Intel Rapids-SCPCUs and NVIDIA A40 GPUs.

Datasets	SinD-TianJin				InD-Bendplatz				RounD-Neuweiler			
Metric / Method	ADE	FDE	APDE	MR	ADE	FDE	APDE	MR	ADE	FDE	APDE	MR
CS-LSTM	1.296	3.549	0.602	0.605	1.593	4.212	0.834	0.635	1.697	4.322	1.011	0.737
STDAN	1.393	4.349	0.646	0.645	1.713	4.771	0.818	0.682	1.502	4.448	0.883	0.749
WSiP	1.168	3.097	0.541	0.558	1.466	3.884	0.559	0.566	1.822	4.575	1.075	0.766
HLTP	1.702	4.742	0.789	0.702	1.614	4.155	0.726	0.675	2.599	7.425	1.359	0.908
BAT	2.154	3.278	1.542	0.715	2.172	4.252	1.038	0.67	2.469	4.264	1.414	0.729
C2F-TP	1.289	3.102	0.565	0.578	1.319	3.332	0.543	0.537	1.792	4.618	0.929	0.707
KI-GAN	0.788	2.112	0.432	0.418	0.893	2.636	0.447	0.474	1.079	2.862	0.692	0.574
AdapTraj	1.074	2.777	0.538	0.498	1.600	4.151	0.752	0.649	1.668	4.087	0.998	0.649
SocialCircle	1.853	4.024	0.832	0.729	2.271	5.667	1.058	0.793	3.252	8.081	2.243	0.895
UniEdge	1.542	3.707	0.731	0.634	2.060	4.985	1.216	0.765	2.019	4.738	1.256	0.765
MTR	0.653	1.960	0.378	0.381	0.920	2.783	0.506	0.509	0.969	2.836	0.666	0.555
HPNet	0.719	**1.765**	0.375	**0.354**	1.001	2.734	0.515	0.478	1.092	2.684	0.664	0.531
G-VTM	**0.647**	1.895	**0.368**	0.367	**0.856**	**2.607**	**0.428**	**0.456**	**0.778**	**2.428**	**0.504**	**0.441**

Table 1: Single Dataset learning. Lower values indicate better performance. **Red**: best, Blue: second-best.

	Datasets	SinD → InD			SinD → RounD			InD → SinD			InD → RounD			RounD → SinD			RounD → InD		
	Metric / Method	ADE	FDE	MR	ADE	FDE	MR	ADE	FDE	MR	ADE	FDE	MR	ADE	FDE	MR	ADE	FDE	MR
5%	CS-LSTM	2.556	6.089	0.779	2.721	6.539	0.824	1.794	5.017	0.707	2.967	7.348	0.911	1.601	4.307	0.692	2.180	5.589	0.734
	WSiP	2.275	5.633	0.651	2.524	6.147	0.873	1.836	4.721	0.745	2.898	7.281	0.910	1.710	4.312	0.635	2.330	5.919	0.736
	AdapTraj	1.647	4.284	0.677	2.008	4.942	0.812	1.351	3.470	0.605	2.110	5.084	0.803	1.276	3.333	0.590	1.638	4.108	0.639
	KI-GAN	1.373	3.760	0.655	1.875	4.584	0.767	1.496	3.550	0.621	1.563	4.996	0.799	1.487	3.426	0.632	1.357	3.913	0.664
	G-VTM	**1.051**	**3.026**	**0.533**	**1.610**	**4.356**	**0.741**	**1.051**	**3.210**	**0.554**	**1.404**	**4.688**	**0.746**	**0.915**	**2.582**	**0.495**	**0.885**	**2.677**	**0.492**
	Imp.(Δ%)	23.5%	19.5%	18.6%	14.1%	4.9%	3.4%	22.2%	7.5%	8.4%	10.2%	6.2%	6.6%	28.3%	22.5%	16.1%	34.8%	31.6%	23.0%
10%	CS-LSTM	2.125	5.414	0.683	2.346	6.173	0.829	1.725	4.617	0.696	2.496	6.564	0.844	1.519	4.061	0.657	1.955	5.105	0.704
	WSiP	2.015	5.157	0.715	2.240	5.557	0.840	1.589	4.041	0.608	2.222	5.371	0.839	1.456	3.652	0.607	2.001	4.957	0.697
	AdapTraj	1.611	4.220	0.698	1.851	4.519	0.744	1.222	3.183	0.591	1.930	4.645	0.772	1.221	3.183	0.565	1.592	4.171	0.688
	KI-GAN	1.262	3.684	0.623	1.468	3.812	0.705	1.317	3.118	0.607	1.265	4.099	0.721	1.195	2.733	0.529	1.353	3.877	0.692
	G-VTM	**0.971**	**2.828**	**0.527**	**1.245**	**3.605**	**0.665**	**0.908**	**2.781**	**0.530**	**1.146**	**3.825**	**0.679**	**0.819**	**2.389**	**0.458**	**0.842**	**2.622**	**0.493**
	Imp.(Δ%)	23.1%	23.2%	15.4%	15.2%	5.4%	5.6%	25.7%	10.8%	10.3%	9.4%	6.7%	5.8%	31.5%	12.6%	13.4%	37.8%	32.4%	28.3%

Table 2: Few-shot learning on 5% and 10% training data. Lower values indicate better performance. **Red**: best, Blue: second-best.

Single-Dataset Learning. We compare G-VTM with other state-of-the-art baselines separately on three datasets. For fairness, all models are set to optimize and output a single deterministic trajectory. As shown in Table 1, G-VTM consistently outperforms most methods in all datasets. Notably, G-VTM achieves these performance gains without relying on diverse future trajectory sampling, highlighting the effectiveness of multimodal (vision and trajectory) semantic fusion.

Few-shot learning. We evaluate the few-shot generalized ability of G-VTM in cross-dataset groups. All models are pre-trained on the source dataset and then fine-tuned on the target dataset by randomly sampling 5% or 10% training data.

We compare G-VTM with baselines that predict a single deterministic trajectory. For fairness, G-VTM is trained to optimize and output the trajectory with the maximum predicted probability. As shown in Table 2, G-VTM consistently outperforms SOTA models in different groups of cross-dataset transfer, achieving an average improvement of 22.2% in ADE, 15.4% in FDE, and 12.6% in MR compared to second-best models with 5% training data. In the few-shot 10% setting, generalized performance further improves by 23.8%, 15.2%, and 13.1%, respectively. These results demonstrate that G-VTM not only predicts a more accurate future trajectory but also enables its classifier to consistently select the most accurate forecast among diverse predictions.

For baselines producing multi-modal K predictions, we optimize the predicted trajectory that has the minimum endpoint error with the ground-truth trajectory. As shown in Table 3, we use the minADE and minFDE as the evaluation metrics. Our method consistently outperforms SOTA baselines. On 5% training data with $K = 5$, G-VTM achieves average improvements of 11.2% in MinADE and 8.7% in MinFDE compared to the second-best baselines. Besieds, G-VTM benefits from increasing K, showing that it can effectively utilize additional multi-modal trajectory hypotheses rather than merely generating redundant predictions. For example, on {SinD → RounD} with 10% data, G-VTM improves MinADE by 37.4%, and MinFDE by 25.1% with K increased. This improvement is also evident in other cross-dataset groups, which shows that G-VTM exhibits a higher generalization upper bound compared to other methods.

Zero-shot Learning. We evaluate the zero-shot generalized performance of G-VTM with KI-GAN and AdapTraj on several cross-dataset groups. With the similar priority-based traffic regulation and different road topologies, G-VTM achieves an average improvement of 9.0% and 7.4% on {InD → RounD} and {RounD → InD}. On {InD → SinD} that has similar road topologies and different traffic regulations,

Datasets			SinD → InD		SinD → RounD		InD → SinD		InD → RounD		RounD → SinD		RounD → InD	
Metric / Method			MinADE	MinFDE	MinADE	MinFDE	MinADE	MinFDE	MinADE	MinFDE	MinADE	MinFDE	MinADE	MinFDE
5%	K=5	SocialCircle	1.722	4.350	2.163	4.965	1.108	2.765	2.017	4.805	1.349	3.367	1.776	4.238
		UniEdge	1.591	3.535	2.093	4.541	1.394	3.206	2.009	4.283	1.273	2.841	1.610	3.614
		MTR	1.633	4.477	1.485	3.493	0.769	2.549	1.357	3.670	0.566	1.559	0.888	2.579
		HPNet	1.321	3.034	1.221	2.997	0.857	2.718	1.411	3.543	0.658	1.678	0.925	2.386
		G-VTM	**1.286**	**2.857**	**1.174**	**2.816**	**0.680**	**2.077**	**1.041**	**3.364**	**0.615**	**1.485**	**0.781**	**2.223**
	K=10	SocialCircle	1.257	3.029	1.517	3.340	0.872	2.120	1.573	3.460	1.020	2.507	1.393	3.315
		UniEdge	1.479	3.194	1.937	4.042	1.238	2.756	1.790	3.772	1.114	2.438	1.435	3.046
		MTR	1.147	2.935	1.127	3.363	0.564	1.494	1.325	3.619	0.560	1.520	0.884	2.404
		HPNet	1.171	2.465	1.005	2.514	0.514	1.541	1.256	3.323	0.544	1.415	0.897	2.345
		G-VTM	**1.068**	**2.266**	**0.890**	**2.174**	**0.489**	**1.373**	**0.995**	**3.203**	**0.483**	**1.090**	**0.671**	**1.883**
10%	K=5	SocialCircle	1.479	3.586	2.045	4.615	1.007	2.515	1.948	4.524	1.310	3.177	1.694	3.991
		UniEdge	1.542	3.429	1.862	4.011	1.312	3.036	1.759	3.811	1.179	2.677	1.500	3.346
		MTR	0.909	2.627	1.498	4.320	0.634	1.885	1.135	3.159	0.545	1.501	0.875	2.578
		HPNet	0.882	2.323	1.227	3.109	1.051	2.470	1.273	3.293	0.518	1.473	0.831	2.240
		G-VTM	**0.865**	**2.085**	**1.034**	**2.486**	**0.532**	**1.616**	**0.763**	**2.522**	**0.540**	**1.297**	**0.694**	**1.912**
	K=10	SocialCircle	1.122	2.712	1.397	3.011	0.816	1.976	1.443	3.179	0.989	2.362	1.246	2.948
		UniEdge	1.323	2.794	1.636	3.331	1.177	2.630	1.575	3.246	1.055	2.300	1.325	2.833
		MTR	1.068	2.930	1.370	3.770	0.525	1.389	1.138	3.103	0.503	1.356	0.868	2.540
		HPNet	0.816	2.121	0.909	2.313	0.554	1.395	1.075	2.701	0.524	1.131	0.816	1.975
		G-VTM	**0.658**	**1.544**	**0.647**	**1.629**	**0.458**	**1.286**	**0.685**	**2.131**	**0.436**	**0.962**	**0.584**	**1.481**

Table 3: Multi-modal few-shot learning on 5% and 10% training data. Lower values indicate better performance.

Datasets	RounD → InD		InD → SinD		InD → RounD	
Metric / Method	ADE	FDE	ADE	FDE	ADE	FDE
AdapTraj	3.503	9.33	3.826	9.601	7.004	16.172
KI-GAN	1.763	5.693	**2.981**	**8.107**	7.037	15.506
G-VTM	**1.578**	**5.202**	3.258	8.788	**6.508**	**14.542**

Table 4: Zero-shot learning.

Vision			Trajectory		SinD		SinD→RounD	
Cross-View Attention	Self-View Conv	ROPE	LSTM & GAT	MoE	ADE	FDE	ADE	FDE
×	×	×	✓	×	1.015	2.937	2.129	5.407
×	×	×	✓	✓	0.924	2.714	1.921	5.197
✓	×	✓	✓	✓	0.736	2.107	1.738	4.778
×	✓	✓	✓	✓	0.701	2.137	1.722	4.914
✓	✓	×	✓	✓	0.681	2.043	1.814	5.125
✓	✓	✓	✓	✓	**0.647**	**1.895**	**1.610**	**4.356**

Table 5: Ablation of G-VTM components.

KI-GAN, which is designed for the SinD dataset, achieves the best performance by explicitly encoding traffic signal states.

Ablation Study. We evaluate the contribution of each module in G-VTM by systematically adding individual components. Results in Table 5 prove that all modules are critical to G-VTM. First, adding vision-related modules consistently improves performance under both in-dataset and cross-dataset learning. In particular, excluding vision modality in G-VTM leads to a significant performance drop (ADE-30% in SinD-Tianjin and ADE-16% in SinD→RounD with 5% training data). Moreover, including the MoE module leads to a significant performance improvement, indicating that MoE is essential for modeling behaviour patterns of road users.

Analysis of the Vision-informed MoE. As shown in Figure 3, we analyze the experts activated in different traffic

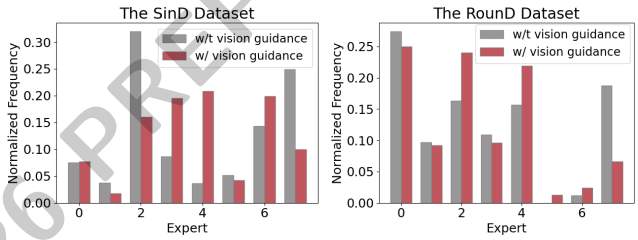

Figure 3: Vision-informed MoE expert analysis.

scenarios, including the signaled intersection and the roundabout. First, the distribution of activated experts varied across different junctions because of distinct traffic behavior patterns. Furthermore, the introduction of visual information enables learning scenario-adaptive behavior patterns. It promotes more experts to be activated, capturing diverse behavioral motions. Besides, rich scenario semantics are provided by the vision modality to correct expert weights, thereby guiding more accurate intention modeling.

Computation Study. G-VTM exhibits competitive computational efficiency with a total of 76.5M parameters, 13190 MiB of memory usage, and 1.268 s/iter of the average inference speed. This efficiency enables scalable deployment across diverse junctions, supporting effective generalization without incurring prohibitive computational overhead.

6 Conclusion

G-VTM is a pioneering exploration of multimodal techniques for generalized trajectory prediction in complex traffic scenarios. Extensive experiments demonstrate its state-of-the-art generalized performance across different junction datasets, especially in few-shot settings. G-VTM supports low-cost deployment and low-latency inference, making it well-suited for practical application in intelligent transportation and vehicle–infrastructure systems.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No. L252034), Frontier Technologies R&D Program of Jiangsu (Grant No. BF2024052), and Chengdu Science and Technology Program (Grant No. 2025-YF08-00097-GX).

Contribution Statement

The authors Xinyue Zhang and Letian Gong contributed equally to this research.

References

[Abdelraouf *et al.*, 2022] Amr Abdelraouf, Mohamed Abdel-Aty, Zijin Wang, and Ou Zheng. Trajectory prediction for vehicle conflict identification at intersections using sequence-to-sequence recurrent neural networks. *arXiv preprint arXiv:2210.08009*, 2022.

[Bock *et al.*, 2020] Julian Bock, Robert Krajewski, Tobias Moers, Steffen Runde, Lennart Vater, and Lutz Eckstein. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1929–1934. IEEE, 2020.

[Cai *et al.*, 2021] Yingfeng Cai, Zihao Wang, Hai Wang, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li. Environment-attention network for vehicle trajectory prediction. *IEEE Transactions on Vehicular Technology*, 70(11):11216–11227, 2021.

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

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

[Huang *et al.*, 2020] Zijie Huang, Yizhou Sun, and Wei Wang. Learning continuous system dynamics from irregularly-sampled partial observations. *Advances in Neural Information Processing Systems*, 33:16177–16187, 2020.

[Huang *et al.*, 2025] Feilong Huang, Zide Fan, Xiaohe Li, Wenhui Zhang, Pengfei Li, Ying Geng, and Keqing Zhu. Tailored meta-learning for dual trajectory transformer: advancing generalized trajectory prediction. *Complex & Intelligent Systems*, 11(3):174, 2025.

[Kawasaki and Tasaki, 2018] Atsushi Kawasaki and Tsuyoshi Tasaki. Trajectory prediction of turning vehicles based on intersection geometry and observed velocities. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 511–516. IEEE, 2018.

[Krajewski *et al.*, 2020] Robert Krajewski, Tobias Moers, Julian Bock, Lennart Vater, and Lutz Eckstein. The round dataset: A drone dataset of road user trajectories at roundabouts in germany. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2020.

[Li *et al.*, 2019] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving. *arXiv preprint arXiv:1907.07792*, 2019.

[Li *et al.*, 2025] Ruochen Li, Tanqiu Qiao, Stamos Katsigiannis, Zhanxing Zhu, and Hubert PH Shum. Unified spatial-temporal edge-enhanced graph networks for pedestrian trajectory prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[Liao *et al.*, 2024a] Haicheng Liao, Yongkang Li, Zhenning Li, Chengyue Wang, Zhiyong Cui, Shengbo Eben Li, and Chengzhong Xu. A cognitive-based trajectory prediction approach for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 9(4):4632–4643, 2024.

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

[Mo *et al.*, 2021] Xiaoyu Mo, Yang Xing, and Chen Lv. Graph and recurrent neural network-based vehicle trajectory prediction for highway driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1934–1939. IEEE, 2021.

[Qian *et al.*, 2024] Tangwen Qian, Yile Chen, Gao Cong, Yongjun Xu, and Fei Wang. Adaptraj: A multi-source domain generalization framework for multi-agent trajectory prediction. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5048–5060. IEEE, 2024.

[Rowe *et al.*, 2023] Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. Fjmp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13745–13755, 2023.

[Schimpf *et al.*, 2023] Nora Schimpf, Zhe Wang, Summer Li, Eric J Knoblock, Hongxiang Li, and Rafael D Apaza. A generalized approach to aircraft trajectory prediction via supervised deep learning. *IEEE Access*, 11:116183–116195, 2023.

[Shi *et al.*, 2022a] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[Shi *et al.*, 2022b] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[Song *et al.*, 2020] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. Pip: Planning-informed trajectory prediction for autonomous driving. In *European conference on computer vision*, pages 598–614. Springer, 2020.

[Sun *et al.*, 2023] Yingbo Sun, Tao Xu, Jingyuan Li, Yuan Chu, and Xuewu Ji. Mmh-sta: A macro-micro-hierarchical spatio-temporal attention method for multi-agent trajectory prediction in unsignalized roundabouts. *IEEE Transactions on Vehicular Technology*, 72(9):11237–11250, 2023.

[Tang *et al.*, 2024] Xiaolong Tang, Meina Kan, Shiguang Shan, Zhilong Ji, Jinfeng Bai, and Xilin Chen. Hpnet: Dynamic trajectory forecasting with historical prediction attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15261–15270, 2024.

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

[Wang *et al.*, 2025] Zichen Wang, Hao Miao, Senzhang Wang, Renzhi Wang, Jianxin Wang, and Jian Zhang. C2f-tp: A coarse-to-fine denoising framework for uncertainty-aware trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12810–12817, 2025.

[Wei *et al.*, 2024] Chuheng Wei, Guoyuan Wu, Matthew J Barth, Amr Abdelraouf, Rohit Gupta, and Kyungtae Han. Ki-gan: Knowledge-informed generative adversarial networks for enhanced multi-vehicle trajectory forecasting at signalized intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7115–7124, 2024.

[Westny *et al.*, 2023] Theodor Westny, Joel Oskarsson, Björn Olofsson, and Erik Frisk. Mtp-go: Graph-based probabilistic multi-agent trajectory prediction with neural odes. *IEEE Transactions on Intelligent Vehicles*, 8(9):4223–4236, 2023.

[Wong *et al.*, 2024] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015, 2024.

[Xu *et al.*, 2022] Yanchao Xu, Wenbo Shao, Jun Li, Kai Yang, Weida Wang, Hua Huang, Chen Lv, and Hong Wang. Sind: A drone dataset at signalized intersection in china. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2471–2478. IEEE, 2022.

[Yang *et al.*, 2025] Jiayu Yang, Jaeyoung Jay Lee, and Constantinos Antoniou. Trajectory prediction for multiple agents in dynamic environments: Factoring in traffic states and driving styles. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[Zhao *et al.*, 2020] Liang Zhao, Yufei Liu, Ahmed Y Al-Dubai, Albert Y Zomaya, Geyong Min, and Ammar Hawbani. A novel generation-adversarial-network-based vehicle trajectory prediction method for intelligent vehicular networks. *IEEE Internet of Things Journal*, 8(3):2066–2077, 2020.